Open Access

Human-mimetic binaural ear design and sound source direction estimation for task realization of musculoskeletal humanoids

Yusuke Omura[1], Kento Kawaharazuka[1*] , Yuya Nagamatsu[1], Yuya Koga[1], Manabu Nishiura[1], Yasunori Toshimitsu[1], Yuki Asano[1], Kei Okada[1], Koji Kawasaki[2] and Masayuki Inaba[1]

Abstract

Human-like environment recognition by musculoskeletal humanoids is important for task realization in real complex environments and for use as dummies for test subjects. Humans integrate various sensory information to perceive their surroundings, and hearing is particularly useful for recognizing objects out of view or out of touch. In this research, we aim to realize human-like auditory environmental recognition and task realization for musculoskeletal humanoids by equipping them with a human-like auditory processing system. Humans realize sound-based environmental recognition by estimating directions of the sound sources and detecting environmental sounds based on changes in the time and frequency domain of incoming sounds and the integration of auditory information in the central nervous system. We propose a human mimetic auditory information processing system, which consists of three components: the human mimetic binaural ear unit, which mimics human ear structure and characteristics, the sound source direction estimation system, and the environmental sound detection system, which mimics processing in the central nervous system. We apply it to Musashi, a human mimetic musculoskeletal humanoid, and have it perform tasks that require sound information outside of view in real noisy environments to confirm the usefulness of the proposed methods.

Keywords: Biomimetic, Environmental sound recognition, Sound source direction estimation, Musculoskeletal humanoid

Introduction

The musculoskeletal humanoids [1–4], which mimic the human body structure in detail, are expected to be used for environmental contact behavior and as subject dummies by using their flexible body structure effectively. In order to use them as substitutes for humans, they need to recognize environments as humans do. Research on vision using movable eyes [5] and tactile sense using the hand [6] and foot [7] with distributed force sensors have been conducted. However, humans use not only visual and tactile senses but also auditory perception to recognize the state of objects out of view or out of touch. In order to have musculoskeletal humanoids perform tasks based on environmental recognition and use them as subject dummies, human mimetic auditory information processing is also essential in addition to vision and tactile senses. There are two major types of information that we can obtain through hearing: sound source directions and sound types. Humans realize sound-based environmental recognition by integrating these pieces of information.

As for existing sound source direction estimation approaches, high-resolution methods, such as the MUSIC (MUltiple SIgnal Classification) method [8], have been proposed, and research using robot audition

*Correspondence: kawaharazuka@jsk.imi.i.u-tokyo.ac.jp
[1] Department of Information Science and Technology, The University of Tokyo, Tokyo, Japan
Full list of author information is available at the end of the article

Omura *et al. ROBOMECH Journal* (2022) 9:17

Page 2 of 15

have been conducted widely. However, most methods are proposed for multi-channel systems with three or more channels, and in the case of human-like inputs with only two channels, the accuracy of sound source direction estimation in 3D is reduced due to the influence of background noise in the real environment. GCC-PHAT (Generalized Cross-Correlation PHAse Transform) [9], which targets two channels of input, uses the time difference between left and right inputs to estimate the sound source direction, but the range of estimated direction is limited. Also, in order to improve the accuracy of those methods, the required sound sampling length becomes longer.

In robot audition, which aims to achieve human-like environmental recognition, environmental sound recognition using a pair of microphones is a particularly important issue. Sound source localization and understanding the meaning of sound in various scenes are important in task realization, and integrating them with other senses enables more accurate environmental recognition. SIG [10], a humanoid that has a human-like head and torso, achieves tracking while filtering sound inputs based on epipolar geometry and vision. The sound recognition system, which integrates sound source localization, sound source separation, and speech recognition [11], achieves simultaneous recognition of multiple speech signals by filters depending on sound source directions. Also, implementation of this method into a robot in the real world is conducted by using a reconfigurable processor [12].
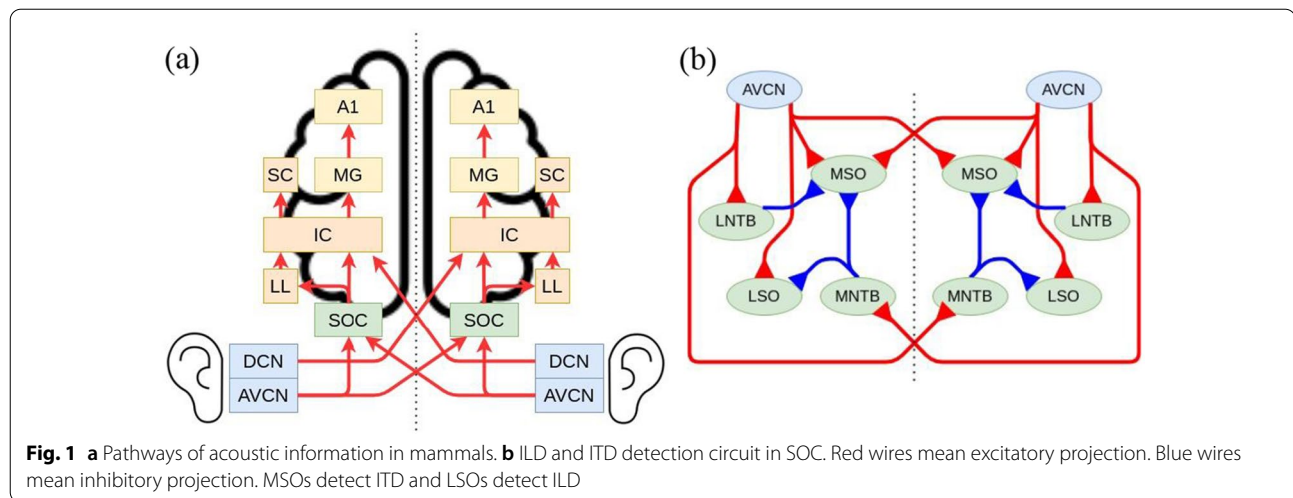
It is widely acknowledged that human sound source direction estimation is related to the complex unevenness in pinnae on the incoming sound. Therefore, research focusing on the shape of the human outer ear has been conducted. In sound source direction estimation using a robot with reflectors shaped like pinnae and a monocular camera [13], the changes in frequency response caused by the reflectors are learned using a self-organizing map. 3D sound source direction estimation is conducted, but it is unsuitable for real environment applications because it requires a single sound source and a broadband signal. The sound source direction estimation method with weighting based on signal-to-noise ratio in addition to GCC-PHAT [14] is applied to SIG-2 [15], a humanoid with human mimetic outer ears. It is possible to estimate the sound source direction for multiple sound sources in the real environment, but there is still the problem that the estimated direction is limited. Sound source localization methods inspired by the human central nervous system [16–18] learn the relationship between the sound source direction and the difference of the left and right sounds. It is possible to estimate the sound source direction in the echoic and noisy environment,

but the estimated direction is limited to the front of the horizontal plane. A two-channel sound source direction estimation method that can be used in the real noisy environment is not limited to broadband sounds, and has a wide range of estimated directions, must be researched.

On the other hand, a number of neural network-based methods for environmental sound detection have been proposed. EnvNet [19] uses sound waves as input to a convolutional neural network (CNN), and acoustic feature maps are formed in the CNN, resulting in highly accurate environmental sound recognition. In addition, the convolutional recurrent neural network-based method [20] realized environmental sound detection by using a model that captures temporal changes in sound. Although these methods achieve high accuracy in detecting environmental sound, they are not designed for real-time use, so the required sound sampling length becomes longer, and they are still not suitable for real-time operation.

The human ear is divided into three parts based on its structural and functional characteristics, the outer ear, middle ear, and inner ear. The outer ear has a pinna with complex unevenness and shows complex frequency response with a sharp gain reduction notch depending on the sound source direction [21]. Also, sound source direction estimation accuracy is significantly decreased when pinna unevenness is filled [22]. In robot audition, implementation of pinnae in a telepresence robot increases sound source localization accuracy of robot users in the median plane [23]. Also, this study says implementing pinnae and head movement increase the sound source localization of telepresence robot users. This complex relationship can be investigated by measuring the Head-Related Transfer Function (HRTF), which describes the relationship between the source sound signal and the actual incoming sound.

Auditory information processing in the central nervous system of mammals is shown in Fig. 1. Auditory information decomposed into frequency components in the cochlea is transmitted to the cochlear nucleus (CN). Signals are transmitted from the anterior ventral cochlear nucleus (AVCN) to the superior olivary complex (SOC), the interaural level difference (ILD), and the interaural time difference (ITD) are extracted in SOC. The lateral nucleus of trapezoid body (LNTB) transmits the contralateral signal as inhibitory projection, and the medial nucleus of trapezoid body (MNTB) transmits the ipsilateral signal as inhibitory projection. ITD is obtained by detecting the simultaneity of sound when the medial superior olive (MSO) receives projections from each side. ILD is obtained by detecting the firing frequency changes by comparing the intensity difference when the lateral superior olive (LSO) receives excitatory projections from

Fig. 1 **a** Pathways of acoustic information in mammals. **b** ILD and ITD detection circuit in SOC. Red wires mean excitatory projection. Blue wires mean inhibitory projection. MSOs detect ITD and LSOs detect ILD

the ipsilateral side, and inhibitory projections from the contralateral side [24]. ILD and ITD extracted in SOC are transmitted to the lateral lemniscus (LL) and the inferior colliculus (IC) and integrated. It is confirmed that there are neurons in IC that respond to sound from specific directions [25]. Auditory information is transmitted from IC to the medial geniculate body (MG) and the primary auditory cortex (A1). Also, auditory information from IC is transmitted to the superior colliculus (SC), which mainly processes visual information, and the spatial map is formed in SC [26].

It can be summarized that the environmental sound recognition function is composed of the following functions;

  (i) Physical effects on frequency response by the complex unevenness of the pinna and frequency decomposition of incoming sounds

  (ii) Sound source direction estimation based on detection of ILD and ITD, and neurons which respond to each sound source directions

  (iii) Environmental sound detection based on the integration of frequency components and temporal changes of incoming sounds.

In this research, we propose a human mimetic auditory information processing system composed of a human mimetic binaural ear, sound source direction estimation system, and environmental sound detection system. The concept of this study is shown in Fig. 2. The rest of this paper is organized as follows. In "Method of human mimetic environmental sound recognition system" section, we explain the proposed system and methods of sound source direction estimation and environmental sound detection. In "Experimental results" section, we apply the proposed method to a musculoskeletal

Fig. 2 The concept of this study
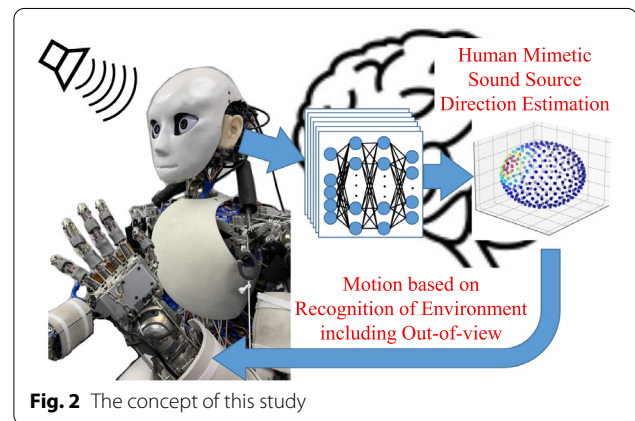
humanoid and conduct task realization experiments that require information out of view. Finally, the discussion and conclusion are presented.

Method of human mimetic environmental sound recognition system

We propose an auditory information processing system as shown in Fig. 3. We explain the components of this system in the following subsections.

Human mimetic binaural ear
In this research, we develop the human mimetic binaural ear unit as shown in Fig. 4 in order to realize the functions (i) and (ii). It is composed of a human mimetic outer ear structure, microphone board, and acoustic processing board and transmits incoming sound information to the humanoid internal computer.

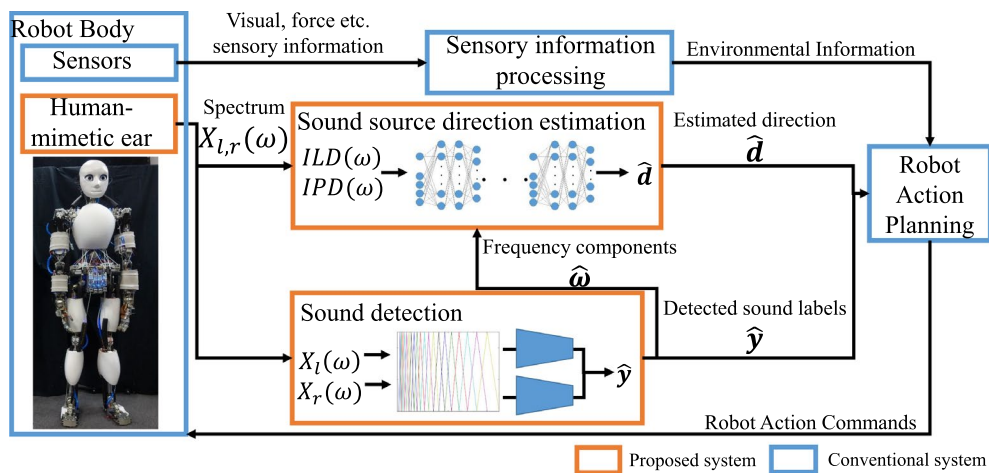Its pinna and cartilaginous part of the ear canal is made of silicone rubber, and its bony part of the ear

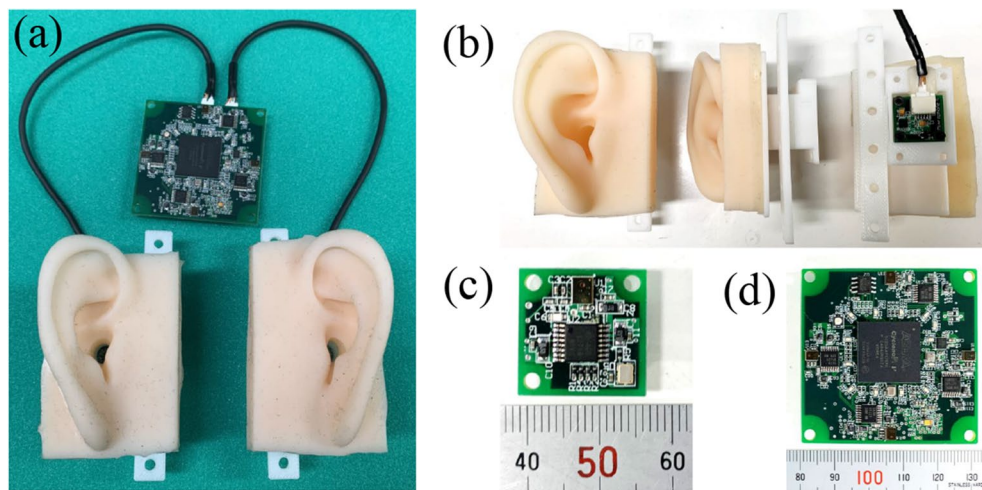**Fig. 3** Overview of proposed system for Human mimetic environmental sound recognition system

Fig. 4 a Overview of developed human mimetic binaural ear unit. **b** Human mimetic outer ear structure. **c** Microphone board. **d** Acoustic processing board

canal and container of microphone board is made of 3D printed ABS resin. The human mimetic outer ear structure has a pinna and ear canal. The distance between the microphone and the open end of the ear canal is approximately 26 mm, and it is about the same length as the human ear canal. The microphone only measures the sound passing its ear canal by covering it with silicone rubber.

The microphone board has a MEMS microphone (ICS-40619) and a 24-bit ADC (ADS1271B) and has human-like acoustic characteristics, as shown in Table. 1. This MEMS microphone has a flat frequency response, particularly from 100 Hz to 10000 Hz,

Table 1 Comparison between the developed microphone board and the human ear

	Mic. board	Human ear
Frequency response[Hz]	$20 \sim 20000$	$20 \sim 20000$
Absolute threshold[μPa]	13.4	~ 20
Acoustic overload point [dBSPL]	132	~ 130
Sensitivity[dBA]	$-39 \; -37$	–
Signal-to-noise ratio [dBA]	67	–

covering everyday sounds and speech. The sampling rate is 44100 Hz.

Omura *et al. ROBOMECH Journal*   (2022) 9:17

Page 5 of 15

The acoustic processing board calculates spectra from incoming sounds by FPGA (Cyclone V). Each spectrum $X_{l,r}$ can be calculated as follows,

$$x'_{l,r}(t) = w_{ham}(t)x_{l,r}(t) \tag{1}$$

$$X_{l,r}(\omega) = \text{FFT}[x'_{l,r}(t)] \tag{2}$$

$$t = 0, \cdots, N-1, \quad \omega = 0, \cdots, N-1 \tag{3}$$

where $x_{l,r}$ is each incoming sound signal, $x'_{l,r}$ is each incoming sound signal after windowing, w_{ham} is a hamming window, $*$ is convolution, FFT is fast Fourier transform, t is time, ω is discrete frequency, and N is the number of FFT points, 2048 in this study. Each spectrum calculated by FFT is transmitted to the humanoid internal computer in a cycle of 62.5 Hz.

Human mimetic sound source direction estimation

Sound source direction estimation described as (ii) is conducted based on left and right spectra. We propose a neural network-based sound source direction estimation method that mimics the process of the human central nervous system. ILD and ITD are projected to IC in humans, but we use interaural phase difference (IPD) instead of ITD in this study. ILD and IPD, inputs for sound source direction estimation, are calculated based on left and right spectra. ILD and IPD can be calculated as follows,

$$\hat{X}_{l,r}(\omega_i) = \frac{X_{l,r}(\omega_i)}{\sqrt{|X_l(\omega_i)|^2 + |X_r(\omega_i)|^2}} \tag{4}$$

$$ILD(\omega_i) = \log|\hat{X}_l(\omega_i)| - \log|\hat{X}_r(\omega_i)| \tag{5}$$

$$IPD(\omega_i) = \begin{pmatrix} \text{Real}[\hat{X}_l(\omega_i)/|\hat{X}_l(\omega_i)|] \\ \text{Imag}[\hat{X}_l(\omega_i)/|\hat{X}_l(\omega_i)|] \\ \text{Real}[\hat{X}_r(\omega_i)/|\hat{X}_r(\omega_i)|] \\ \text{Imag}[\hat{X}_r(\omega_i)/|\hat{X}_r(\omega_i)|] \end{pmatrix} \tag{6}$$

where $\hat{X}_{l,r}$ is left and right normalized spectra at discrete frequency ω_i, Real is the operation to take out the real part, and Imag is the operation to take out the imaginary part. In order to use the spectra obtained by FFT as the input of the neural network with as little modification as possible, the ILD and IPD are determined by trial and error.

For each frequency, sound source existences $\boldsymbol{P}(\boldsymbol{\omega_i})$ can be calculated by Sound Source Direction Estimation Network (SSDENet) shown in Fig. 5, using calculated ILD and IPD as input, as follows,

$$\boldsymbol{P}(\omega_i) = \text{SSDENet}_i(ILD(\omega_i), IPD(\omega_i)) \tag{7}$$

$$\boldsymbol{P}(\omega_i) = \big(P(\boldsymbol{d}_1, \omega_i) \ \dots \ P(\boldsymbol{d}_D, \omega_i)\big)^T \tag{8}$$

where SSDENet_i is SSDENet at discrete frequency ω_i, $\boldsymbol{d}_k$ is a vector which expresses 3-dimensional sound source direction, $P(\boldsymbol{d}_k, \omega_i)$ is a sound source existence at direction $\boldsymbol{d}_k$ and discrete frequency ω_i, and D is the number of directions $\boldsymbol{d}_k$ at which estimation is conducted. SSDENet is a neural network consisting of 4 fully connected layers, with the number of nodes in each layer being 5, 500, 500, and 326, and the activation function is sigmoid. Since SSDENet outputs sound source existences of all
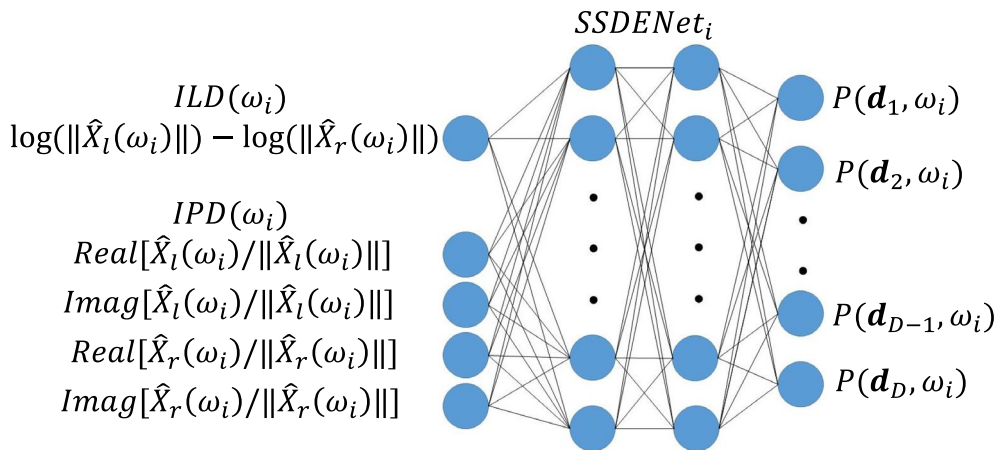

Fig. 5 Structure of SSDENet

directions at once and is structured so that the sound source directions affect each other, it is expected to reduce false estimates more than existing engineering methods that estimate each direction independently.

Estimated sound source direction $\hat{\boldsymbol{d}}$ can be calculated as follows,

$$\hat{\boldsymbol{d}} = \underset{\boldsymbol{d}_k}{\text{argmax}}\, P(\boldsymbol{d}_k) \quad (k = 1, \ldots, D) \tag{9}$$

$$P(\boldsymbol{d}_k) = \sum_i c_i P(\boldsymbol{d}_k, \omega_i) \tag{10}$$

$$c_i = \begin{cases} 1 & (\text{if } \omega_i \text{ is valid}) \\ 0 & (\text{otherwise}) \end{cases} \tag{11}$$

where c_i is a coefficient indicating that the discrete frequency ω_i is valid for estimation. ω_i can be set to match each sound to be estimated. The sound source direction can be estimated while excluding unnecessary frequency components of the subject sound by analytically selecting the frequency components. By estimating only the frequencies of the sound source signal, the influence of other sounds can be suppressed, and the directions of multiple sound sources with different frequency components can be estimated. In practice, the sound source direction estimation is conducted on the frequency components $\hat{\boldsymbol{\omega}}$ output by the binaural environmental sound detection described below.

In order to train SSDENets, we use pre-measured HRTFs and generate training data by placing virtual sound sources. When a virtual sound source is placed in direction $\boldsymbol{d}_h$, the spectra $X_{l,r}^{train}$ and the sound source existence P^{train} are calculated as follows,

$$X_{l,r}^{train}(\omega_i) = \frac{A_{l,r}(\boldsymbol{d}_h, \omega_i)}{\sqrt{|A_l(\boldsymbol{d}_h, \omega_i)|^2 + |A_r(\boldsymbol{d}_h, \omega_i)|^2}} s + n_{l,r} \tag{12}$$

$$P^{train}(d_k, \omega_i) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\Delta d_{k,h}^2}{2\sigma^2}\right) \tag{13}$$

where $A_{l,r}$ are left and right HRTF, s is spectrum of virtual sound source, $n_{l,r}$ are left and right background noise spectra, $\Delta d_{k,h}$ is the angle between the direction $\boldsymbol{d}_k$ and the direction $\boldsymbol{d}_h$, and σ is the variance of the existence distribution. By learning the sound source existences as a distribution with a peak in the correct sound source direction, it is expected for sound source direction to influence the close directions. The loss function of learning is the mean squared error, and the learning rule is Adam.

Binaural environmental sound detection

Environmental sound detection described as (iii) is conducted based on the convolutional neural network (MelCNN) using each Mel spectrogram as input. The structure of MelCNN is based on logMel-CNN [27], as shown in Fig. 6, and consists of two convolutional layers and two fully-connected layers. LogMel-CNN is a simple neural network used in environmental sound recognition. The input of logMel-CNN is a Mel spectrogram, and it can be easily modified in both time and frequency domains. We think that logMel-CNN can be used as a detector by modified to target short sounds and running periodically in practice. MelCNN uses a Mel spectrogram as input and outputs the existence probability of each sound by using sigmoid as the activation function of the output layer. The Mel spectrogram is a series of 25

Fig. 6 Structure of MelCNN based on log-MelCNN [27]

Omura *et al. ROBOMECH Journal*   (2022) 9:17

Page 7 of 15
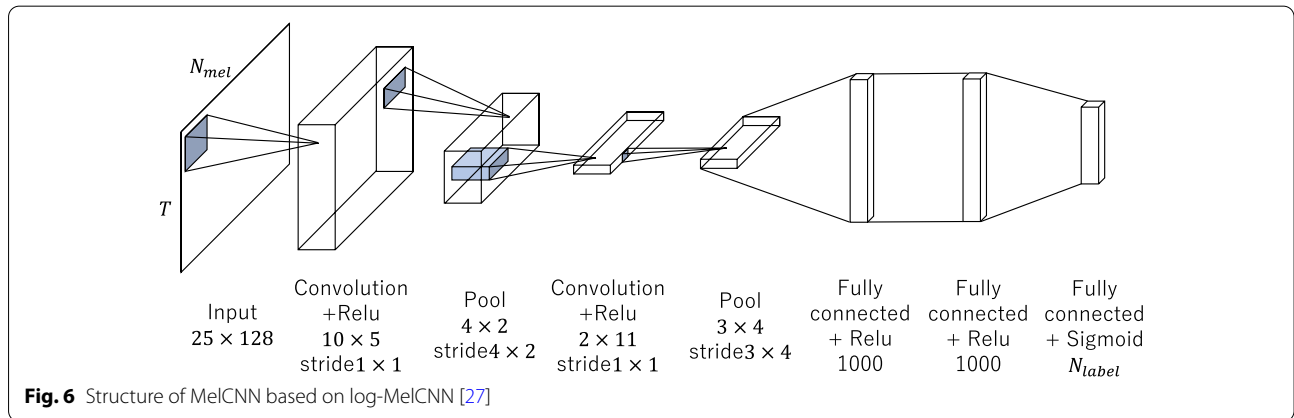
Mel spectra with 128 points on the mel-scale. This Mel spectrogram is normalized to use relative intensity as the input because the volume of incoming sound can easily vary depending on the distance of the sound source in the real environment. The existence probability of the target sound is defined as 0 when the sound is absent and as 1 when the sound is present. MelCNN outputs the existence probabilities of each sound respectively. In order to predict co-occurring labeled sounds, we use each sound data and mixed sound data for the training of MelCNN. In case multiple labeled sounds co-occur when MelCNN is used in a real environment, the outputs of MelCNN are existences probabilities of each sound, and multiple sound detection can be expected. Left and Right Mel spectrograms are input to MelCNN, respectively. The existence probabilities of each sound are calculated by summing the outputs weighted by sound volume ratio as follows,

$$p(y_i) = \frac{V_l p_l(y_i) + V_r p_r(y_i)}{V_l + V_r} \tag{14}$$

where y_i is the label of sound, $p(y_i)$ is the existence probability of sound y_i, $p_{l,r}(y_i)$ is the left and right existence probabilities of sound y_i, and $V_{l,r}$ are the left and right sound volume. This weighting is intended to prioritize the result of a more audible side. The loss function of learning is the KL divergence, and the learning rule is Adam.

Also, the characteristic frequency components $\hat{\omega}$ of each sound are calculated during train data processing. When MelCNN predicts, detected sound labels and corresponding frequency components $\hat{\omega}$ are output, and the frequencies are used for determining which frequency is used for sound source direction estimation.

Experimental results
Effects of human mimetic outer ear structure
Change of frequency response in the median plane
We investigate whether the pinnae of the human mimetic binaural ear unit affect the frequency response in the median plane. A dummy head, as shown in Fig. 7(a), is used in the experiment. The dummy head can rotate in azimuth angle ϕ and elevation angle θ by two servo motors.

A loudspeaker is placed 1.5 m in front of the dummy head to play white noise, while the elevation angle is varied from $-90°$ to $90°$. In the experiment, we compare the outer ear structure with and without pinnae, as shown in Fig. 7(b). The results of this experiment are shown in Fig. 7(c). While moving the neck in elevation, there is little change in the frequency response in the absence of pinnae. However, in the case of outer ear structure with pinnae, the frequency response changes significantly as the elevation changes. In particular, notches occur clearly in the high-frequency band above approximately 10000 Hz, suggesting that introducing
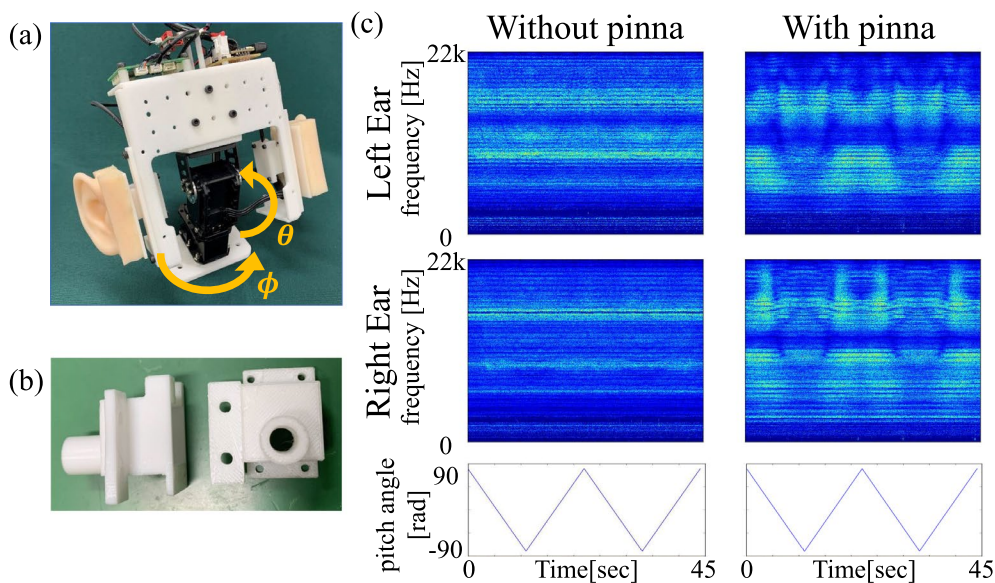

Fig. 7 **a** Dummy head with developed ear unit. **b** Outer ear structure without pinna. **c** The difference of frequency response with and without pinna

the human mimetic outer ear structure can induce notches like humans.

HRTF measurement
In order to confirm that the developed human mimetic binaural ear unit causes complex frequency response, we measured the HRTF of the unit. HRTF measurement is conducted in a conference room with the noise level of 35 dBSPL and the reverberation time RT_{60} of 550~600 msec. HRTF measurement is taken in 326 directions around the dummy head as shown in Fig. 8, and the angle between each direction and the nearest direction is $10° \sim 11.8°$.
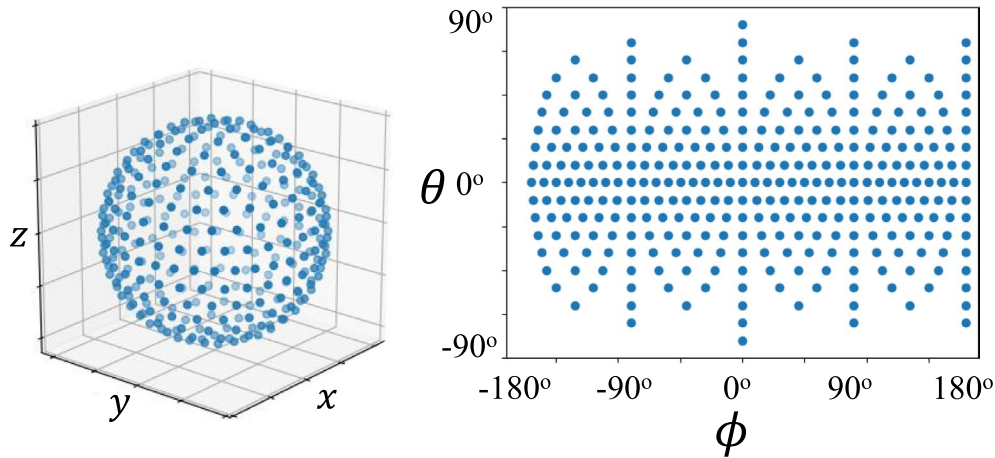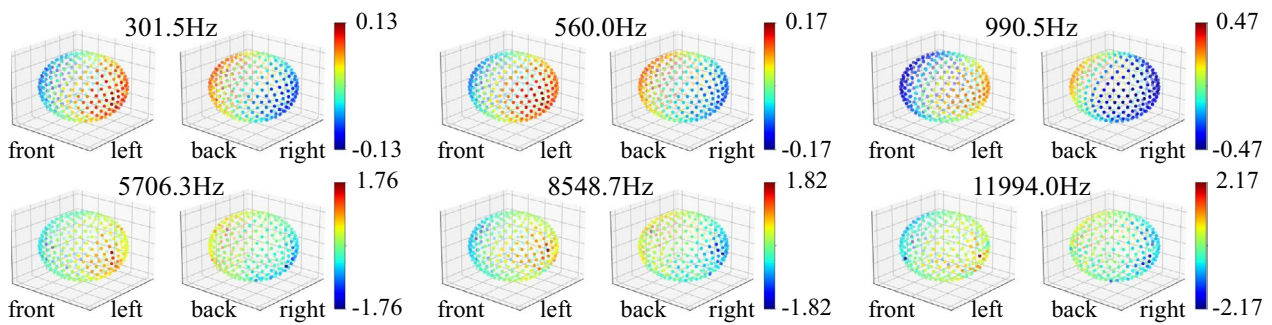

Fig. 8 Directions of HRTF measured
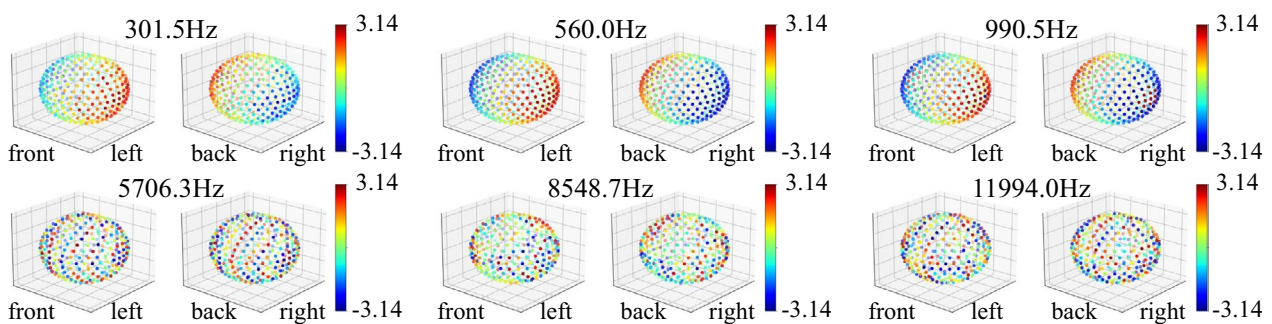

Fig. 9 ILDs calculated from HRTFs for each frequency

Fig. 10 IPDs calculated from HRTFs for each frequency

Omura *et al. ROBOMECH Journal* (2022) 9:17

Page 9 of 15

The results of this experiment are shown in Figs. 9 and 10. ILD and IPD of this experiment are calculated as follows,

$$ILD_{HRTF}(\omega) = \log |A_l(\boldsymbol{d}, \omega)| - \log |A_r(\boldsymbol{d}, \omega)| \quad (15)$$

$$IPD_{HRTF}(\omega) = \arg \frac{A_l(\boldsymbol{d}, \omega)}{A_r(\boldsymbol{d}, \omega)} \quad (16)$$

where arg is the argument of the complex.

First, we describe the result of ILD. In lower frequency bands of 301.5, 560.0, and 990.5 Hz, ILD changes gently from left to right, and the range of ILD is relatively small, ranging from 0.13 to 0.47. On the other hand, in higher frequency bands of 5706.3, 8548.7, 11994.0 Hz, ILD changes in such a way that there are sharp peaks on both sides, and the range of ILD increases from 1.76 to 2.17, indicating that the difference between left and right sides increased.

Next, we describe the result of IPD. Like ILD results, IPD changes gently from left to right in lower frequency bands. On the other hand, in higher frequency bands, IPD changes complicatedly because the period of sound is much smaller than the arrival time difference.

Sound source direction estimation using SSDENet
Simulation

We compare the performance with the MUSIC method [8] in sound source direction estimation when some types of sounds are generated from virtual sources located in the direction where HRTFs are measured. In this experiment, a single source is assumed. In the MUSIC method, the sound source direction is estimated by summing the spatial spectra for detected frequency, and the frequency components used for sound source direction estimation are selected based on the ratio of the eigenvalues of the spatial correlation matrix [28]. The steering vectors for each frequency in the MUSIC method are the HRTFs measured at each frequency in the previous experiment. Spatial smoothing is not applied because the size of the

correlation matrix is 2×2. SSDENets for each frequency component are trained by 1000 sounds generated randomly. The sounds generated by the virtual sound source are sine waves, triangle waves, square waves, and sawtooth waves with fundamental frequencies of 500, 1000, and 2000 Hz, and white noise.

The results of sound source direction estimation for each wave are shown in Fig. 11. The mean sound source direction estimation errors of the proposed method are smaller than those of the MUSIC method under all conditions. In particular, the errors of the MUSIC method for sine waves and triangle waves are around 90°, which is comparable to the random case, but the results of the proposed method are relatively accurate in many directions.
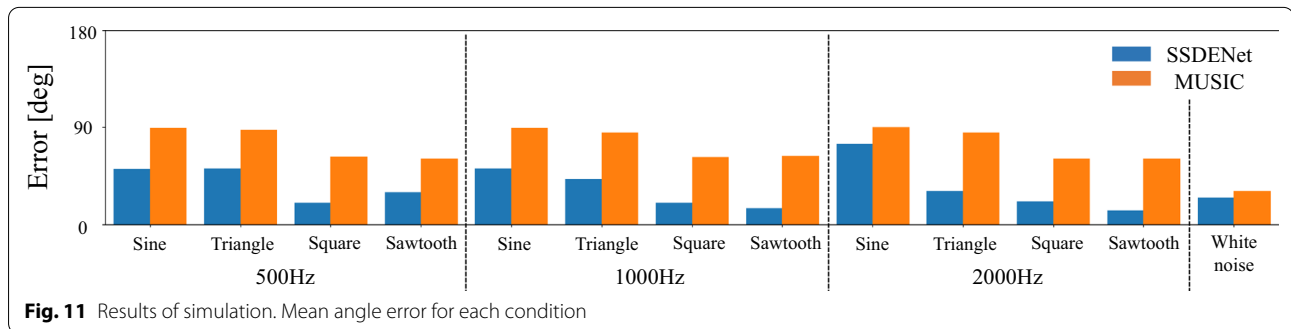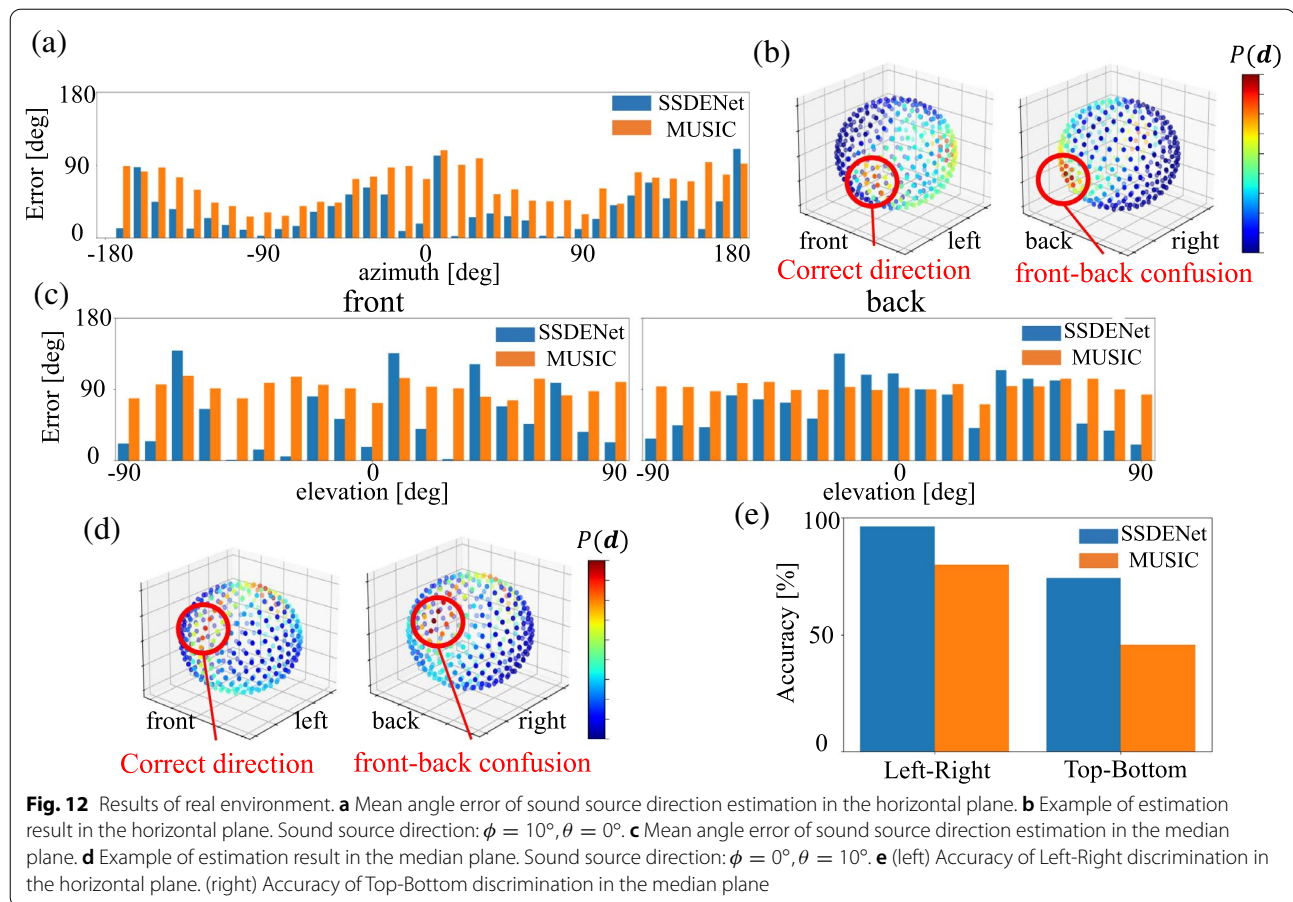
Real environment

We investigated the accuracy of the sound source direction estimation in the real environment using the dummy head shown in Fig. 7a. This experiment is conducted in a room with the noise level of 58 dBSPL and the reverberation time RT_{60} of 270~320 msec.

A loudspeaker is placed 1.5 m in front of the dummy head to play white noise of around 73~76 dBSPL. We investigate the following three points in this experiment,

- Sound source direction estimation error in the horizontal plane
- Sound source direction estimation error in the median plane
- Discrimination between left and right in the horizontal plane, and between top and bottom in the median plane.

The results of this experiment are shown in Fig. 12e. As shown in Fig. 12a, the error of the proposed method is smaller than those of the MUSIC method in many directions of the horizontal plane. The results of sound source existences for the direction where the error of the proposed method is large are shown in Fig. 12b. Although

Fig. 11 Results of simulation. Mean angle error for each condition

Fig. 12 Results of real environment. **a** Mean angle error of sound source direction estimation in the horizontal plane. **b** Example of estimation result in the horizontal plane. Sound source direction: $\phi = 10°, \theta = 0°$. **c** Mean angle error of sound source direction estimation in the median plane. **d** Example of estimation result in the median plane. Sound source direction: $\phi = 0°, \theta = 10°$. **e** (left) Accuracy of Left-Right discrimination in the horizontal plane. (right) Accuracy of Top-Bottom discrimination in the median plane

there is a peak in the correct sound source direction, there is also a large peak in the front-back symmetry direction.

In the median plane, the errors of the MUSIC method are around 90° in any direction, which shows that the MUSIC method is not capable of sound source direction estimation in the median plane. On the other hand, the proposed method can estimate sound source direction with a sufficiently small error depending on the direction. The results of sound source existences for the direction where the error of the proposed method is large are shown in Fig. 12d. As in the case of the horizontal plane, there is a peak in the correct sound source direction, but there is also a large peak in the front-back symmetry direction.

The accuracies of left-right discrimination in the horizontal plane and top-bottom discrimination in the median plane are shown in Fig. 12e. The proposed method can estimate the left and right of sound source direction with high accuracy. In top-bottom discrimination, the accuracy of the MUSIC method is around 50 %, which means that the MUSIC method is not capable of

sound source top-bottom discrimination. On the other hand, the proposed method can discriminate top and bottom though not as good as left and right.

Task realization based on detection of environment including out-of-view

The musculoskeletal humanoid used in this experiment is Musashi [29]. We investigate that the musculoskeletal humanoid can realize tasks that need out-of-view information in the real complex environment. This experiment contains tasks that require environmental sound detection of not only types of sound but also sound directions. We use SSDENets trained in the former section. MelCNN used in this experiment is trained by six types of sound, car horn, alarm, ratchet sound, gearshift manipulating sound, key turning sound, and the human voice. The 50 samples are recorded for each label. The F score of MelCNN for these sounds is 0.87. In this experiment, we deal with the driving noise of Musashi as noise. We record the driving noise and calculate the driving spectra in advance. We set the thresholds at 10 times the driving spectra for each frequency and apply the
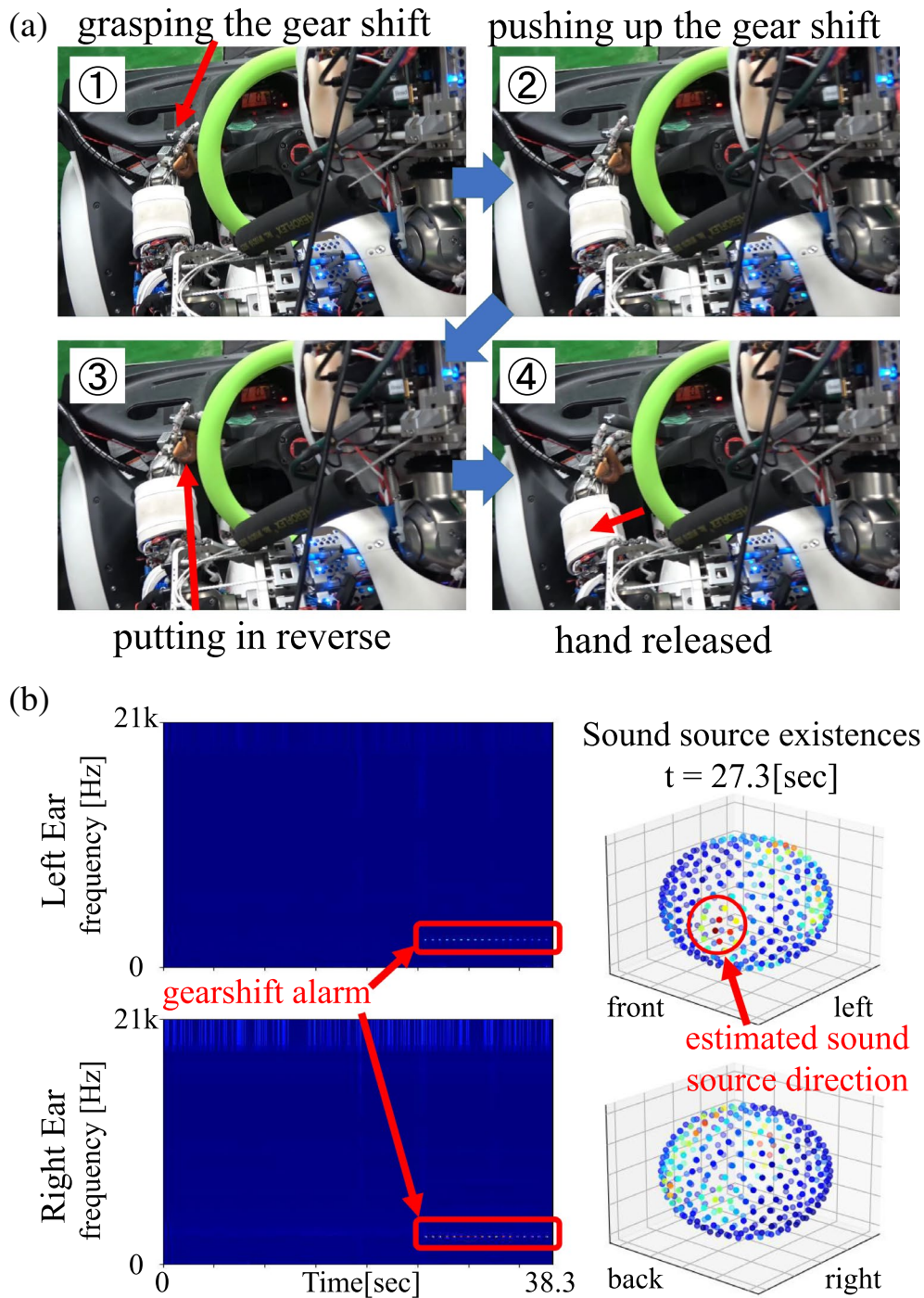
Fig. 13 a Putting the car into reverse. **b** Melspectrograms and result of alarm sound direction estimation

proposed method to spectra exceeding the thresholds. We use incoming sound and noise as training data for SSDENet and expect Musashi predicts the sound source direction in the real environment with his driving noise.

Manipulation of gearshift

Fig. 13 shows the behavior of the gearshift manipulation. Musashi needs to operate the gearshift in this task and recognize that his operation causes the alarm. The
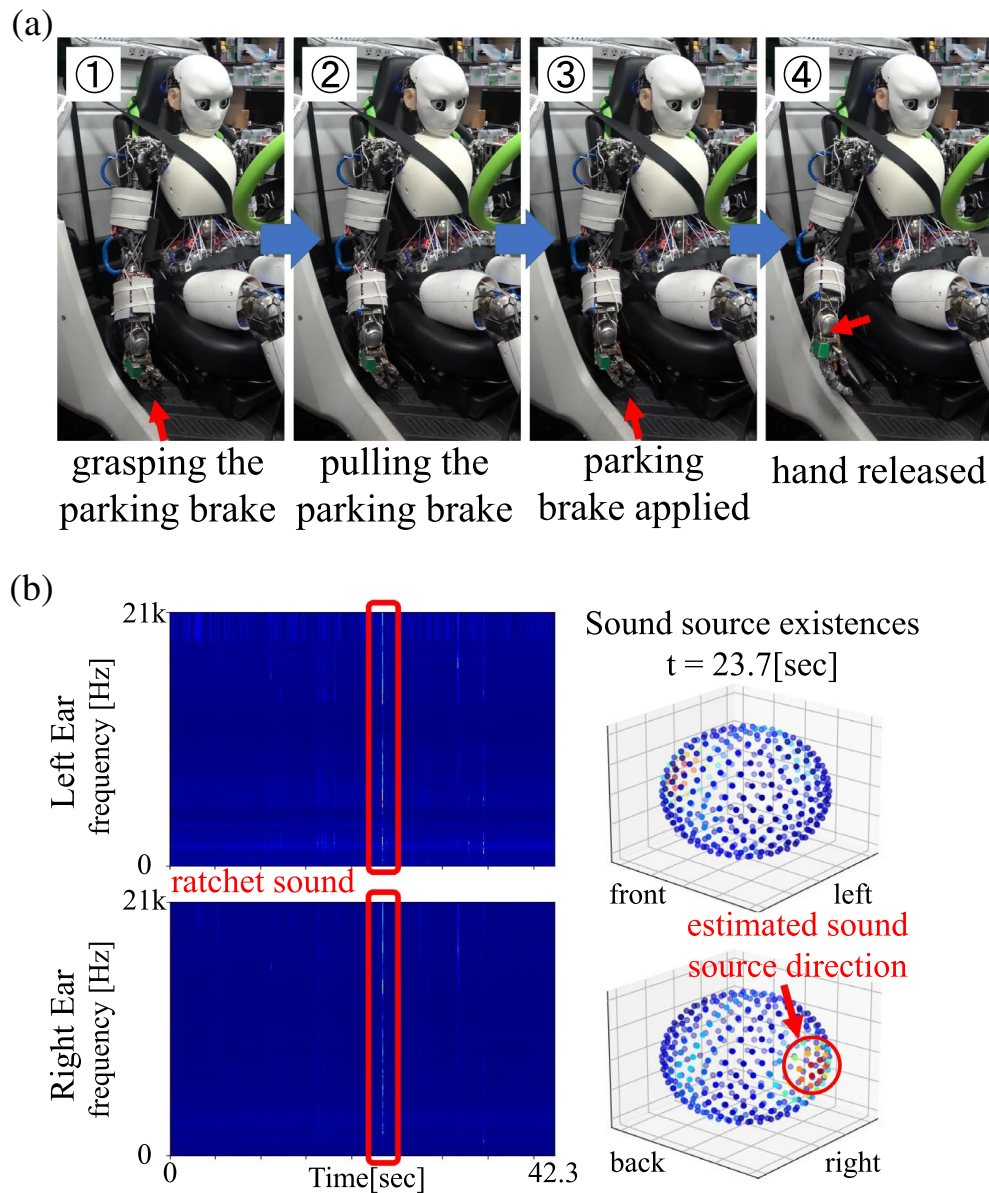
Omura *et al. ROBOMECH Journal*   (2022) 9:17

Page 12 of 15

Fig. 14 a Applying the parking brake. **b** Melspectrograms and result of ratchet sound direction estimation

alarm sound occurs in front of Musashi. Musashi moves to put the gearshift into reverse and takes the left hand away when Musashi recognizes the alarm sound from the front. At first, the gearshift is grasped with the left hand as the initial state, and Musashi starts pushing up the gearshift. At around 25 sec, the gearshift is put into reverse, and the alarm sound starts. Musashi recognizes the alarm sound direction to be in front at 27.3 sec and releases the left hand from the gearshift to complete the motion.

Manipulation of parking brake

Fig. 14 shows the behavior of the parking brake manipulation. Musashi needs to operate the parking brake in this task and recognize that his operation causes the ratchet sound. The ratchet sound occurs right side of Musashi. Musashi pulls up the parking brake and takes the right hand away when Musashi recognizes the ratchet sound from right. At first, Musashi grasps the parking brake by the right hand as the initial state and starts pulling up the parking brake. At around 23.5 sec, the parking brake is
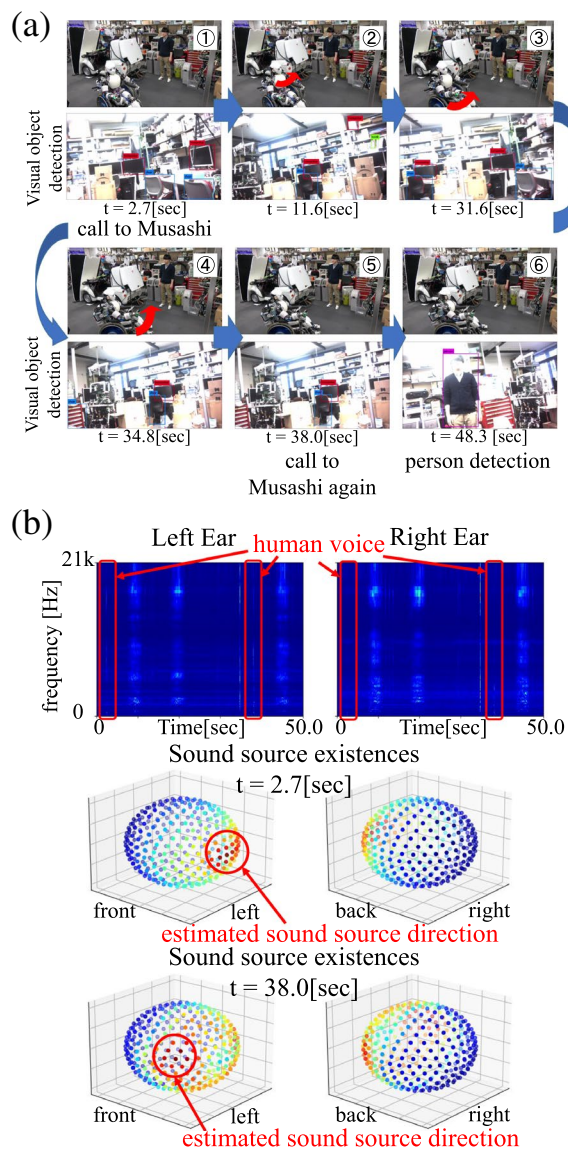
Omura *et al. ROBOMECH Journal* (2022) 9:17

Page 13 of 15

Fig. 15 a Musashi turns in the direction of the voice. **b** Melspectrograms and results of voice direction estimation

applied, and the ratchet sound is heard. Musashi recognizes the ratchet sound direction to be in the right at 23.7 sec and releases the right hand from the parking brake to complete the motion.

Response to calls

Fig. 15 shows the behavior of the response to call. In order to turn to the direction of call, Musashi moves the neck and eyes and manipulates the wheelchair to turn the body around, and Musashi captures the person who calls Musashi in the view. The visual detection of humans is performed for the right eye view,

and [30] is used for the detection. Initially, at around 2 sec, a person calls from the left side of Musashi, and Musashi detects the voice and recognizes that the call comes from the left side. After that, Musashi moves the neck and eyes at around 11 sec to check the left side, but Musashi cannot catch the person in the field of view, and then Musashi turns the body to the left side by manipulating the wheelchair for around 30 sec. The person calls again at around 38 sec, and Musashi recognizes that the voice came from the left front. Musashi moves the neck and eyes to check the left side again, and Musashi captures the person in its field of view and finishes the behavior of the response to call at around 48 sec.

Discussion

We discuss the results obtained from the experiments of this study. First, we describe the effects of the human mimetic outer ear structure. The human-like structures of the pinna and ear canal produce changes of frequency response in the median plane depending on the elevation and complex changes of HRTF. The frequency response changes in the median plane due to the pinna shape show effect similar to the pinna notch in humans. The pinna notches are seen from about 5000 Hz in the lower frequency range and are clearly visible in the higher frequency range, which corresponds to the previous investigation [21]. Also, in terms of the effect on the HRTF in the frequency domain, at lower frequencies, ILD changes little, and the effect by ITD (IPD) is large, while the effect by ILD increases and ITD (IPD) has complex changes at higher frequencies, which supports "Duplex Theory" [31]. The human mimetic binaural ear unit developed in this study reproduces the characteristics of the human outer ear.

Second, we describe the sound source direction estimation using SSDENet. In the simulation with multiple types of sound sources, the estimation errors are smaller than those of the MUSIC method. In the existing sound source direction estimation methods such as the MUSIC method, each frequency and direction is calculated independently. On the other hand, SSDENet has a structure that includes the relationship between directions to be estimated in the proposed method so that it is expected for the false estimation results to be suppressed. Also, SSDENet works robustly in the real environment because it uses data with background noise components in training. Both the MUSIC method and the proposed method using SSDENet show higher accuracy for sounds with more frequency components such as square waves, sawtooth waves, and white noise than for sounds with fewer frequency components such as sine waves and triangle

waves. The reason is that the effect of the phase diversity of IPD becomes smaller as the number of frequency components increases. In the real environment, the accuracy of the proposed method is higher than that of the MUSIC method in the horizontal plane. Although the errors in the median plane are not as large as that of the MUSIC plane in any direction, the estimation errors are larger than those of the horizontal plane, and it can be said that it is only for reference in practical use. Also, there are front-back confusions in both the horizontal and median planes. The accuracy of human sound source direction estimation is lower for top-bottom and front-back directions than for left-right directions. It can be said that the proposed method shows a similar trend of accuracy in sound source direction estimation as humans. However, humans improve their sound source localization accuracy by rotating and tilting their heads. In order to realize human-like sound source direction estimation, recognition combined with motion should be addressed in future work.

Finally, we describe task realization based on environment recognition, including out-of-view. We realize the acquisition of environment information and task realization based on human-like auditory information processing for each action. The high accuracy of the proposed system enables auditory recognition even in a cluttered real environment with background noise.

Conclusion

In this research, we proposed a human mimetic auditory environmental recognition system consisting of a human mimetic binaural ear, sound source direction estimation system, and environmental sound detection system. The developed human mimetic binaural ear unit, which consists of the human mimetic outer ear structure, the microphone board that mimics human hearing characteristics, and the acoustic processing board that performs frequency decomposition with low latency, shows complex frequency response depending on the sound source direction like the human ear. The proposed sound source direction estimation method mimics detection of ILD and ITD at SOC and response to each direction at IC in human auditory information processing. In contrast to the existing engineering method, the proposed method uses neural networks that include the relationship between directions and shows that it is possible to estimate the sound source direction using only two ears roughly. By implementing the proposed system, we realized for the musculoskeletal humanoid to recognize the environment, including out-of-view areas, and perform some tasks which require recognition of objects out of view. In future works, we will work on the human mimetic wide-range environment recognition and task realization by integrating the proposed method with sensing modes, such as visual and tactile senses.

Author contributions
YO and KK proposed the concept of the human mimetic auditory information processing system. YO and YN developed the human mimetic ear unit. YO, KK, YK, MN and YT supported experiments of this research. YA, KO, KK and MI supported the whole development of this research. All authors read and approved this manuscript.

Declarations

Author details
[1]Department of Information Science and Technology, The University of Tokyo, Tokyo, Japan. [2]Toyota Motor Corporation, Toyota, Japan.

References
1. Nakanishi Y, Ohta S, Shirai T, Asano Y, Kozuki T, Kakehashi Y, Mizoguchi H, Kurotobi T, Motegi Y, Sasabuchi K, Urata J, Okada K, Mizuuchi I, Inaba M (2013) Design approach of biologically-inspired musculoskeletal humanoids. Int J Adv Rob Syst 10(4):216–228
2. Wittmeier S, Alessandro C, Bascarevic N, Dalamagkidis K, Devereux D, Diamond A, Jäntsch M, Jovanovic K, Knight R, Marques HG, Milosavljevic P, Mitra B, Svetozarevic B, Potkonjak V, Pfeifer R, Knoll A, Holland O (2013) Toward anthropomimetic robotics: development, simulation, and control of a musculoskeletal torso. Artif Life 19(1):171–193
3. Jäntsch M, Wittmeier S, Dalamagkidis K, Panos A, Volkart F, Knoll A (2013) Anthrob—a Printed anthropomimetic robot. In: Proceedings of the 2013 IEEE-RAS international conference on humanoid robots, pp. 342–347
4. Asano Y, Kozuki T, Ookubo S, Kawamura M, Nakashima S, Katayama T, Iori Y, Toshinori H, Kawaharazuka K, Makino S, Kakiuchi Y, Okada K, Inaba M (2016) Human mimetic musculoskeletal humanoid kengoro toward real world physically interactive actions. In: Proceedings of the 2016 IEEE-RAS international conference on humanoid robots, pp 876–883
5. Makabe T, Kawaharazuka K, Tsuzuki K, Wada K, Makino S, Kawamura M, Fujii A, Onitsuka M, Asano Y, Okada K, Kawasaki K, Inaba M (2018) Development of movable binocular high-resolution eye-camera unit for humanoid and the evaluation of looking around fixation control and object recognition. In: Proceedings of the 2018 IEEE-RAS international conference on humanoid robots, pp 840–845
6. Makino S, Kawaharazuka K, Kawamura M, Fujii A, Makabe T, Onitsuka M, Asano Y, Okada K, Kawasaki K, Inaba M (2018) Five-fingered hand with wide range of thumb using combination of machined springs and variable stiffness joints. In: Proceedings of the 2019 IEEE/RSJ international conference on intelligent robots and systems, pp 4562–4567
7. Shinjo K, Kawaharazuka K, Asano Y, Nakashima S, Makino S, Onitsuka M, Tsuzuki K, Okada K, Kawasaki K, Inaba M (2019) Foot with a core-shell structural six-axis force sensor for pedal depressing and recovering from foot slipping during pedal pushing toward autonomous driving by humanoids (in press). In: Proceedings of the 2019 IEEE/RSJ international conference on intelligent robots and systems

8. Schmidt R (1986) Multiple emitter location and signal parameter estimation. IEEE Trans Antennas Propag 34(3):276–280. https://doi.org/10.1109/TAP.1986.1143830
9. Knapp C, Carter G (1976) The generalized correlation method for estimation of time delay. IEEE Trans Acoust Speech Signal Process 24(4):320–327
10. Nakadai K, Lourens T, Okuno HG, Kitano H (2000) Active audition for humanoid, pp 832–839
11. Nakadai K, Matsuura D, Okuno HG, Tsujino H (2004) Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots. Speech Commun 44(1–4):97–112
12. Kurotaki S, Suzuki N, Nakadai K, Okuno HG, Amano H (2005) Implementation of active direction-pass filter on dynamically reconfigurable processor . IEEE, pp 3175–3180
13. Nakashima H, Mukai T (2005) 3d sound source localization system based on learning of binaural hearing. In: 2005 IEEE international conference on systems, man and cybernetics, vol 4. IEEE, pp 3534–3539
14. Kim U, Nakadai K, Okuno HG (2015) Improved sound source localization in horizontal plane for binaural robot audition. Appl Intell 42(1):63–74
15. Yamamoto S, Nakadai K, Valin J-M, Rouat J, Michaud F, Komatani K, Ogata T, Okuno HG (2005) Making a robot recognize three simultaneous sentences in real-time. In: 2005 IEEE/RSJ International conference on intelligent robots and systems. IEEE, pp 4040–4045
16. Heckmann M, Rodemann T, Joublin F, Goerick C, Scholling B (2006) Auditory inspired binaural robust sound source localization in echoic and noisy environments. IEEE, pp 368–373
17. Youssef K, Argentieri S, Zarader J-L (2013) A learning-based approach to robust binaural sound localization. IEEE, pp 2927–2932
18. Dávila-Chacón J, Liu J, Wermter S (2018) Enhanced robot speech recognition using biomimetic binaural sound source localization. IEEE Trans Neural Netw Learn Syst 30(1):138–150
19. Tokozume Y, Harada T (2017) Learning environmental sounds with end-to-end convolutional neural network. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 2721–2725
20. Iqbal T, Kong Q, Plumbley M, Wang W (2018) Stacked convolutional neural networks for general-purpose audio tagging. DCASE2018 Challenge
21. Musicant AD, Butler RA (1984) The influence of pinnae-based spectral cues on sound localization. J Acoust Soc Am 75(4):1195–1200
22. Gardner MB, Gardner RS (1973) Problem of localization in the median plane: effect of pinnae cavity occlusion. J Acoust Soc Am 53(2):400–408
23. Toshima I, Aoki S (2009) Possibility of simplifying head shape with the effect of head movement for an acoustical telepresence robot: Telehead. IEEE, pp 193–198
24. Grothe B, Pecka M, McAlpine D (2010) Mechanisms of sound localization in mammals. Physiol Rev 90(3):983–1012
25. Semple MN, Aitkin LM, Calford MB, Pettigrew JD, Phillips DP (1983) Spatial receptive fields in the cat inferior colliculus. Hear Res 10(2):203–215
26. Palmer A, King A (1982) The representation of auditory space in the mammalian superior colliculus. Nature 299(5880):248–249
27. Piczak KJ (2015) Environmental sound classification with convolutional neural networks. In: 2015 IEEE 25th international workshop on machine learning for signal processing (MLSP). IEEE, pp 1–6
28. Mohan S, Lockwood ME, Kramer ML, Jones DL (2008) Localization of multiple acoustic sources with small arrays using a coherence test. J Acoust Soc Am 123(4):2136–2147
29. Kawaharazuka K, Makino S, Tsuzuki K, Onitsuka M, Nagamatsu Y, Shinjo K, Makabe T, Asano Y, Okada K, Kawasaki K, Inaba M (2019) Component modularized design of musculoskeletal humanoid platform musashi to investigate learning control systems. In: Proceedings of 2019 IEEE/RSJ international conference on intelligent robots and systems, pp 7300–7307
30. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv
31. Rayleigh L (1907) On our perception of sound direction. Lond Edinb Dublin Philos Mag J Sci 13(74):214–232

Publisher's Note