# Intuitive Environmental Perception Assistance for Blind Amputees Using Spatial Audio Rendering

Xuhui Hu, *Student Member IEEE*, Aiguo Song, *Senior Member IEEE*, Hong Zeng, *Member, IEEE*, and Dapeng Chen, *Member, IEEE*

***Abstract*—Vision and touch are essential sensory systems for human to interact with the environment. For the blind amputees, how to quickly and intuitively convey the environmental information to them is one of the key issues for recovering their daily living ability. Inspired by the auditory localization ability of human, we constructed a virtual scene almost identical to reality, and concurrently added a virtual sound source to the interactive object. Leveraging the method of spatial audio rendering (SAR), the three-dimensional motion of the virtual sound source can be vividly simulated in real-time. Finally, a myoelectric prosthetic control system was developed to assist blind amputees in their daily activities, The Fitts' law test on target localization was conducted on both SAR and voice prompt (VP) based path guidance methods, the results indicate that SAR significantly improves the information transfer rate. The results of prosthetic control test show that SAR reduces the completion time by half than the VP, while restoring the natural grasping path. With the advantage of intuitive and rich perception, the SAR demonstrated the potential applications for blind amputees to reconstruct the control and sensory loops.**

***Index Terms*—Human-robot interaction, Prosthetics, Sensory Feedback, Spatial Audio.**

## I. Introduction

THE EYES and hands of humans are indispensable sensory systems for learning, adapting, and transforming the environment. For the upper-limb amputees, their missing motor function can be partially restored by wearing a myoelectric prosthetic hand. However, most of the commercial products neglected the haptic function of the robotic hand because users can rely on their vision as haptic substitution. Therefore, once the loss of both haptic and visual perception, it will bring great challenges to their daily life.

Manuscript received April 8, 2021; revised July 16, 2021 and October 21, 2021; accepted January 23, 2022. Date of publication January 26, 2022; date of current version February 22, 2022. This article was recommended for publication by Associate Editor A. Forner and Editor P. Dario upon evaluation of the reviewers' comments. This work was supported in part by the National Natural Science Foundation of China under Grant 91648206, Grant 61673105, and Grant 62003169; and in part by the Jiangsu Key Research and Development Plan under Grant BE2018004-4. *(Corresponding author: Aiguo Song.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the ethical committee of Southeast University, and performed in line with the Declaration of Helsinki.

Xuhui Hu, Aiguo Song, and Hong Zeng are with the State Key Laboratory of Bioelectronics and Jiangsu Key Laboratory of Remote Measurement and Control, School of Instrument Science and Engineering, Southeast University, Nanjing 210096, Jiangsu, China (e-mail: a.g.song@seu.edu.cn).

Dapeng Chen is with the School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, Jiangsu, China.

This article has supplementary downloadable material available at https://doi.org/10.1109/TMRB.2022.3146743, provided by the authors.

Digital Object Identifier 10.1109/TMRB.2022.3146743

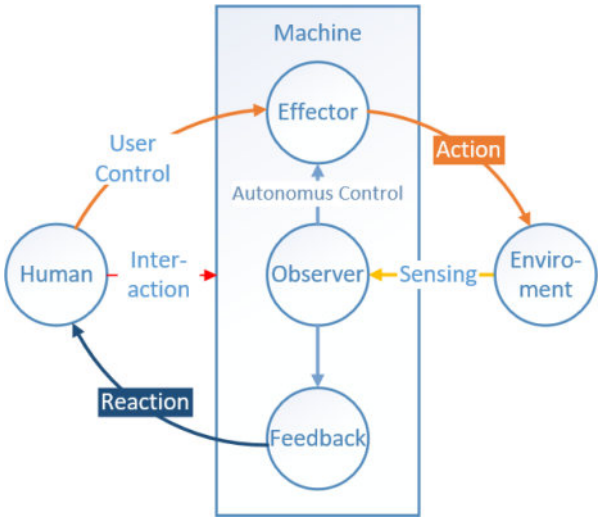


Fig. 1. The framework of environmental perception assisting system.

There is a high incidence of blindness and hand amputation in the military or high risk occupations [1], Krukenberg procedure [2] was an alternative technique to surgically separate the radius and ulna of the distal forearm, and the patients can control the movement of radius and ulna like a pincer. Meanwhile, the skin wrapping around two bones realizes the tactile perception to the environment. Decades of efforts have brought about great progress on intelligent bionic prosthetic hand: in order to build the visual sensation substitution, the method of multimodal information fusion based on computer vision, inertial [3] and eyes tracking [4] was proposed to assist the manipulating of prosthetic hand. With the rich perception of RGBD camera and the excellent recognition of deep learning, the method of computer vision improved the grasping accuracy in the daily life scenarios [5], [6]. However, these studies mainly aimed at improving the autonomous dexterity of the machine, lacking the environmental feedback to the user.

An environmental perception assisting system for blind amputees is shown in Fig. 1, where the human can interact with the machine or directly control it to perform actions on the environment. Meanwhile, the environmental information perceived by the machine is not only used for autonomous control of the system, but also transmitted to the human. The machine contains three sub-components: 1) an observer unit for environmental sensing, including a camera module

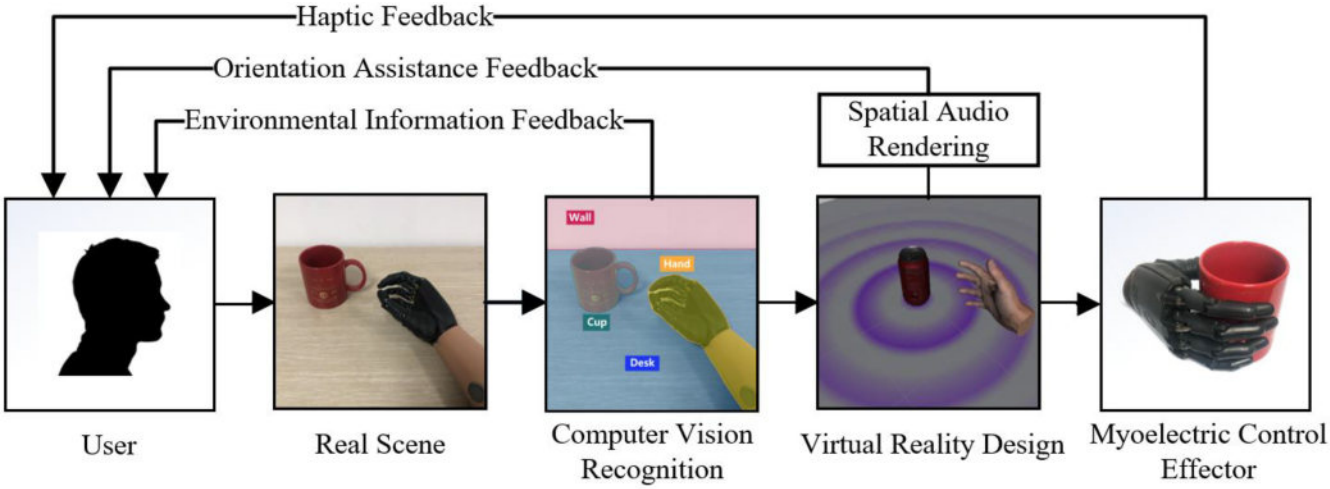


Fig. 2. A typical workflow diagram of the environmental perception assisting system.

(as visual substitution) and a haptic sensory module (integrated within the prosthetic hand); 2) an effector unit for direct action to the environment (i.e., myoelectric controlled prosthetic hand); 3) a feedback unit that transmits the interactive reactions to the human.

In order to rebuild the closed environmental perception loop, researchers used Google Glass as the observer, and the methods of visual augmented-reality (AR), voice prompt and vibrotactile encoding was used to transmit the sensory information from computer vision to the user [7]. Nevertheless, the visual AR is not effective for the blindness, and the other feedback strategies transform the visual information into a sequence of vibration or voice, which greatly shrinks the information transfer rate (ITR) and requires the user to learn the preset feedback modalities.

In purpose of achieving much faster and intuitive sensory feedback, recent studies have attempted to encode the machine information into sequences of electrical pulses that directly stimulate the sensory nervous system in an invasive or noninvasive form, which can generate multiple phantom sensation such as pressure, tingling and vibration [8]–[11]. However, these researches mainly focused on producing the haptic feedback of the robotic hand. Assuming a blind amputee is using his prosthetic hand to drink a cup of tea, so the first stage is to approach the cup, which is of great challenge without the visual assistance. In general, the ITR of voice instruction is insufficient to convey the accurate orientation information, an alternative method is using wearable devices to generate force/tactile stimuli for orientation guidance [12]–[14]. However, the disabled may not easy to wear too many complicated devices on their own. More importantly, this method creates an indirect sensory surrogate, which requires a long period of training to build sensory mapping.

In addition to the visual localization, our auditory system can also help to determine the location of the target. As can be imagined, if a 3D sound could be rendered from the position of the cup, he/she could easily locate the cup with his/her auditory perception to the sound source. Inspired by this idea, we designed a virtual scene that rebuilt the spatial information of the real environment. For the interactive object in the reality (e.g., the cup), a sound source was added into the corresponding virtual object. We used a method called spatial audio rendering (SAR) which generated the 3D surround sound according to the virtual scene, giving the user an illusion that the sound is coming from the object. As a result, these virtual auditory cues functioned as an intuitive orientation guidance to help the user to find the real object. The proposed spatial audio based sensory surrogate offers a novel pathway that may help the blind amputees to regain their natural grasping. Comparing to the same auditory based voice prompt (VP) method, SAR has greatly improved the information transfer rate and grasping efficiency. More significantly, SAR empowers the user with active perception, allowing a greater control robustness.

## II. Methods

### A. System Framework

The proposed environmental perception assisting system, as shown in Fig. 1, centers on the interaction between human, machine, and environment. Specifically, for the assistance of blind amputees, a typical system workflow diagram is shown in Fig. 2. At the very beginning, the speech instruction is issued by the user to initiate the computer vision recognition from the real scene captured with the camera. Then the recognized environmental information, such as the type (e.g., indoor) and content (e.g., cup, hand) of the real scene, will be told to the user by speech interface module. The user can further specify the task (e.g., "fetching the cup") based on the results of the voice prompts, and then the system will execute the corresponding task according to the user's requirement, here we mainly focus on the object localization task. Given the proposed real-time spatial audio rendering techniques rely on a virtual environment, the real environmental information of the (i.e., the spatial coordinates of the object and the prosthetic hand) is used to build a virtual scene, and then triggering SAR to provide the orientation cues. As the object approaching task is finished, the user can control the prosthetic hand to grab the

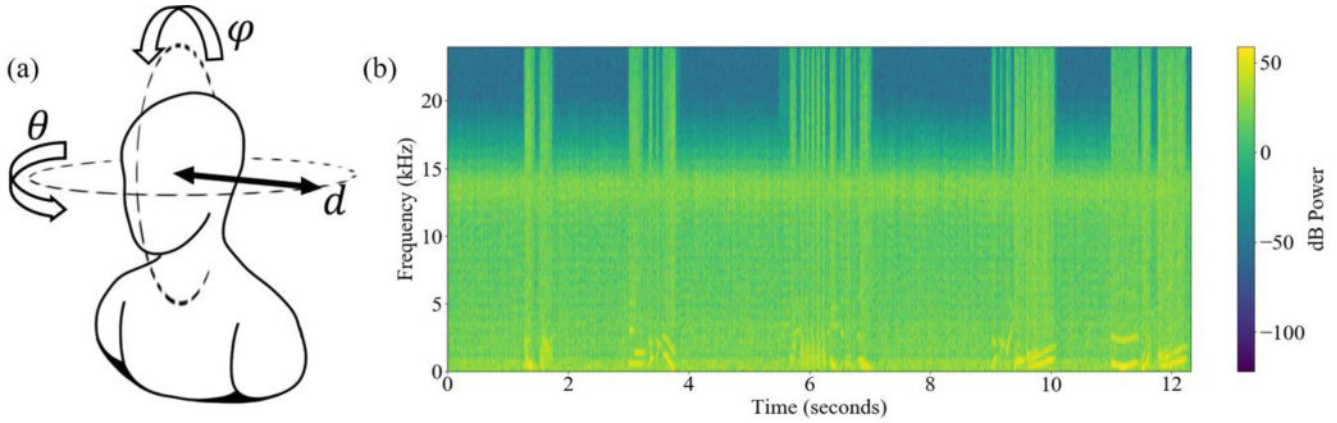


Fig. 3. (a) Definition of the three-dimensional information ($d$, $\theta$, $\varphi$) of the sound source (b) The spectrogram of looping spatial audio guidance cues.

object. At this time, the haptic information of the robotic hand will give feedback to the user. Finally, the proposed system realizes the complete closed loop from active motor control to environmental perception.

Since the core of this paper is the object reaching and grasping, the details of "Computer Vision Recognition" in Fig. 2 would not discussed any further, the remaining subsections were organized as follows: Firstly, the principle of human auditory localization based on SAR was introduced in Section II-B. Secondly, through the study of human orientation perception to the spatial audio, the virtual reality scene was designed in Section II-C. Finally, the myoelectric control effector was introduced in Section II-D to verify the feasibility of the proposed method for daily living.

### *B. The Principle of Auditory Localization by SAR*

Our auditory system has inborn spatial perception ability, even when the vision was affected by occlusion or surrounding light, through listening with our ears and adjusting the listening direction, we can also distinguish the three-dimensional location of the sounding object: the distance ($d$), the azimuth angle ($\theta$), and the elevation angle ($\varphi$) [15], as is shown in Fig. 3(a).

The reason for human to be able to distinguish the azimuth angle is that our brain can feel the subtle interaural phase differences by timing and intensity cues [16], which refers to Interaural Time Difference (ITD) and Interaural Intensity Difference (IID) [17].

$$ITD = \frac{r}{c}(sin\theta + \theta)\ f < 1500\text{Hz} \tag{1}$$

$$IID = 20\log\frac{|A_L(s,\theta)|}{|A_R(s,\theta)|}\ f > 1500\text{Hz} \tag{2}$$

$$A_L(s,\theta) = \frac{\left(1+\cos\left(\theta+\frac{\pi}{2}\right)\right)s + 2c/r}{s + 2c/r} \tag{3}$$

$$A_R(s,\theta) = \frac{\left(1+\cos\left(\theta-\frac{\pi}{2}\right)\right)s + 2c/r}{s + 2c/r} \tag{4}$$

where $r$ is the head radius, $c$ is the acoustic velocity, $A_L(s,\theta)$ and $A_R(s,\theta)$ are a pair of transfer function that specify, for a given azimuth angle $\theta$ and sound frequency $f$, the attenuation of the sound by head shadowing. We assumed that $\theta$ equals to zero degree when the sound source is directly in front of the head. When the sound frequency is up to approximately 1500Hz, the wavelength of the sound becomes comparable to the diameter of the head, and ITD cues for azimuth angle become ambiguous. Meanwhile, the magnitude of high-frequency sounds are easily attenuated by the head shadowing, resulting in a gradual intensity difference between two ears. Therefore, IID is used to calculate the interaural phase difference when the sound frequency is above 1500Hz.

The perception to the elevation angle are mainly derived from the monaural cues, which is well explained by the pinna filtering effect theory [18]. The shape of human pinna (external ear) is very special, which makes sounds coming from different directions bounce off in different ways. As a result, the reflected waves reaching the eardrum will generate an orientation-related frequency spectrum, which can be processed by auditory nerves. These spectrum clues generated by pinna filtering effect can be presented as a Head-Related Transfer Function (HRTF), which are commonly specified as a minimum-phase FIR filter [19]. In order to synthesis the directional sound from single channel audio, the sounds received by two ears were constructed separately.

$$s_l(t) = s(t - ITD) * h_l(\theta, \varphi, t) \tag{5}$$

$$s_r(t) = s(t) * h_r(\theta, \varphi, t) \tag{6}$$

where $s(t)$ is the time-domain single channel audio, $s_l(t)$ and $s_r(t)$ are the synthesized sound that output from headphones, $h_l$ and $h_r$ are the minimum phase impulse responses measured at azimuth $\theta$ and elevation $\varphi$. The minimum phase assumption allows to uniquely specify an HRTF's phase alone, and to separate ITD from the FIR specification of HRTF [20]. Here, ITD is defined to be negative for sounds arriving at the left ear first. HRTF can be obtained from measuring the head-related impulse response $h(t)$, where an impulse sound source $\Delta(t)$ was placed somewhere and the response signal was recorded by "Dummy head" simultaneously. Finally, The Fourier transform of $h(t)$ gives the equation of HRTF. Here we used the publicly available HRTF from open source databases [21].

Human's perception to the distance of a sound source is affected by multiple factors. Since we only use sound

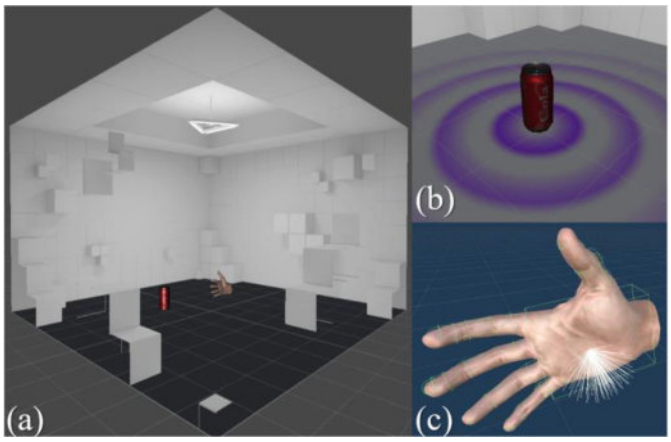


Fig. 4. The Virtual Reality Design (a) Audio Room (b) The target object with virtual sound source (c) The virtual hand with distance and force sensing zone.

as a guidance for spatial localization, only a static single sound source in the close-up range was discussed. On this premise, the distance perception ability is mainly affected by: 1) The loudness and the spectrum of the sound, which have already become common senses. 2) The sound reflection in the environment, specifically the ratio between the direct sound (arrives straight to the ears) and the reflected sound (arrives after the reflections), which will give an indication about the distance of the sound source. If there is no reflected sound, it is confusing whether the change in the loudness of the sound is caused by the distance or its power [22]. As a result, on the one hand, we need to select the appropriate volume and spectrum to the sound source. On the other hand, a well-designed virtual environment is essential, because the size and material of the scene may influence the optimal distance perception of the user.

In this paper, a looping audio clip was added into the target object, the spectrogram is shown in Fig. 3(b). If the frequency composition of a sound is too simple, people will feel harsh and tired. Here, a sound cue with both high-frequency components (from 12 kHz to 15 kHz) and low-frequency components (from 200 Hz to 3 kHz) was used. The maximum level of the sound is around 50 dB, mainly concentrated in the low-frequency band, which is the most sensitive frequency band of the human ear. Here we used "Resonance Audio" (Google Inc.) as the audio rendering engine, which can render full-sphere surround sound in real-time by Ambisonics format [23]. Furthermore, it supports HRTF-based surround sound decoding, geometry-based reverb rendering, as well as the directivity customization of sound source.

### C. Virtual Reality Design

The virtual scene mentioned above was designed for carrying the SAR, which is shown in Fig. 4. All the interactions took place in an audio room (Fig. 4(a)), in which a target object (Fig. 4(b)) and a virtual hand (Fig. 4(c)) were placed. The constructed virtual scene was consistent with the real environment information. For the interactive object in the reality, a sound source was added into corresponding virtual object, producing the 3D surround sound through real-time sound rendering. The directivity pattern of the sound source was designed into a circular shape, which is represented by purple circular effect in Fig. 3(b), so that the sound effects we hear from all directions are consistent. In order to simplify the complexity of the experiment, the position of the target and the hand was limited to a two-dimensional horizontal plane.

Based on our common sense to the sound localization, we usually adjust the orientation of our ears (i.e., the listener) by rotating the head to better determine the direction sound source, sometime it is also necessary to move the body to get closer when at a distance. However, for the desktop object reaching tasks, unnecessary effort is expended. Therefore, we set the user controller (i.e., virtual hand) as the listener, whose orientation and position can be controlled by easily moving the hand, contributing to a better position perception.

As mentioned in Section II-B, people cannot effectively perceive the distance when there are only a few reflected sounds. Therefore, we used an enclosed cubic scene, as is shown in Fig. 4(a) to enhance sound reflection. This virtual scene was built on the Unity platform (Version. 2019.4.15f1c1), and ran on Windows 10 operation system (CPU: Intel Core i5-9400F, RAM: 8GB).

### D. Myoelectric Control Effector

Here, the myoelectric control effector was developed to evaluate the performance of SAR by integrating the force/tactile and proximity perception. We used a fixed Web camera (C930e, Logitech), which has 90 degrees of diagonal field of view (dFoV), hovering above the desk to capture the two-dimensional horizontal coordinates of target and hand. The visual field is about $58\times118$ cm, and the resolution of the video stream is $1280\times720$ pixels, with 30 fps of frame rate. The target object was placed under the visual field, with a positioning tag (Tag36h11, Apriltag) affixed on its top. The subject put on a pair of wireless BlueTooth earbuds (AirPods Pro, Apple), which supports the 5.1 surround sound [24]. Meanwhile, a bebionic hand (Ottobock) was fixed on the stump of the user, which includes another positioning tag on the prosthesis socket to locate the position of the hand. Finally, the pixel coordinates of two positioning tags in the live video stream were transformed into the horizontal coordinates of the target object and the hand in the virtual scene respectively. In order to overcome the image distortion caused by the wide-angle camera, the perspective correction was made to unified the coordinates of the two tags with the real coordinate system

In order to control the prosthetic hand, two surface electromyography (sEMG) sensors were closely attached to a pair of antagonistic muscles on the radial and ulnar sides of the right upper-limb, the state of the art myoelectric control strategy (SOA) [25], [26] was utilized to open or close the robotic hand. A piezo-resistive film force sensor (FSR400, Interlink Electronics) was installed on the thumb of the prosthetic hand to detect the force/tactile signal when grasping an object. A laser-ranging sensor was set in the palm of the robotic hand

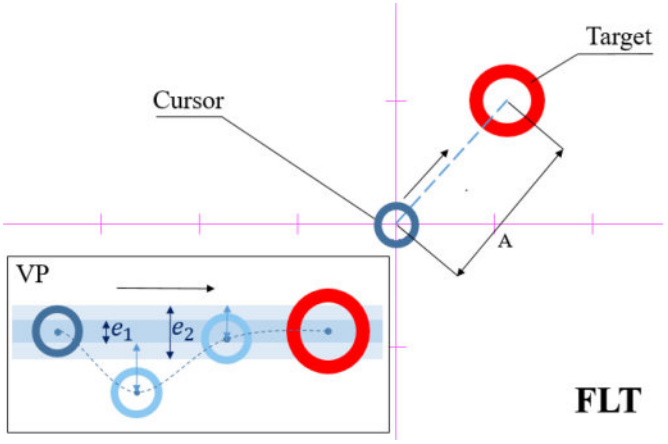


Fig. 5. Fitt's law test design.

to provide a more precise distance between the object and the end of the prosthetic hand than visual positioning. When the sensing threshold is reached, the system would prompt the completion of the reaching task. The environmental perception and interaction of this system still included two sequential steps: 1) Moving the arm to the position to be grasped by following the spatial audio cues. 2) Controlling the robotic hand to open or close through their sEMG signals. When each step was finished, the corresponding sound effects functioned as status prompts to feedback the step results to the subject.

## III. Experiments

In this section, a virtual experiment based on Fitt's law test (FLT) was firstly conducted to quantify the environmental perception performance of SAR. After that, a prosthetic control test (PCT) in the realistic environment was conducted to verify the feasibility of the proposed method. The content of PCT reflects one of the most basic actions in people's daily life, that is reaching and grasping the target object on the table, by controlling a myoelectric prosthetic hand when blind folded. Meanwhile, we designed another Voice Prompt (VP) based feedback strategy as a control group. Informed consent form was obtained from all the participated subjects, all the experiments were approved by the Ethical Committee of the university and conformed to the Declaration of Helsinki.

### A. Fitt's Law Test

Fitt's law is a descriptive and predictive model of human movement primarily used in human–computer interaction and ergonomics. It was motivated by investigating whether human performance could be quantified using a metaphor from the field of information theory. According to Fitt's law, the act of performing a target selection task is similar to transmitting information through the human channel, and the difficulty of a target selection task could be quantified using the information metric "bits". Therefore, the rate of information transmission is able to measure the performance of human movement, which is more commonly used as "throughput".

We designed a two-dimensional Fitt's law test to quantify the performance of orientation perception, as is shown in Fig. 5. The pipeline of FLT was designed as follows: The subject was blind folded and used the mouse to move a blue circle cursor freely on a two-dimensional plane, which was displayed on a computer screen. This 2D plane was mapped into the azimuth plane (shown in Fig. 3a) of human, the sound source was set into the red circle, and the listener was set into the blue circle, the audio room is the same as in PCT. The aim of the test is to leverage the auditory cues to move the cursor to a red circle target as quickly as possible. The radius of the red target was 1.5 times that of the blue cursor. As long as the blue cursor was inside the red target and the space bar was pressed at the same time, the task was considered successful, so a certain positioning error was allowed for the blue circle to move into the red circle. On each trial, the target is randomly repositioned and the cursor always starts from the origin of the coordinate axis. Throughput (*TP*) is calculated according to the index of difficulty (*ID*) and the corresponding movement time (*MT*):

$$ID = ln\left(\frac{A}{\sqrt{2\pi e}\cdot\sigma} + 1\right) \tag{7}$$

$$TP = \frac{1}{N}\sum_{i=1}^{N}\frac{ID_i}{MT_i} \tag{8}$$

where A is the movement amplitude, represented by the initial distance between the cursor and the target. Since the target contains a movement tolerance radius, $\sigma$ represents the standard deviation of the position error between the final cursor and the target in all successful trials.

Since the coordinates where the cursor falls into a target satisfy a two-dimensional normal distribution [27], [28]. The denominator term in Equation (7) is called effective width, which expresses a condition that target width is analogous to noise, and the distribution is normal with 96% of the hits falling within the target and 4% of the hits missing the target. Equation (7) is an evolution from Shannon's Theorem 17. Therefore, the information of each test trial is expressed in the ratio of movement amplitude (A) and target width.

In Equation (8), *TP* can be estimated by taking the mean of the ratio between *ID* (in bits) and *MT* (in seconds), the subscripts indicate the index of the trial sequence. Alternatively, *TP* can be obtained by drawing a two-dimensional scatter plot. Here we used simple linear regression to fit a linear function with respect to ID and MT, thus obtaining the throughput based on its slope. Finally, *TP* can objectively reflect the information capacity (in bits/s) of the feedback strategy, independent of the information content (e.g., distance or orientation).

### B. Voice Prompts Guidance

Here we introduced another mainstream pathway guidance strategy, the voice prompts strategy (VP), to make a comparison with SAR. This strategy is based on an intuitive idea that we could use few voice cues to guide the users to follow the path. In the preliminary experiment, we found that the subjects would get tired when they received too many kinds of directional cues, and it is also difficult for them to accurately move along the non-orthogonal direction. Therefore, only four instructions was used in VP, i.e., "forward", "backward", "left"

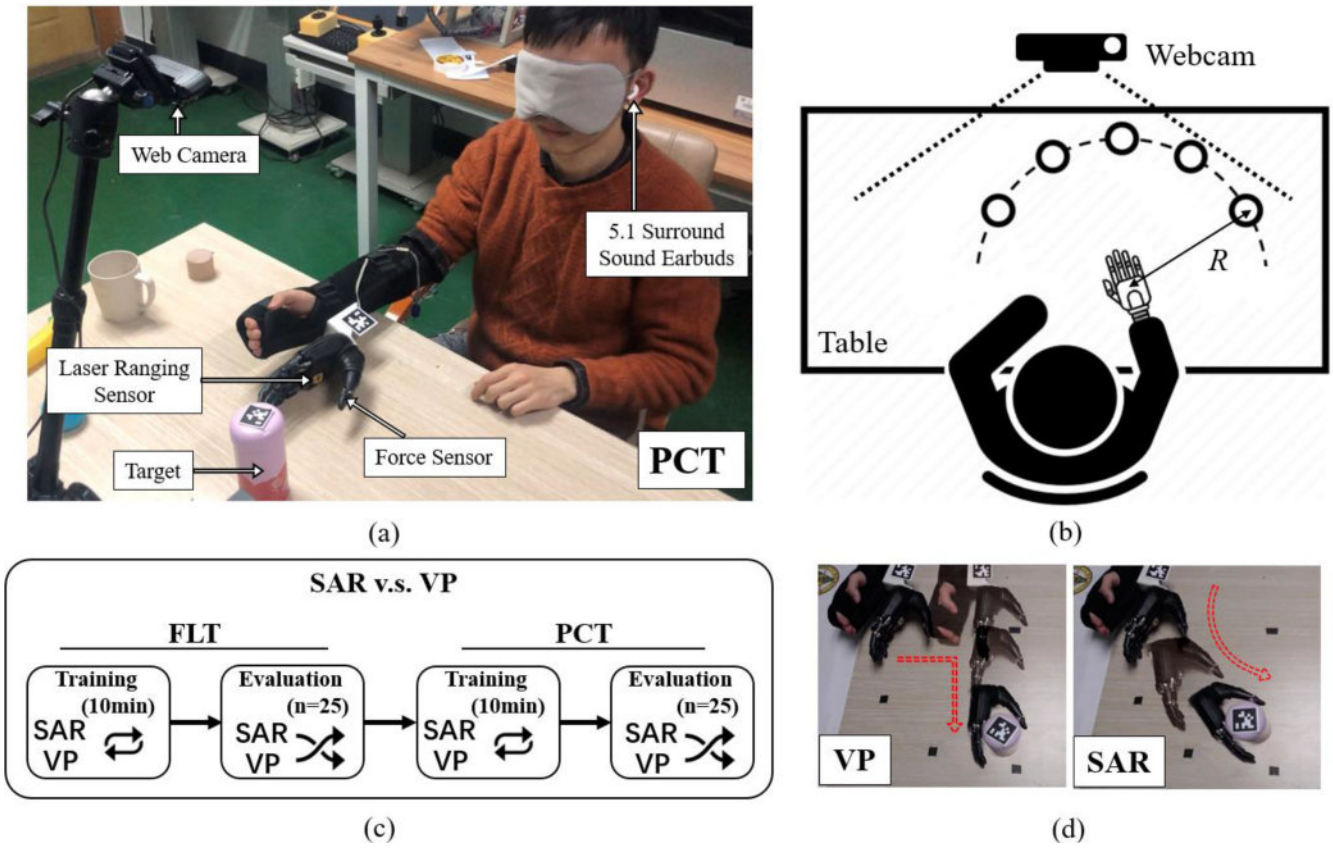


Fig. 6. (a) Prosthetic control system design. (b) Target position design in PCT. (c) Experiment session design. (d) Assistance path based on VP and SAR.

and "right", and we used three different frequencies of beeps to prompt the distance to each turning point. As a result, the reaching pathway planned by the system was an orthogonal polyline. At the same time, the voice prompt for guiding direction was played every six seconds to prevent auditory fatigue caused by high-frequency voice prompts. In the prosthetic control test, when the prosthetic hand is above 15cm away from the turning point, the frequency of beep is 0.4Hz; and 1Hz is used when between 15cm and 5cm; and 2Hz when within 5cm. The user would get a prompt sound effect when one of the steps in the sequential task was accomplished, such as arriving at the position to be grabbed or grabbing success, which was the same in SAR. In addition, VP would prompt the user when they arrived at the turning position.

In the VP strategy, since it is difficult for the blind folded subjects to accurately move their arms along the orthogonal direction, if the expected positioning error is too strict, the voice prompts will be frequently triggered during the movement, which causes fatigue and irritation to the subject. Therefore, we added an error hysteresis filter. When the system determines that the subject has deviated from the preset trajectory, the system will prompt the subject to return to the predetermined trajectory with a strict dead band e1. Once the subject is within e1, the system uses a relatively wider error band e2 for subsequent movement.

In the specific experiments, VP's speech and sound cues were not applied to spatial audio rendering (SAR), so subjects were only able to localize by distinguishing the meaning of the sound effects and understanding the speech content.

### C. Experimental Design

Sixteen able-bodied and normal hearing individuals (25±3 years old) were informed and participated in this experiment,the experimental setup is shown in Fig. 6(a). An experimenter was responsible for placing the grasping target at a location on the table (randomly generated by the system), then the subject needed to control the robotic hand to grab the cup as fast as possible. Before the experiment, the subject was blind folded and put on the earbuds, the subject's right arm was fitted with the prosthesis socket, which can constrain the motion of the wrist to simulate the isometric contraction of the residual muscle in transradial amputees. Meanwhile, none of them knew the layout of the scene in advance, nor can they have any perception to the environment with their hands.

All the participants were tested over the course of two experimental tests, as is shown in Fig. 6(c). The first is the Fitt's law test (FLT), the second is prosthetic control test (PCT). Each test contained one training session and one evaluation session. During each training session, the subject needed to learn two feedback strategies (VP and SAR). The experimenter would guide the arm of the subject to approach the target as a teaching procedure, the training time for each strategy was within 10 minutes. After the training session, the evaluation session was carried out, the subject needed to independently follow

the auditory cues to play 25 trials of virtual game as shown in Fig. 5, and complete 25 trials for real grasp.

Since it is difficult to precisely describe the non-orthogonal location into words, the main difference between VP and SAR lies in the performance of the non-orthogonal target localization. As is shown in Fig. 6(d), SAR can give users intuitive perception and achieve a more natural grasping trajectory, while VP needs to generate the grasping path in advance, and adopts a non-optimal right-angle path for reliable grasping. We studied the influence of the target location on the grasping performance of PCT. As shown in Fig. 6(b), we drew a semi-circular arc from 30 degrees to 150 degrees on the table with the initial coordinates of the prosthetic hand as the center of the circle, every 30 degrees adjacent to each other was set as a possible target position. The table only contains one target in each trial, which was randomly positioned among the five possible location, making each target appear at the same distance but in different orientation. There was also a 30 seconds of maximum time limit for each trial and the start of each trial was controlled by the user.

PCT is more difficult than FLT, because the subject need to adjust the posture of the prosthetic to precisely align the target. In SAR, the subject can actively adjust the posture of the prosthetic hand according to the position of the sound source. However, adding a posture prompt in VP not only increases the complexity of the subject's manipulation, but also wastes much time on the calibration of grasping posture. As a result, we canceled the posture prompts of the prosthetic hand in VP, the subject needs to keep the arm posture fixed when moving.

In FLT, the ideal range of ID is approximately 0∼7 bits, if the range is too narrow, it may under-fit the linear regression with respect to ID and MT, making the TP fluctuate drastically between subjects. Therefore, the span of the movement amplitude (i.e., distance of "A" in Fig. 5) in FLT needs to be larger than that (i.e., distance of "R" in Fig. 6(b)) in PCT, as well as a smaller effective width than that in PCT. As a result, if the effective width was reduced while kept the loudness of the sound source constant, in our preliminary experiment we found that the subjects were difficult to quickly distinguish the exact coordinates of the target although they were already very close to the sound source. Meanwhile, when the distance is far, it is also difficult for the subjects to determine the direction of the sound source due to the very low sound volume. Therefore, we added a new experimental control group, vSAR, which is a strategy based on SAR and the same directional speech cues as in VP (the distance cues of VP was not involved). In addition, the mouse movement sensitivity was also adapted according to the effective width of the target, so that the subjects would not over-shot frequently.

Through the comparison of VP and SAR, it seems that the performance of SAR is inherently better than that of VP, it is not only because of the limited size of the instruction dictionary in VP that force the subjects to follow a longer reaching path, but also it is essentially a discrete command feedback, which has a lower ITR compared with SAR. As a result, the comparability between these two methods has become unfair. Therefore, we added another feedback strategy that hopes to balance VP and SAR, denoted as cVP. Here, the direction cues of cVP still used the discrete four-direction voice command as in VP, but the distance cue was replaced by a continuous guide audio as is used in SAR group (as shown in Figure 3(b)), but there is no spatial audio effect. We normalized the distance and linearly mapped it to the volume of the guide audio, where the low volume indicates a long distance, and a high volume indicates a short distance. We hoped that this continuous guide audio would improve the efficiency on FLT and PCT.

In each trial of FLT, one of the four guidance strategies (VP, cVP, SAR, vSAR) was randomly adopted to prevent the subject's proficiency from the long-term mono-strategy. If the cursor was outside the target zone as pressing the space bar, or not finished the localization within 30 seconds, the current trial would be determined as a failure and terminated immediately. After that, the target object would be randomly repositioned, and the next trial will start immediately.

### D. Statistical Analysis

Statistical analysis was performed using PRISM (GraphPad Inc.) or custom Python scripts. Effect sizes are reported throughout the article as information transfer rate (*TP*), to highlight across-condition (e.g., VP versus SAR) differences. Additionally, the root-mean-square error (RMSE) was used to quantify the fitness of TP curve, the Pearson correlation coefficient ($r$) was used to evaluate the correlation between MT and ID. One-way repeated-measures ANOVAs were used with the main effects of feedback strategy. All behavioral were first evaluated with the Shapiro-Wilk test to test for the normality of the residuals of a standard ANOVA. If the P value of the majority of all multiple comparisons was less than 0.05, then a rank-transformed ANOVA was used. Otherwise, a standard ANOVA was used. Finally, a Tukey's post hoc test was used to correct for multiple comparisons.

## IV. Results

### A. Fitt's Law Evaluation

We collected a total of 400 trials in FLT from Sixteen subjects, the success rate is shown in Fig. 7(a), where the successful trial needed to meet two conditions: 1) Localizing the cursor within the effective width of the target; 2) The completion time was above the top 96% of total trials. The results show that the success rate under all three feedback strategies exceeds 90%, but VP group is lower than SAR group (VP vs. vSAR: $p = 0.0259$; VP vs. SAR: $p = 0.0238$; cVP vs. vSAR: $p = 0.0104$; cVP vs. SAR: $p = 0.0266$).

In terms of localization accuracy, the absolute error violin plots were shown in Fig. 7(b), where the columns of "error_x" and "error_y" represent the localization error along the x and y axis, "error" column represents the $\ell^2$-norm of error vector ($error\_x$, $error\_y$). The experimental results show that the SAR method can significantly improve the localization accuracy in the horizontal direction (x axis) compared with the VP method. However, there is no significant difference in the vertical direction (y axis).

In terms of localization efficiency, we plotted the violin plots of the three feedback strategies with respect to completion

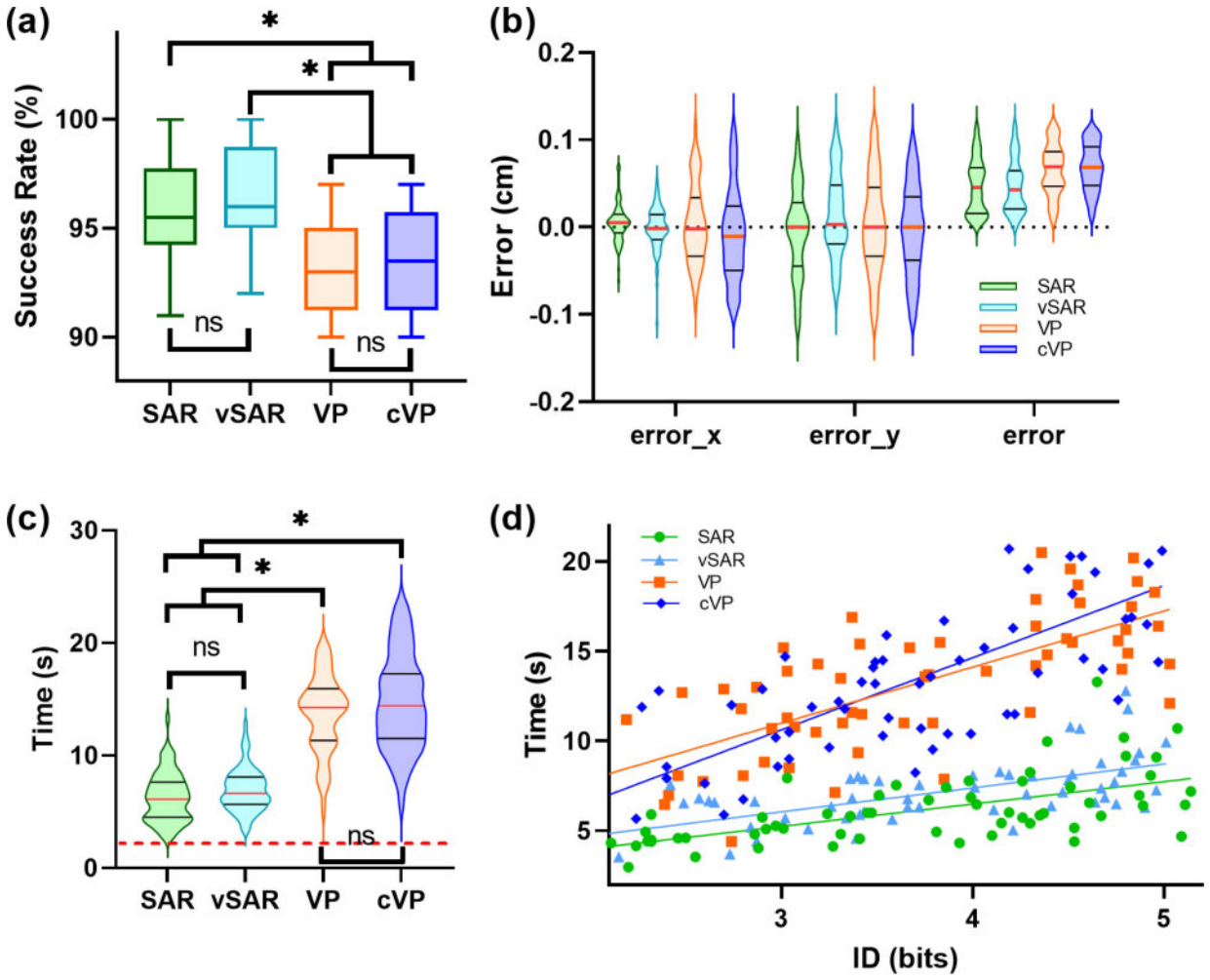


Fig. 7. Experimental results in FLT. (a) The success rate of four feedback strategies. (b) The violin plots of absolute error in localization. (c) The violin plots of completing time. (d) The scatter plot based on MT and ID and throughput line. For the sake of simplicity of the scatter figure, only part of the sample points are shown. Tukey's HSD post hoc test: *P $<$ 0.05.

time (Fig. 7(c)). Additionally, we measured the average completion time of common visual feedback as the highest level, represented by the red dash line. It is shown that SAR and vSAR have significantly faster completion time than VP and cVP ($p$<0.0001), and there is no significant difference between SAR and vSAR ($p = 0.1488$), or VP and cVP ($p = 0.3014$).

The scatter plot based on MT and ID was shown in Fig. 7(d). The fitting results show that the ID has a more significant positive correlation with MT in VP ($r = 0.7227$) and cVP ($r = 0.7615$) than in SAR ($r = 0.4988$) and vSAR ($r = 0.5123$). These results indicate that in SAR group, the complexity of task caused by distance is reduced, and the performance of the subjects is closer to that by visual perception, while the completion time of VP is more affected by distance.

By fitting the linear regression function, TP was calculated for the three feedback strategies, where VP was 0.320 bits/s, cVP was 0.249 bits/s, SAR was 0.803 bits/s, and vSAR was 0.750 bits/s, indicating that the information transfer rate of SAR was significantly larger than VP. In addition, the RMSE of linear regression using VP was 2.652, cVP was 2.879, SAR was 1.552, and vSAR was 1.416, indicating that the inclusion of speech cues makes the completion time for different IDs more stable than single SAR, while the completion time of VP and cVP shows a much significant fluctuation.

### *B. Prosthetic Control Result*

After the FLT experiment, it is found that the differences within SAR group (SAR and vSAR) and VP group (VP and cVP) were not significant, and the subjects were able to establish a more accurate environment perception mapping during a period of learning, although they had insensitive direction perception at the beginning of the evaluation session. Therefore, only SAR and VP were compared in the PCT experiments.

The successful rate of PCT is shown in Fig. 8(b), it is shown that the average grasping successful rate for VP and SAR is lower than that of FLT, and there is no significant difference in the grasping success rate between the two feedback strategies ($p = 0.1078$). However, the main differences were focused on the grasping path and task completion time.

Fig. 8(a) shows the fetching paths from one of the representative subjects, it shows that the SAR feedback has a more natural grasping path. Since the prosthetic hand in PCT is a right hand, the paths shows that for targets on the right side

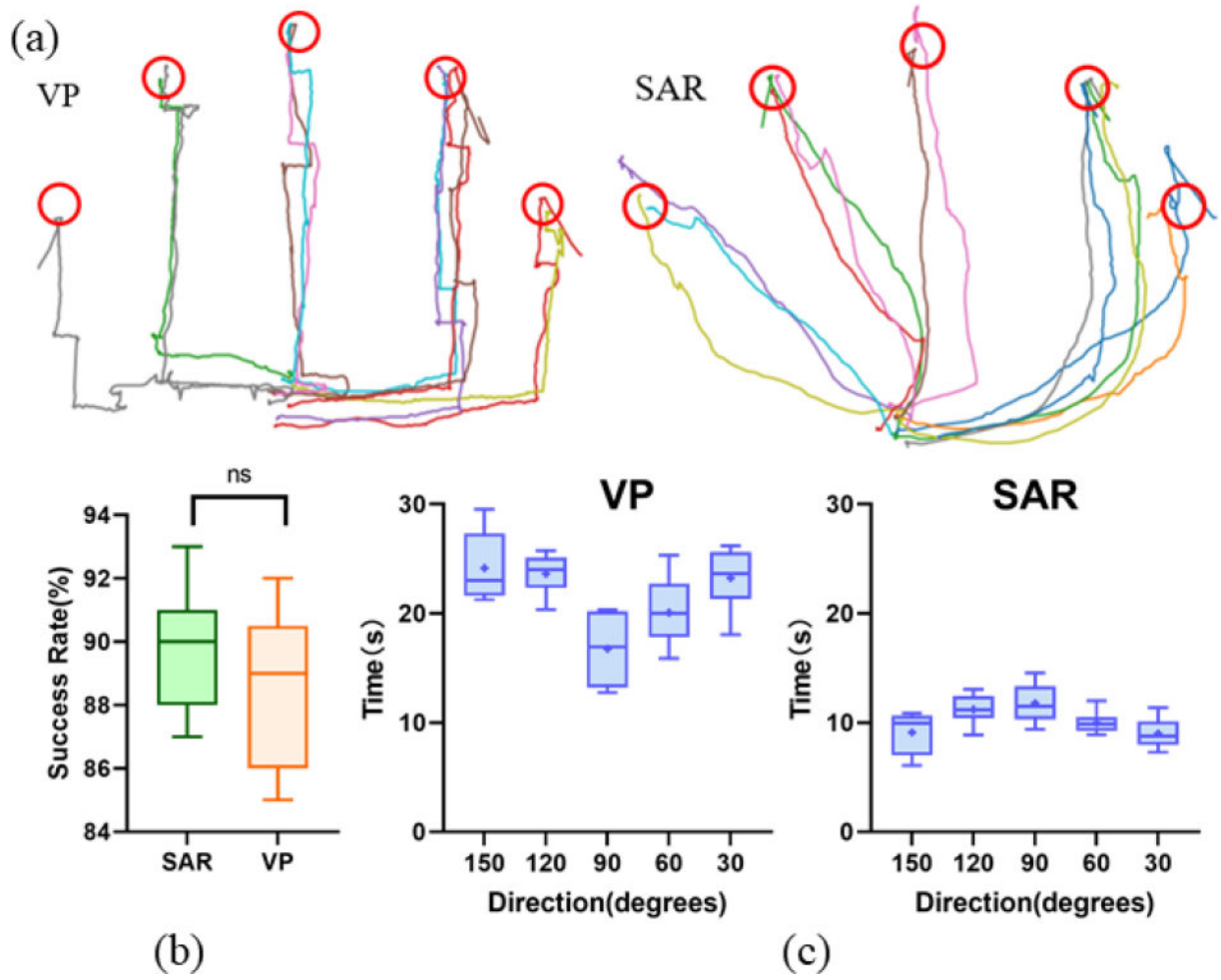


Fig. 8. Experimental results in PCT. (a) The fetching paths by VP and SAR feedback. (b) The success rate of VP and SAR feedback. (c) The box plots of completing time for target localization in different directions.

of the hand, the subject can still follow the natural grasp path, getting around to the rear side of the target to complete a reliable grasp. For targets on the left side, the prosthetic hand is more easily aligned with the target, so the subject can also approach the object with a shorter path.

For the paths under the VP feedback (Fig. 8(a)), it is found that there are many right-angle movement on both of the horizontal and vertical paths, this is because the subject was correcting the cumulative errors during the movement. Furthermore, for those targets far away from the left side of the prosthetic hand, a greater body shift was inevitable in order to maintain a fixed hand posture, which caused considerable inconvenience in practical grasping.

We calculated the completion time for reaching and grasping the target in different directions, as is shown in Fig. 8(c). When using VP feedback, the fetching time for different directions shows significant differences. The completion time to grasp the target on the direction of 90 degrees is the fastest, and the completion time increases as the direction of the target deviates from 90 degrees. For the objects located on the palm side of the prosthetic hand (on the direction of 120 and 150 degrees), subjects had to move their bodies due to the need of maintaining the arm posture, resulting in a longer completion time than the dorsal side of the prosthetic hand. In contrast, SAR is less affected by target orientation than VP, and the average completion time is almost doubled compared to the VP approach.

Experimental results show that the proposed SAR can provide users with intuitive environmental information without relying on their inborn vision. More importantly, the user can interact with the environment at a higher information transmission rate. Finally, the developed prosthetic control system provides a more natural assistance to the user.

## V. Discussion

By analyzing the experimental results, it is shown that the difference of success rate between VP and SAR is not significant, it mainly because the subjects aimed at completing each trial and there is no strict time limit, so both of the feedback sources can guide the subjects to finish the task eventually. A more important evaluation index is the localization efficiency, the experimental results of FLT and PCT show that SAR can save half of the completion time compared with VP.

In order to render the spatial audio, the virtual scene only needs to the spatial information of the target objects in the real environment. The configuration of the virtual scene (the size and the material of the scene) can enhance the user's

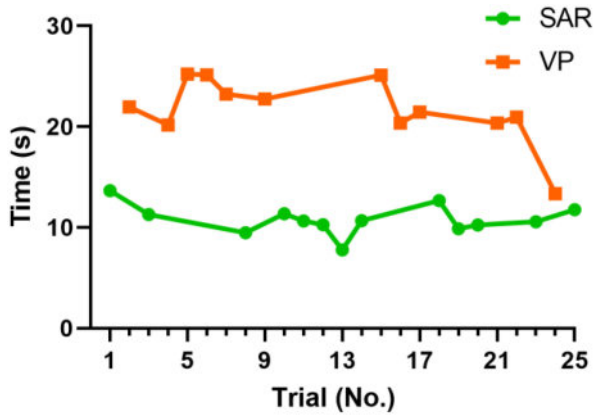


Fig. 9. Learning effect of SAR and VP.

location perception, so there is no need to realistically simulate the environment, which reduces the burden of real-time scene modeling.

In the PCT experiment, although having the wider effective width and shorter movement amplitude than in FLT, the task success rate of PCT was slightly lower than that of FLT, this is because the prosthetic hand increased the difficulty of interaction between human and environment.

The reason for the longer completion time in VP is two-fold. On the one hand, the language expression is inherently slower than the instinct to sound localization. The voice prompts needs to be spaced to prevent auditory fatigue from frequent disturbance. As a result, the subjects often had to wait for the cue for the next move, which reduced the efficiency. In cVP, we replaced the discrete distance feedback with a continuous sound effect, but there was no significant improvement. In addition, the completion time of VP is more influenced by the orientation of the target. When VP was located roughly in the orthogonal direction, the difference between the system-planned path and the optimal path is not significant. While in terms of non-orthogonal orientation localization, the guiding method using stepwise alignment of each orientation is significantly slower than the shortest path. Furthermore, as the complexity of the environment increases, such as obstacles, or movement along a specific trajectory, subjects need to receive more intensive cueing information, which is prone to fatigue and boredom, which also indicates that VP is not suitable for interaction in complicate scenarios.

On the other hand, SAR can give users the freedom of active perception, that is, the reaching path is no longer planned by the system and users do not need to passively receive the moving commands from the system. Moreover, the subjects can actively perceive the approximate position of the object even if it is still some distance away. As a result, the user can adjust the posture of the prosthetic hand in advance. Otherwise, the prosthetic hand may not correctly align with the target, then the outer edge of the hand can easily knock over the object.

The potential learning effects may have taken place over the significant number of trials that participants were asked to perform across experiments. Fig. 9 shows the completing time of one representative subject during the consecutive 25 trials of PCT, where VP or SAR was randomly selected as the feedback strategy in each trial. It is found that the completion time of the two methods did not decrease with the increase of the number of trials, indicating the learning effect is not significant across the experiment trials.

Besides, the proposed environmental perception assisting system has more to be explored. The current experiment used able bodied participants in place of blind amputees, our future work will involve members of the actual target population in order to properly validate the usability and usefulness of the system. Furthermore, we will put effort into real-time spatial audio techniques based on wearable devices such as Hololens 2 (Microsoft Ltd.) [29] and Realsense (Intel Ltd.). In terms of environment perception, when facing more complex interaction scenarios, it is obviously not enough to rely on SAR alone, so the inclusion of full hand coverage haptic sensor can better solve the problem of target localization in close range. In terms of system integration, the inclusion of first-view camera device and the portable embedded hardware can enhance the wearability of the system. Given some of the complex applications such as multipath complications or fading issues, we will optimize the system functions to support the two-armed collaboration capability. For the disabled with multi-sensory impairments, this study also has some limitations. For example, the subjects need to have at least an intact 3D hearing system (a hypothesis that can be too strong for those blind amputees if this is because of a traumatic event as in the army). In addition, in order to achieve the best spatial audio perception performance, it is better to require subject-specific HTFT, which requires more expensive calibration and debugging procedures. Even so, among the existed wearable sensory feedback solutions, SAR is still an effective and convenient method that allows subjects to intuitively perceive the environment. In the future, it may be possible to use bone conduction or cochlear implant technology, which can directly transfer the spatial audio to the disabled.

## VI. Conclusion

In this paper, we proposed an intuitive environmental perception assistance method for blind amputee based on spatial audio rendering (SAR). The method of SAR expresses the real environmental information in a virtualized way. Through the full use of the human instinct on sound localization, the user can obtain the natural spatial information perception to the interacted objects. Comparing to voice prompts (VP), SAR significantly improves the information transfer rate, and reduces the completion time by half than the VP while restoring the natural grasping path. The experimental results demonstrated the potential applications for blind amputees to reconstruct the control and sensory loops. This study presents an extended perspective on the real-time multi-modal perception feedback technology, the integration of wearable robotics, and the environmental perception based on the vision and haptics of the machine, which contributes to the proprioceptive control to the prosthetic robots.

## REFERENCES


[1] G. L. Krahn, "WHO world report on disability: A review," *Disabil. Health J.*, vol. 4, no. 3, pp. 141–142, 2011.

[2] R. L. Schatz and M. P. Rosenwasser, "Krukenberg kineplasty: A case study," *J. Hand Therapy*, vol. 15, no. 3, pp. 260–265, 2002.

[3] M. Markovic, S. Dosen, D. Popovic, B. Graimann, and D. Farina, "Sensor fusion and computer vision for context-aware control of a multi degree-of-freedom prosthesis," *J. Neural Eng.*, vol. 12, no. 6, Dec. 2015, Art. no. 66022.

[4] M. Cognolato *et al.*, "Gaze, visual, myoelectric, and inertial data of grasps for intelligent prosthetics," *Sci. Data*, vol. 7, p. 43, Feb. 2020.

[5] G. Ghazaei, A. Alameer, P. Degenaar, G. Morgan, and K. Nazarpour, "Deep learning-based artificial vision for grasp classification in myoelectric hands," *J. Neural Eng.*, vol. 14, no. 3, Jun. 2017, Art. no. 036025.

[6] C. Shi, D. Yang, J. Zhao, and H. Liu, "Computer vision-based grasp pattern recognition with application to myoelectric control of dexterous hand prosthesis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 9, pp. 2090–2099, Sep. 2020.

[7] M. Markovic, H. Karnal, B. Graimann, D. Farina, and S. Dosen, "GLIMPSE: Google glass interface for sensory feedback in myoelectric hand prostheses," *J. Neural Eng.*, vol. 14, no. 3, 2017, Art. no. 036007.

[8] C. Antfolk, M. D'Alonzo, B. Rosen, G. Lundborg, F. Sebelius, and C. Cipriani, "Sensory feedback in upper limb prosthetics," *Expert Rev. Med. Devices*, vol. 10, no. 1, pp. 45–54, Jan. 2013.

[9] D. Zhang, H. Xu, P. B. Shull, J. Liu, and X. Zhu, "Somatotopical feedback versus non-somatotopical feedback for phantom digit sensation on amputees using electrotactile stimulation," *J. Neuroeng. Rehabil.*, vol. 12, p. 44, May 2015.

[10] L. Zollo *et al.*, "Restoring tactile sensations via neural interfaces for real-time force-and-slippage closed-loop control of bionic hands," *Sci. Robot.*, vol. 4, no. 27, May 2019, Art. no. eaau9924.

[11] L. E. Osborn *et al.*, "Prosthesis with neuromorphic multilayered e-dermis perceives touch and pain," *Sci. Robot.*, vol. 3, no. 19, Jun. 2018, Art. no. eaat3818.

[12] D. J. Ni, A. Song, L. Tian, X. N. Xu, and D. F. Chen, "A walking assistant robotic system for the visually impaired based on computer vision and tactile perception," *Int. J. Social Robot.*, vol. 7, no. 5, pp. 617–628, Nov. 2015.

[13] J. Salazar, K. Okabe, and Y. Hirata, "Path-following guidance using phantom sensation based vibrotactile cues around the wrist," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2485–2492, Jul. 2018.

[14] G. Flores, S. Kurniawan, R. Manduchi, E. Martinson, L. M. Morales, and E. A. Sisbot, "Vibrotactile guidance for wayfinding of blind walkers," *IEEE Trans. Haptics*, vol. 8, no. 3, pp. 306–317, Jul.–Sep. 2015.

[15] C. Roads and J. Strawn, *The Computer Music Tutorial*. Cambridge, MA, USA: MIT Press, 1996.

[16] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press, 1997.

[17] K. Watanabe, K. Ozawa, Y. Iwaya, Y. Suzuki, and K. Aso, "Estimation of interaural level difference based on anthropometry and its effect on sound localization," *J. Acoust. Soc. America*, vol. 122, no. 5, pp. 2832–2841, 2007.

[18] D. W. Batteau, "The role of the pinna in human localization," *Proc. Royal Soc. London Series B Biol. Sci.*, vol. 168, no. 1011, pp. 158–180, 1967.

[19] A. Kulkarni, S. K. Isabelle, and H. S. Colburn, "On the minimum-phase approximation of head-related transfer functions," in *Proc. Workshop Appl. Signal Process. Audio Accoust.*, 1995, pp. 84–87.

[20] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space," *J. Audio Eng. Soc.*, vol. 49, no. 4, pp. 231–249, Apr. 2001.

[21] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2001, pp. 99–102.

[22] A. W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature*, vol. 397, no. 6719, pp. 517–520, 1999.

[23] F. Zotter and M. Frank, *Ambisonics : A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality* (Springer Topics in Signal Processing). Cham, Switzerland: Springer, 2019.

[24] T. Lund, "Enhanced localization in 5.1 production," in *Audio Engineering Society Convention 109*, Audio Eng. Soc., New York, NY, USA, 2000.

[25] D. Farina *et al.*, "The extraction of neural information from the surface EMG for the control of upper-limb prostheses: Emerging avenues and challenges," *IEEE Trans Neural Syst. Rehabil. Eng.*, vol. 22, no. 4, pp. 797–809, Jul. 2014.

[26] Y. Gu, D. Yang, Q. Huang, W. Yang, and H. Liu, "Robust EMG pattern recognition in the presence of confounding factors: Features, classifiers and adaptive learning," *Expert Syst. Appl.*, vol. 96, pp. 208–217, Apr. 2018.

[27] P. M. Fitts and B. K. Radford, "Information capacity of discrete motor responses under different cognitive sets," *J. Exp. Psychol.*, vol. 71, no. 4, pp. 475–482, Apr. 1966.

[28] I. S. Mackenzie, "Fitts' throughput and the remarkable case of touch-based target selection," in *Proc. Int. Conf. Human-Comput. Interact.*, 2015, pp. 238–249.

[29] M. Eckert, M. Blex, and C. M. Friedrich, "Object detection featuring 3D audio localization for Microsoft HoloLens," in *Proc. 11th Int. Joint Conf. Biomed. Eng. Syst. Technol.*, vol. 5, 2018, pp. 555–561.