International Journal of Social Robotics
https://doi.org/10.1007/s12369-020-00718-w

Human–Robot Collaborative Control in a Virtual-Reality-Based Telepresence System

Jianhao Du[1] · **Ha Manh Do**[2] · **Weihua Sheng**[3]

Accepted: 26 October 2020

Abstract
In this paper, we develop a robotic telepresence system to provide remote users with immersive embodiment in local environments through a custom-designed mobile robot. The proposed telepresence system uses a virtual reality (VR) device to connect a remote user to the robot. Three dimensional visual data from a RGB-D camera are rendered for real-time stereoscopic display in the VR device, which forms a deeply-coupled human machine system and creates an immersive experience of telepresence. Based on a user study, it is found that better user experience can be achieved by allowing the robot to track the speaker while being aware of the intention of the remote user. To this end we propose a human-robot collaborative control framework based on human intention recognition and sound localization. The intentions of head movement of the remote user are inferred based on the motion of the VR device using hidden Markov models. The speaker is tracked through sound source localization using a microphone array. A collaborative control scheme is developed to fuse the control from the robot and the remote user. Experiments are conducted in both one-to-one and one-to-two remote conversation scenarios. The results show that the proposed system can significantly improve the immersiveness and performance of robotic telepresence systems, therefore greatly enhancing the user experience of such telepresence systems.

Keywords Telepresence · Virtual reality · Sound localization · Collaborative control

1 Introduction

Telepresence robots [1–3] allow people to "transport" themselves to remote locations and accomplish certain tasks such as having a conversation and conducting teleoperation. Telepresence robots have many potential applications. For example, doctors can interview and examine patients remotely, or even perform surgery from distant locations. Search and rescue teams can conduct exploration and inspection via robots in hazardous environments. Students can attend school from their homes when they have to be absent due to sickness. A typical scenario of robotic telepresence is that a remote user controls a robot in a local environment and interacts with local users via the robot, as illustrated in Fig. 1. The remote user experiences his/her presence in the local environment through sensory stimulus provided by the different sensors (cameras, microphones, etc.) on the mobile robot and a rendering device such as a head mounted display (HMD). The local users can also feel the existence of the remote user via the robot avatar which mimics the appearance and actions of the remote user. Thus, a robotic telepresence system involves both human-human interactions and human-robot interactions.

A traditional robotic telepresence system [3–5] is essentially a video conferencing system implemented with a mobile robot, which consists of a joystick and a computer at the remote site, an LCD screen, a web camera, a microphone, and speakers on the robot at the local site, allowing "face-to-face" communication between the remote and local users. Such robotic telepresence systems lack immersiveness and are usually not intuitive to operate for the remote user, which also causes poor experience to the local users. In this

✉ Weihua Sheng
weihua.sheng@okstate.edu

Jianhao Du
jdu@mathworks.com

Ha Manh Do
ha.do@louisville.edu

[1] Mathworks Inc., Natick, USA

[2] Department of Electrical and Computer Engineering, University of Louisville, Louisville, USA

[3] School of Electrical and Computer Engineering, Oklahoma State University, Stillwater, USA

Published online: 09 November 2020

Fig. 1 The concept of robotic telepresence: a remote user interacts with local users via a remotely controlled robot

paper we study the problem of how to develop a more immersive and user-friendly robotic telepresence system, so that it can help improve the user experience of the participants involved. There are two key problems that a robotic telepresence system should address. The first problem is how to develop a proper human-robot interface to allow the remote user to gain an immersive experience of telepresence. The second problem is how to facilitate smooth and easy control of the telepresence robot by the remote user, which will create better user experience for both the remote and local users.

For the first problem, we propose that the virtual reality technique can be utilized to enhance the remote user's experience of telepresence. Usually virtual reality devices provide sensory stimulus from simulated environments and allow the user to interact with the simulated environments. In recent years, the virtual reality technique progresses rapidly and many VR devices have been developed, such as the Cardboard [6] from Google and the Oculus Rift [7] from Facebook. These devices are excellent interfaces for robotic telepresence which provide immersiveness, realisticness and interactivity. The visual data from the cameras on the robot can be displayed in realtime to the user via the VR device, while the user can send commands to the robot via the head movement collected by the motion sensor in the VR device.

For the second problem, it is desirable to introduce some local intelligence into the robot, which allows the robot to take some actions based on its local sensor input so that the human-robot interaction experience can be improved. Therefore the robot has its own decisions and the remote user may have his/her intentions, which leads to a deeply coupled human-machine system. To deal with any potential conflict, we propose a collaborative control mechanism that allows the robot and the remote user to coordinate with each other smoothly and efficiently. Unlike direct teleoperation, the robot needs to predict user's intentions and then assists the user to fulfill his/her intentions.

This project considers the application of telepresence in a tele-conferencing scenario, as depicted in Fig. 1. The *main contribution* of this paper has three aspects. Firstly, we propose and develop a robotic telepresence platform which connects human and robot closely by taking advantage of the virtual reality device. The VR device facilitates immersive audio visual stimuli and intuitive control to the remote user. Secondly, we introduce an approach to infer the remote user's intentions based on the head movement captured by the VR device. We also develop a local intelligence in the telepresence robot that tracks the speakers in the local environment using a microphone array. Finally, a two-stage collaborative control scheme is proposed to fuse the human intentions and robot's local control, which facilitates smooth human-robot collaboration in telepresence. To the best of our knowledge, this is the first work to combine human intentions and robot local intelligence in an immersive telepresence setting.

This paper is organized as follows: in the next section, the related work is introduced. In Sect. 3, we present an overview of the proposed virtual-reality-based robotic telepresence system. In Sect. 4 we conduct a user study and the collaborative control framework is explained in detail. Experimental results are demonstrated in Sect. 5. Section 6 concludes the paper and outlines the future work.

2 Related Work

A telepresence robot can be defined as an avatar in a distant environment that is operated by a remote user and carries out certain tasks, for example, facilitating social interactions between the remote user and one or more local users. Many research work has been focused on how to develop an intuitive interface to connect the remote user and the robot closely, and thus to improve the user experience. Adalgeirsson and Breazeal [8] designed and evaluated a telepresence robot named MeBoT which had social expressions. They found that people felt more psychologically focused and more engaged in the interaction with their remote partners when they were embodied in a socially expressive way. Escolano et al. [9] developed a telepresence system with a noninvasive brain-computer interface (BCI) to provide a user with presence in remote environments through a mobile robot. The system inferred the human intention using the BCI decoding of task-related orders and collaborated with the robot. Sirkin and Ju [10] conducted a study on how augmented movement capability would improve the user experience in telepresence meetings. The robot had both a display screen and a moving base. Martins et al. [11] designed a teleoperation system based on a field robot and a head-mounted display was used for immersive display. The video streams were transferred from a pair of stereo cameras located on the robot and the tracker on the HMD provided head orientation which was used for operation. Similarly, Kratz et al. [12] developed a mobile telepresence robot for navigation tasks. The robot was able to perform immersive navigation using head-tracked stereoscopic video and a head-mounted display. A user study was conducted and the results showed the improvements of user experience. In [13], a simple telepresence robot was designed to perform navigation tasks under the remote control of a human operator. The operator interacted with a virtual environment mapped to robot's

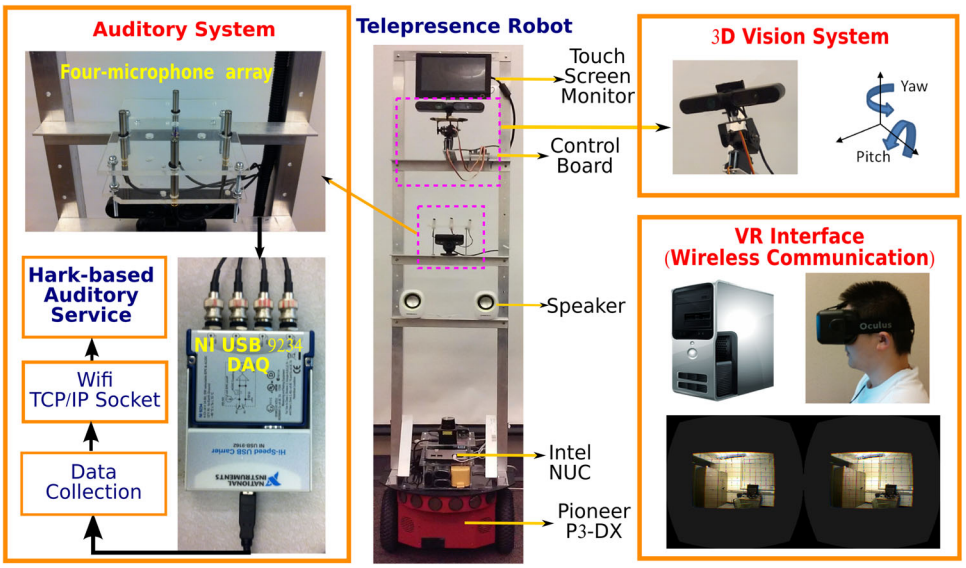

Fig. 2 The proposed VR-based robotic telepresence system

real world which improved usability and reduced power consumption. Recently a platform named DORA [14] attempted to track the movements of a user's head with an advanced camera system. The user was able to gain immersive experiences by wearing an Oculus Rift goggle.

On the other hand, in a robotic telepresence system, the robot can provide assistance while the user is teleoperating the robot. In these cases, we have a human-robot collaborative control system which has received much interest recently. Such collaborative control systems can infer user intentions based on sensor observations, and the intentions can be fused with the robot decisions to obtain the final control. Stiefelhagen et al. [15] developed a system to estimate the visual focus of attention of human users from multiple cues in a televideo meeting. Both visual and audio data were used to predict user intentions. Gao et al. [16] proposed a shared autonomy system based on a mobile robot. The robot recognized the user intentions by estimating the task using the context information and provided motion assistance accordingly. Carlson et al. [17] developed a shared control system for a brain-computer interfacing (BCI) controlled wheelchair. The robot interpreted BCI commands by considering the context of the surroundings perceived through vision. The vision information provided reliable cues to the shared controller, allowing the synthesized BCI interface to drive safely in an indoor environment. Almeida et al. [18] proposed a human-embodiment method for tele-operating a mobile robot. The human intentions were inferred from user's body postures captured through vision.

Overall, although there are many existing robotic telepresence systems, most of them are not user friendly and therefore not appealing to the users. The key factors of building an ideal telepresence robot are a friendly interface that tightly connects the remote user and the robot, and an efficient scheme for easy and intuitive control. In this paper, we propose to use an Oculus Rift goggle as the interface to provide a tight connection between the remote user and the robot. We use a RGB-D camera and an array of four microphones to enhance the sensing capability of the robot. Moreover, we propose a novel collaborative control framework based on human intention recognition and sound localization to improve the experience of the users.

3 System Overview

In this section, we first describe the overall design of the proposed virtual-reality-based robotic telepresence system. Then we introduce the pose integration and stereoscopic rendering for the display of virtual reality, as well as the approach for sound localization.

3.1 Overall Design

The proposed VR-based robotic telepresence system is shown in Fig. 2. The system has two main components: the mobile platform and the remote station. The mobile platform consists of a Pioneer mobile robot base [19], a compact mini-computer, a microphone array (4 microphones), a touch-screen monitor, two speakers and a RGB-D camera. The mobile robot base provides the ability to move around in the environment. The compact mini-computer (Intel NUC [20]) receives all the sensor data and sends commands to the robot. The microphone array is used to collect sound signals. The RGB-D camera is used to capture the color and depth images necessary for 3D data display. To enlarge the field of view (FoV), the RGB-D camera is mounted on a pan-tilt unit controlled by a control board using the Pulse Code Modula-

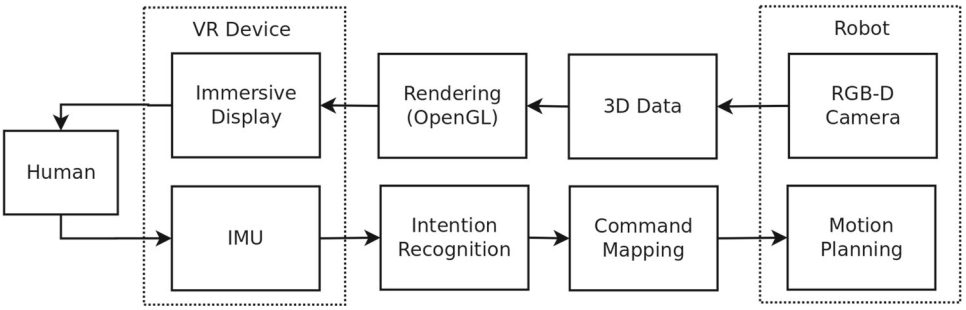

Fig. 3 The functional block diagram of the proposed virtual-reality-based robotic telepresence system

tion (PCM) signals. The camera is calibrated with the pan-tilt unit using external markers through a motion capture system. Additionally, the touch screen and speakers are used for displaying the information of any user connected to the system. On the remote site, a head-mounted display for virtual reality (Oculus DK2 [7]) is adopted to connect the human to the robot via a workstation with wireless communication. The Oculus DK2 has dual lenses to provide a stereoscopic 3D perspective and an IMU for orientation tracking.

The functional block diagram of the proposed virtual-reality-based robotic telepresence system is shown in Fig. 3. The system is divided into two parts. The first part is data acquisition and visualization. The 3D point cloud data are captured by the RGB-D camera mounted on the robot and rendered using Open Graphics Library (OpenGL [21]), and then transmitted to the Oculus Rift goggle for immersive display to the human. The current pose of the robot is aligned with the virtual viewpoint in the Oculus Rift goggle, which allows the human to explore the 3D data. The second part is human intention recognition for robot control. The intentions of the human head movement are inferred from the inertial measurement unit (IMU) data in the Oculus Rift goggle using hidden Markov models (HMMs). Then the intention results are translated into the control commands to guide the robot to interact with the environment or other humans. An application scenario of such an immersive VR-based telepresence system is shown in Fig. 4, in which one remote user engages one or two local users in a conversation via the robot.

Fig. 4 The application scenario of an immersive VR-based telepresence system (top images: 1 vs. 1 conversation. Bottom images: 1 vs. 2 conversation)

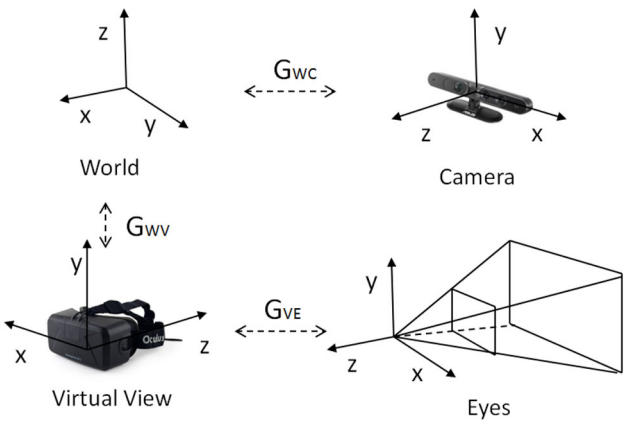

Fig. 5 Defined frames for pose integration

3.2 Pose Integration and Stereoscopic Rendering

To display the data from the camera to the human, we implement a rendering pipeline to integrate the 3D data and the virtual reality device using the OpenGL library. The pipeline takes the images as an input and gives the output to the VR device. The 3D point cloud data are reconstructed based on the color and depth images from the RGB-D camera. The movements of the camera and the VR device are coordinated in the pipeline. Before rendering the 3D point cloud to the VR device, the pose of the RGB-D camera in the real world needs to be associated with the viewpoint in the virtual environment for display. Several coordinates are defined as shown in Fig. 5. The world frame is defined as the initial position of the telepresence robot in the real world. The relative transformation matrix $\mathbf{G_{WC}}$ between the world frame

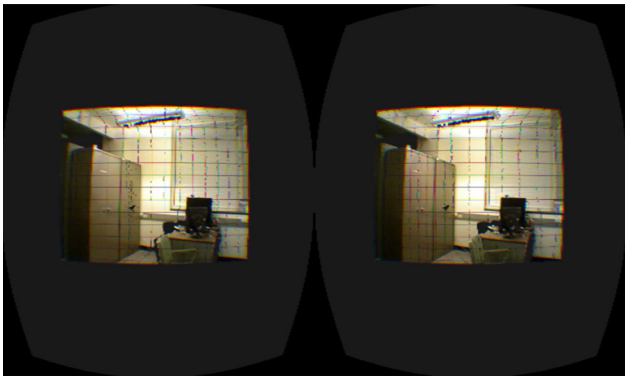

Fig. 6 The views in Oculus Rift

and the camera frame is obtained from the system calibration process. The transformation matrix $\mathbf{G_{WV}}$ represents the VR device's relative motion with respect to the world frame, which is calculated using the data (roll, pitch and yaw) from the IMU in the VR device. $\mathbf{G_{VE}}$ is the projection matrix from the 3D world on a screen or display, which is calculated as follows.

$$\mathbf{G_{VE}} = \begin{bmatrix} \frac{1}{a \cdot \tan(\beta/2)} & 0 & 0 & 0 \\ 0 & \frac{1}{\tan(\beta/2)} & 0 & 0 \\ 0 & 0 & \frac{-z_1 - z_2}{z_1 - z_2} & \frac{2z_1 z_2}{z_1 - z_2} \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad (1)$$

a is the screen aspect ratio and β is the angle of the field of view in the vertical direction. z_1 and z_2 are the near and far distance in z axis respectively.

After pose integration, the 3D data is rendered in a shared frame and then projected in split-screen stereo with half of the screen used for each eye as shown in Fig.6. The viewpoint is adjusted to remove the lens effects by applying a radial scaling function

$$f(r) = k_0 + k_1 r^2 + k_2 r^4 + k_3 r^6, \quad (2)$$

where r is the distance from each distorted pixel to the lens center and (k_0, k_1, k_2, k_3) are the distortion parameters. The stereoscopic rendering is implemented based on both the OpenGL and the Oculus SDK. The OpenGL provides the handle of the reconstructed 3D point cloud data to the Oculus SDK and then the Oculus SDK uses it to display for both left and right eyes, including shade generation, distortion compensation, etc.

3.3 Sound Localization and Separation

As shown in Fig. 2(left), the audio data are obtained using a microphone array that was built with 4 G.R.A.S IEPE (Integrated Electronic Piezoelectric) microphones [22] and an NI USB-9234 DAQ (Data Acquisition) [23]. This set of microphones has high-sensitivity at 50 mV/Pa, a wide frequency range up to 20 kHz, and a large dynamic range topping at around 135 dB. The DAQ is a USB-based four-channel C Series dynamic signal acquisition module for high-accuracy audio frequency measurements from IEPE and non-IEPE sensors. It can deliver a dynamic range of 102dB, incorporate programmable AC/DC coupling and IEPE signal conditioning for accelerometers and microphones, as well as digitize signals at rates up to 51.2 kHz per channel with built-in antialiasing filters that automatically adjust to the sampling rate.

The auditory software is developed based on HARK [24], which is an open source audition software consisting of modules for acoustic signal processing, sound localization and separation, speech recognition, and audio streaming. The data collection program is developed to capture the audio data from the microphones, filter them out, and send them to an audio stream receiver through a TCP/IP socket for sound localization and separation.

Sound localization is implemented based on the GEVD-MUSIC (Generalized EigenValue Decomposition-Multiple Signal Classification) method [25]. This method localizes sound sources by computing an eigenvalue decomposition vector of the correlation matrix between the inputs signal channels, then calculating MUSIC spectrum of this vector and the impulse responses (transfer functions) of microphones. The DoAs (Direction of Arrival) which have the largest values of the spectrum power are the sound source direction results.

The transfer function generally varies depending on the shape of the room and the relative positions between microphones and sound sources [26]. However, when ignoring acoustic reflection and diffraction, and given that the relative position of microphones and sound sources is known, the transfer function $H_D(k_i)$ is limited only to the sound source direction and calculated by the following Equation [26]:

$$H_{Dm,n}(k) = exp\left(\frac{-j2\pi\omega}{c} r_{m,n}\right) \quad (3)$$

where c is the speed of sound; ω is the frequency in the frequency bin k; $r_{m,n}$ is the difference between the distance from the microphone m to the sound source n and the distance from the reference point of the coordinate system to the sound source n.

The sounds from Ns sources are affected by the transfer function of each microphone $H_i(k)$ in space and perceived by M microphones as expressed by the following equation:

$$X_i(k) = \sum_{j=1}^{Ns} H_i(k) S_j(k) + N_i(k), \quad i = 1, 2, \ldots, M \quad (4)$$

where $S_j(k)$ is the Fourier transform of the jth sound source at the frequency k; $N_i(k)$ is the additive noise that includes environmental noise and electronic noise in each microphone. Sound source separation extracts the sound in each direction that is estimated by the sound localization from the recorded sound $X(k)$. The Fourier transform of separated sound $Y_j(k)$ is obtained from the following equation:

$$Y_j(k) = W_j(k)X(k) \tag{5}$$

The separation matrix $W_j(k)$ is estimated by Geometric-Constrained High-order Source Separation (GHDSS) [27] which has the highest total performance in various acoustic environments.

4 Collaborative Control for Telepresence Robot

Traditional telepresence robots are based on video conferencing and teleoperation, which have several limitations, including the lack of close engagement, difficulty of manipulation, etc. An ideal telepresence robot should represent the remote user in the environment while perceiving and interacting with local users. In order to find the best control strategy of the robot, we conduct a user study, in which a remote user controls the telepresence robot to have conversations with one or more local users.

4.1 User Study

The user study evaluates two different control strategies. The first is teleoperation which directly maps the head movement to the rotation of the camera. We call it the "teleoperated" method. The second one is that the robot has its own local intelligence which allows it to track the speakers through sound localization. We assume the only sounds in the environment are the voices from the local users. We call it the "sound-guided" method. A total of 10 human subjects participated in the study who are graduate students from our department. For each method, the user performs a conversation task through the telepresence robot as the remote user and the local user, respectively. After the task, the user is asked to answer several questions to assess each method, using a scale of zero to ten. We divide the questions into two categories based on the user's role: remote or local.

1. Questions for the remote user

   Q1 (Convenience): How convenient is it to use the telepresence system? (10 for the most convenient)
   Q2 (Workload): How much is the workload? (10 for the lowest)
   Q3 (Immersiveness): How immersive is your feeling during the conversation? (10 for the most)

2. Questions for the local user

   Q4 (Engagement): How much do you feel engaged with the operator via the robot? (10 for the most)
   Q5 (Embodiment): How much do you feel the embodiment of the remote operator? (10 for the most)

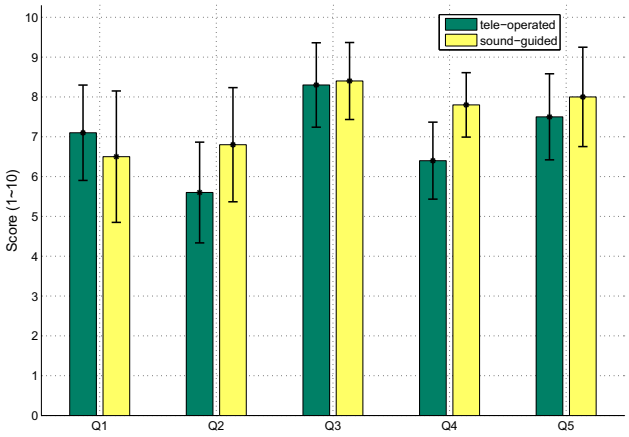

Fig. 7 The user study results of Q1 (Convenience), Q2 (Workload), Q3 (Immersiveness), Q4 (Engagement) and Q5 (Embodiment)

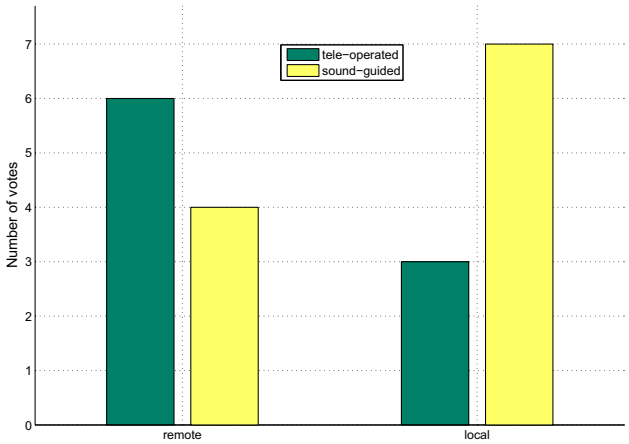

Fig. 8 The results of users' preferences for the two methods

The results of the user study are shown in Fig. 7 and the users' preferences for the two methods are shown in Fig. 8. The "teleoperated" method gives more manipulability but also induces heavier workload. The "sound-guided" method incurs less workload on the remote user. Both methods have high scores on immersive experience. For the local user, the "sound-guided" method has higher scores in engagement and embodiment. From Fig. 8, we can see the participants prefer the "teleoperated" method as a remote operator and the "sound-guided" method as a local user. We conclude that people want more manipulability from the remote side and

need more embodiment from the local side. Users feel more comfortable when the robot is changing its viewpoint according to the sound sources as in the "sound-guided" method. Although the "teleoperated" method provides direct control, the remote operator gets tired easily and sometimes it is difficult for the remote user to find the local user due to the inability to track the sound source.

4.2 Collaborative Control Framework

Based on the above user study we find that the two methods have both advantages and disadvantages. Therefore we introduce a shared control framework to combine these two methods. The main challenge is that the robot should be able to decide its actions based on two different control inputs: the remote operator's control and the robot's local control. To solve this problem, we propose a two-stage collaborative control framework as shown in Fig. 9. The first stage is the prediction of the remote user's intention by fusing the observations from the user input (head movement) and the microphone (sound), which predicts what the remote user wants to do. The second stage is the coordination based on the evaluation of the prediction and control commands, which derives the actions that the robot should take.

In the first stage, the prediction of the remote user's intention is modeled by the posterior probability $P(\mathbf{x}|\mathbf{O_h}, \mathbf{O_s})$. $\mathbf{x}$ is the intention of the remote user. In the case of daily conversation in telepresence, $\mathbf{x}$ is defined as $\{left, middle, right\}$, which indicates the direction the remote user wants to look at. The definition of the intention can vary to adapt to different scenarios. $\mathbf{O_h}$ and $\mathbf{O_s}$ are the observations from the human and the microphone array respectively. The observation $\mathbf{O_s}$ from the microphone array is the yaw angle of the detected human voice, which is obtained by the GEVD (Generalized Eigen Value Decomposition) method [28]. The observation $\mathbf{O_h}$ from human is the head movement (roll, pitch and yaw) obtained from the IMU inside the VR device.

The posterior probability is inferred using the Bayesian theorem as shown in Eq. (6).

$$P(\mathbf{x}|\mathbf{O_h}, \mathbf{O_s}) \propto P(\mathbf{O_h}, \mathbf{O_s}|\mathbf{x}) \cdot P(\mathbf{x}) \tag{6}$$

We assume $\mathbf{O_h}$ and $\mathbf{O_s}$ are independent given $\mathbf{x}$, so we have the fusion rule as follows:

$$p(\mathbf{x}|\mathbf{O_h}, \mathbf{O_s}) \propto P(\mathbf{O_h}|\mathbf{x}) \cdot P(\mathbf{O_s}|\mathbf{x}) \cdot P(\mathbf{x}) \tag{7}$$

Due to the lack of the prior knowledge on the distribution of $\mathbf{x}$, the prior $P(\mathbf{x})$ is set as a uniform distribution. Based on the Bayesian theorem, $P(\mathbf{x}|\mathbf{O_h}, \mathbf{O_s})$ is inferred through posterior fusion of two observations [29].

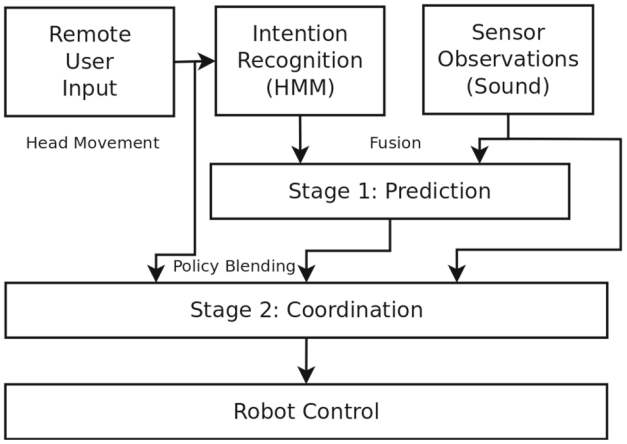

Fig. 9 The block diagram of the proposed two-stage collaborative control framework

$$\begin{aligned} P(\mathbf{x}|\mathbf{O_h}, \mathbf{O_s}) &\propto \frac{P(\mathbf{x}|\mathbf{O_h})}{P(\mathbf{x})} \cdot \frac{P(\mathbf{x}|\mathbf{O_s})}{P(\mathbf{x})} \cdot P(\mathbf{x}) \\ &\propto P(\mathbf{x}|\mathbf{O_h}) \cdot P(\mathbf{x}|\mathbf{O_s}) \end{aligned} \tag{8}$$

$P(\mathbf{x}|\mathbf{O_h})$ is derived through intention recognition using hidden Markov models (HMMs) [30]. The intentions are inferred from the head movement of the remote user. The head movement is characterized by a sequence of rotation angles obtained from the IMU inside the Oculus Rift. The rotation angles are roll, pitch and yaw with respect to the coordinate system of the Oculus Rift. Similarly, we define three intentions: *turn left*, *stay* and *turn right*, which correspond to the user's true state. To recognize the intentions, we use HMMs to characterize the head movements. HMM is a mathematical model of stochastic processes in terms of a directed structure of states and observations, which can be parametrized as $\lambda = (A, B, \pi)$. A is the state transition probability distribution, B is the observation probability distribution, π is the initial state distribution. The objective is to maximize the probability of $P(O|\lambda)$ and the problem can be solved by the Baum-Welch algorithm [31] in the training stage. Since the head movements are within relatively short range we set the number of the states as 5 in the model. The training data are aligned with the first observation and the length of the sequence is not fixed. In the real-time application, we apply a sliding window of observations to realize continuous recognition. Then $P(\mathbf{x}|\mathbf{O_h})$ is inferred based on the trained HMMs.

The range of the direction of arrival (DoA) of the sound sources is $(-\pi, \pi]$ which is divided into three regions ($left$, $middle$ and $right$): $(-\pi, -b]$, $(-b, b]$ and $(b, \pi]$. 0° is aligned with the current pose of the camera. b is set to 0.087 (5°) based on the accuracy of the sound localization algorithm. We use the angle θ to represent the real location of the

sound, which is defined as a circular normal distribution [32]:

$$p(\theta|\mathbf{O_s}) = f(\theta|\mathbf{O_s}, \lambda) = \frac{e^{\lambda \cos(\theta - \mathbf{O_s})}}{2\pi I_0(\lambda)} \tag{9}$$

λ is the concentration factor which is analogous to $1/\sigma^2$ in Gaussian distribution. $I_0(\lambda)$ is the modified Bessel function of order 0. The cumulative distribution function is calculated as:

$$F(\theta|\mathbf{O_s}, \lambda) = \frac{1}{2\pi}\left(\theta + \frac{2}{I_0(\lambda)}\sum_{j=1}^{\infty} I_j(\lambda)\frac{sin(j(\theta - \mathbf{O_s}))}{j}\right) \tag{10}$$

The probability $P(\mathbf{x}|\mathbf{O_s})$ is modelled as:

$$\begin{aligned} P(\mathbf{x} = left|\mathbf{O_s}) &= P(-\pi < \theta \leq -b|\mathbf{O_s}, \lambda) \\ &= F(-b|\mathbf{O_s}, \lambda) - F(-\pi|\mathbf{O_s}, \lambda) \\ P(\mathbf{x} = middle|\mathbf{O_s}) &= P(-b < \theta \leq b|\mathbf{O_s}, \lambda) \\ &= F(b|\mathbf{O_s}, \lambda) - F(-b|\mathbf{O_s}, \lambda) \\ P(\mathbf{x} = right|\mathbf{O_s}) &= 1 - P(\mathbf{x} = left|\mathbf{O_s}) \\ &\quad - P(\mathbf{x} = middle|\mathbf{O_s}) \end{aligned} \tag{11}$$

Therefore $P(\mathbf{x}|\mathbf{O_h}, \mathbf{O_s})$ can be inferred by $P(\mathbf{x}|\mathbf{O_h})$ and $P(\mathbf{x}|\mathbf{O_s})$ after normalization.

In the second stage, the coordination is based on the evaluation of the prediction in the previous stage. We use the difference of the entropy between the human original intention and the fused results to evaluate the prediction, which is denoted as D_p. A larger difference means better prediction and less uncertainty.

$$\begin{aligned} D_p = &\sum_{\mathbf{x}} P(\mathbf{x}|\mathbf{O_h}) \log P(\mathbf{x}|\mathbf{O_h}) \\ &- \sum_{\mathbf{x}} P(\mathbf{x}|\mathbf{O_h}, \mathbf{O_s}) \log P(\mathbf{x}|\mathbf{O_h}, \mathbf{O_s}) \end{aligned} \tag{12}$$

The final control output U_f is calculated as follows.

$$U_f = \begin{cases} U_h & D_p < 0 \\ (1-\alpha)U_h + \alpha U_s & D_p \geq 0 \end{cases} \tag{13}$$

U_h and U_s represent the control inputs which move the camera to a location based on the user's head movement and sound location. α is a coordination factor that defines the amount of control from the user which takes value in [0, 1]. $\alpha = 1$ gives the user full control and $\alpha = 0$ allows the robot to control itself. α is calculated using a sigmoid function of D_p for the purpose of smooth shift of the control weight.

$$\alpha = \frac{1}{1 + e^{-c_1 D_p + c_2}} \tag{14}$$

Although the remote operator can hear the voices from the local user, he/she cannot infer the sound location especially when there is no human subject in the field view of the camera. In this case, it is highly likely that there is a conflict between the human and the robot, which causes insufficient movements from the human operator. So we can draw an arrow in the 3D image displayed on the Oculus Rift to indicate the sound location which helps the operator find the local user quickly.

5 Experiments and Results

5.1 Experimental Setup

We evaluate the proposed two-stage collaborative control framework on our VR-based robotic telepresence system. Oculus Rift Development Kit 2 is used as the VR device which has a resolution of 1920 × 1080 per eye and a field of view of 90° horizontal and 110° vertical. We use the ASUS Xtion Pro Live sensor as the RGB-D camera. Due to the limited bandwidth, the images from the RGB-D camera are compressed, JPEG compression for the color image and PNG compression for the depth image respectively. The sound localiztion algorithm is implemented based on an open source audition software HARK [24]. All the data are synchronized and processed in real time.

5.2 HMMs-Based Intention Recognition

Firstly, we test the performance of HMM-based intention recognition. We define five simple intentions: *turn left*, *turn right*, *turn up*, *turn down* and *other (not moving)*. The training data are collected from two human users and manually labelled. During the training stage, the length of the sequence is not restricted. In the real-time testing stage, the length of the sequence is fixed to 20 observations using a slide window method. We also add a motion detection threshold to remove the noise. Four human users participated in the testing stage. Table 1 shows the statistical results. The HMM-based method can distinguish two intentions even part of the sequence is similar, such as turn left and turn right. Overall the accuracy is above 92%. Figure 10 shows the online recognition results. Both the ground truth and the recognition results are shown in the figure. Most of the intentions are recognized. There is a slight delay in the testing results compared to the ground truth because of the processing time. The recognition is quite robust and it can be easily extended to more intentions.

Table 1 Results of HMM-based intention recognition

Intentions	Recognized results					Accuracy
	Right	Left	Down	Up	Other	
Right	0.95	0.01	0	0	0.04	0.95
Left	0.03	0.92	0	0	0.05	0.92
Down	0	0	0.95	0.02	0.03	0.95
Up	0	0	0.02	0.96	0.02	0.96
Other	0.02	0.03	0.02	0.01	0.92	0.92

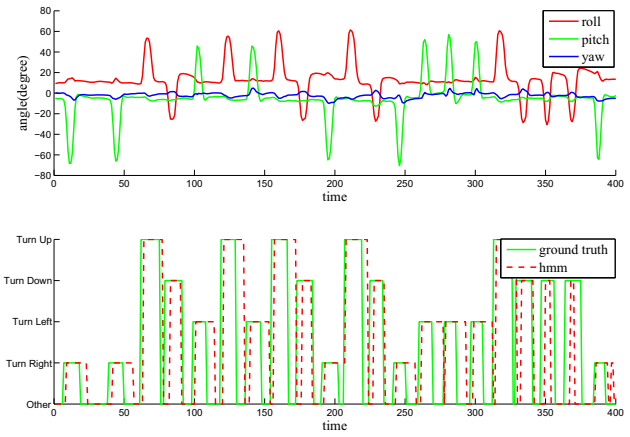

Fig. 10 The online recognition results using HMMs

5.3 Sound Localization

Sound localization is tested using a sound simulation system and a motion capture system (OptiTrack [33]). To fully evaluate the accuracy of the sound localization, the speaker is placed at different directions (0°, ±45°, ±90°, and ±135°) and distances (0.5 m, 1 m, 2 m and 3 m) with respect to the robot. The motion capture system obtains the relative locations between the speaker and the robot, which are treated as the ground truth. For each location, we run the sound localization algorithm 10 times and calculate the mean and the standard deviation which are shown in Table 2. From Table 2, we can see that the detection errors are small in the same distance and not very sensitive to the direction of the sound sources. However, the errors increase with the distance. The standard deviation of errors is less than 2° at 0.5 m and less than 4° at 3 m away from the robot. Based on the results, we set $1/\lambda$ to be 0.003 ($\sigma = 3°$) for the sound detection model as mentioned in Sect. 4.2.

5.4 Overall Experiments

The experiments consist of two scenarios of daily conversation: one remote user vs. one local user (1 vs. 1) and one remote user vs. two local users (1 vs. 2). The remote user wears the Oculus Rift to view the rendered 3D scene of the local site while having an audio-only Skype chatting with the local user(s). We assume only one local user speaks at any time. Ten participants were involved in the experiments. They are graduate students from School of ECE at Oklahoma State University. Most of the participants have limited experience of virtual reality devices. Before the experiment, a detailed instruction manual was provided to help the participants use the VR device. 1 vs. 2 conversation was conducted after 1 vs. 1 conversation. All the participants were involved in both scenarios. For each scenario, three trials were conducted for each participant. In the 1 vs. 1 conversation, one local user speaks to the telepresence robot from different angles with respect to the robot, while the remote user talks through the robot. The results are shown in Fig. 11. The first row is the speed of the head movement in the yaw direction. The second row is the result of sound localization. The third row is the decision output using the proposed collaborative control method where 1 means no sound is detected, 2 means the control is based on fusion which implies that the robot is assisting the user, 3 means the prediction is not good so the user takes over the control. The fourth row is the location of the motor which indicates the viewpoint of the camera. For most of the times, the robot provides assistance to the user when there is any sound detected. The remote user can reject the robot's control if there is a conflict with robot's prediction, such as the period from frame 19 to frame 97 in Fig. 11. In the 1 vs. 2 conversation, two local users stand at two different locations. The results are shown in Fig. 12. In the second row, the blue and red dots indicate different users, and the black dots represent the false detections. Although the sound location is quite accurate at most of the times, there are some noises that cause false detections. Even with false detection, our method can still handle the situation to avoid unnecessary movements. The results are similar to those of the 1 vs. 1 conversation. If the prediction of human intention is correct then the robot provides assistance. The human can take over the control if the robot's prediction is not correct.

Table 2 Results of sound localization

Distance	Errors	Direction				
		0°	±45°	±90°	±135°	Sum
0.5 m	Mean (°)	−0.3	−0.1	−0.2	0.2	−0.2
	Std (°)	1.5	2.0	1.9	1.6	1.7
1 m	Mean (°)	0.6	−0.8	−0.2	0.5	−0.1
	Std (°)	2.2	2.1	2.3	2.0	2.2
2 m	Mean (°)	0.1	0.2	−0.1	0.1	−0.3
	Std (°)	3.1	2.9	3.0	2.7	2.9
3 m	Mean (°)	1.8	0.3	−1.1	−0.9	0.5
	Std (°)	4.2	3.6	4.0	3.7	3.9

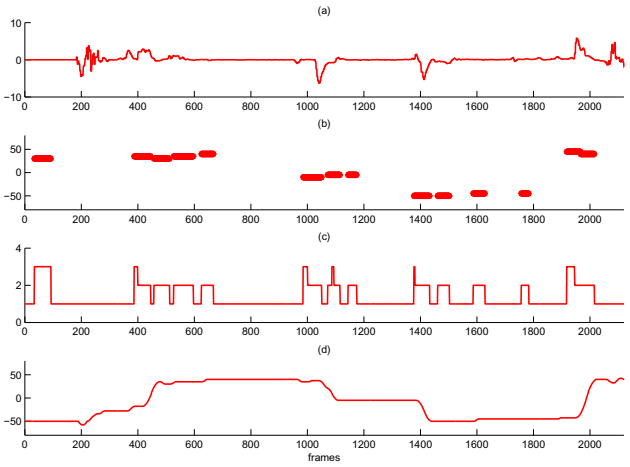

Fig. 11 The results of a 1 versus 1 conversation. **a** Speed of human head movement (°), **b** sound location (°), **c** decision output, **d** motor location (°)

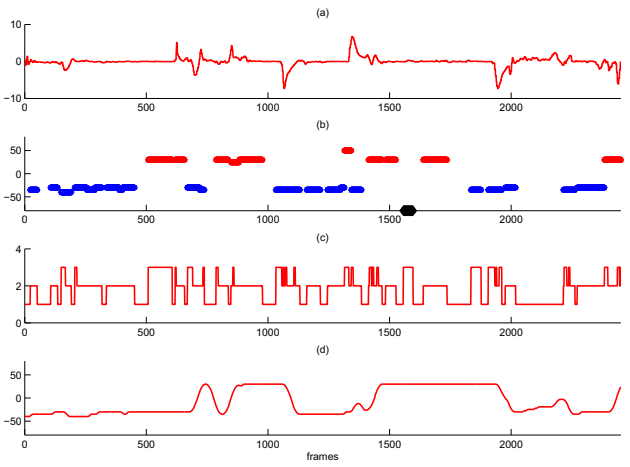

Fig. 12 The results of a 1 versus 2 conversation. **a** Speed of human head movement (°), **b** sound location (°), **c** decision output, **d** motor location (°)

Then we evaluate the system in a scenario when the local user does not appear in the field of view of the camera at the beginning. The goal is to let the remote user find the local user who is speaking. A hint in the form of a red arrow will show up on the display to indicate the sound location as shown in Fig. 13 (Top). The experiments are conducted on 10 participants in a total of 18 trials using both the "teleoperated" method and the proposed collaborative control method. The average search time is 15.83*s* for the "teleoperated" method and 7.89*s* for the collaborative control method. With the hint of the sound location, the search time is significantly reduced. We also conduct a user study with the same questions mentioned in Sect. 4.1. The results are shown in Fig. 14. The proposed collaborative control method has the highest score for convenience and a similar score in workload compared with the "sound-guided" method. The manipulability on the remote user side is improved with more convenience and less workload, while the engagement and embodiment on the local user side are also maintained.

Fig. 13 An example of showing hint in a human search experiment. Top: Hint (red arrows) appears in the search; bottom: human subject is found. (Color figure online)

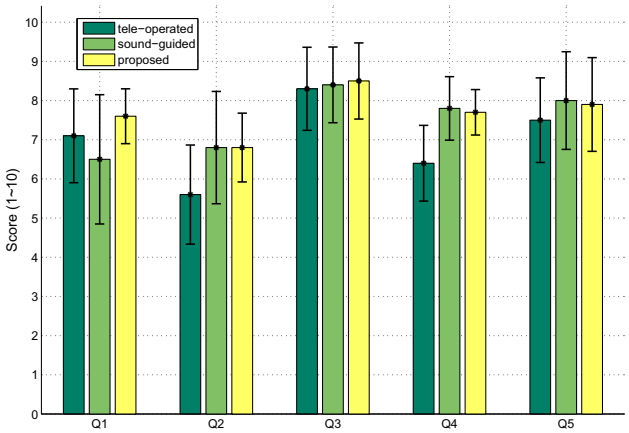

Fig. 14 The results of the user study. Q1 (Convenience), Q2 (Workload), Q3 (Immersiveness), Q4 (Engagement) and Q5 (Embodiment)

6 Conclusions and Future Work

In this paper, we develop a virtual-reality-based robotic telepresence system with both visual and audio input that aims to improve the user experience in telepresence. We adopt an

Oculus Rift goggle as the interface that connects the user to the robot to provide immersive feelings and infer head movement intentions. Sound localization is implemented with a microphone array to allow the robot to track the speaker in a conversation. Based on the findings of a user study, a human-robot collaborative framework is proposed to integrate the remote user's control and the robot's control. The experimental results show that the proposed system can provide better user experience for telepresence tasks. The proposed collaborative control framework can be applied to other human-machine coupled control systems. The proposed VR-based telepresence system has great potential in many social robot applications. The future work will focus on further improving the user friendliness of the robot and conducting more experiments with more human subjects as well as collecting more side-effect data, including fatigue, control frustration, etc. We will also consider the application of such telepresence robots in elderly care and telemedicine.

Compliance with Ethical Standards

Conflict of interest This material is based upon work supported by the National Science Foundation under Grant Nos. CISE/IIS 1427345, CISE/IIS 1910993, EHR/DUE 1928711, and CISE/IIS/SCH 1838808. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.This project is also partially supported by the Natural Science Foundation of China-Shenzhen Basic Research Center Project No. U1713216 and the Open Research Project of the State Key Laboratory of Industrial Control Technology, Zhejiang University, China (No. ICT20026). The authors declare that they have no conflict of interest.

References

1. Goodrich MA, Schultz AC (2007) Human-robot interaction: a survey. Found Trends Hum-Comput Interact 1(3):203
2. Tsui KM, Desai M, Yanco H, Uhlik C et al (2011) Exploring use cases for telepresence robots. In: Human–Robot Interaction (HRI), 2011 6th ACM/IEEE international conference on. IEEE, pp 11–18
3. Kristoffersson A, Coradeschi S, Loutfi A (2013) A review of mobile robotic telepresence. Adv Hum-Comput Interact 2013:3
4. Paulos E, Canny J (2001) Social tele-embodiment: understanding presence. Auton Robots 11(1):87
5. Michaud F, Boissy P, Labonte D, Corriveau H, Grant A, Lauria M, Cloutier R, Roux MA, Iannuzzi D, Royer MP (2007) Telepresence robot for home care assistance. In: AAAI Spring symposium: multidisciplinary collaboration for socially assistive robotics, California, USA, pp 50–55
6. Google Cardboard (2015). https://www.google.com/get/cardboard/
7. Oculus Rift (2015). https://www.oculus.com
8. Adalgeirsson SO, Breazeal C (2010) Mebot: a robotic platform for socially embodied presence. In: Proceedings of the 5th ACM/IEEE international conference on Human–robot interaction. IEEE Press, pp 15–22
9. Escolano C, Antelis JM, Minguez J (2012) A telepresence mobile robot controlled with a noninvasive brain–computer interface. IEEE Trans Syst Man Cybern Part B Cybern 42(3):793
10. Sirkin D, Ju W (2012) Consistency in physical and on-screen action improves perceptions of telepresence robots. In: Proceedings of the seventh annual ACM/IEEE international conference on human–robot interaction. ACM, pp 57–64
11. Martins H, Oakley I, Ventura R (2014) Design and evaluation of a head-mounted display for immersive 3D teleoperation of field robots. Robotica 33:2166
12. Kratz S, Vaughan J, Mizutani R, Kimber D (2015) Evaluating stereoscopic video with head tracking for immersive teleoperation of mobile telepresence robots. In: Proceedings of the tenth annual ACM/IEEE international conference on human–robot interaction extended abstracts. ACM, pp 43–44
13. Kato Y (2015) A remote navigation system for a simple telepresence robot with virtual reality. In: 2015 IEEE/RSJ international conference on Intelligent Robots and Systems (IROS). IEEE, pp 4524–4529
14. DORA (2020). http://doraplatform.com
15. Stiefelhagen R, Yang J, Waibel A (2001) Estimating focus of attention based on gaze and sound. In: Proceedings of the 2001 workshop on Perceptive user interfaces, ACM, pp 1–9
16. Gao M, Oberlander J, Schamm T, Zollner JM (2014) Contextual task-aware shared autonomy for assistive mobile robot teleoperation. In: 2014 IEEE/RSJ international conference on Intelligent Robots and Systems (IROS 2014). IEEE , pp 3311–3318
17. Carlson T, Monnard G, del R. Millán J (2011) Vision-based shared control for a BCI wheelchair. Int J Bioelectromagn 13(EPFL–ARTICLE–168977):20
18. Almeida L, Patrao B,Menezes P, Dias J (2014) Be the robot: Human embodiment in tele-operation driving tasks. In: 2014 RO-MAN: the 23rd IEEE international symposium on robot and human interactive communication. IEEE, pp 477–482
19. Pioneer P3-DX (2016). http://www.mobilerobots.com/ResearchRobots/PioneerP3DX.aspx
20. Intel NUC (2015). http://www.intel.com/content/www/us/en/nuc/overview.html
21. OpenGL (2015). https://www.opengl.org
22. G.R.A.S Microphones (2015). http://www.gras.dk
23. NI USB-9234 (2015). http://sine.ni.com/nips/cds/view/p/lang/en/nid/204481
24. Nakadai K, Takahashi T, Okuno HG, Nakajima H, Hasegawa Y, Tsujino H (2010) Design and implementation of robot audition system 'HARK'-open source software for listening to three simultaneous speakers. Adv Robot 24(5–6):739
25. Nakamura K, Nakadai K, Asano F, Ince G (2011) Intelligent sound source localization and its application to multimodal human tracking. In: 2011 IEEE/RSJ international conference on Intelligent Robots and Systems. IEEE, pp 143–148
26. Okuno HG, Nakadai K, Takahashi T, Takeda R, Nakamura K, Mizumoto T, Yoshida T, Lim A, Otsuka T, Nagira K, Itohara T, Bando Y (2014) Hark document version 2.1.0. Tech. rep., Kyoto University
27. Nakadai K, Ince G, Nakamura K, Nakajima H (2012) Robot audition for dynamic environments. In: 2012 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC). IEEE, pp 125–130
28. Nakamura K, Nakadai K, Asano F, Hasegawa Y, Tsujino H (2009) Intelligent sound source localization for dynamic environments. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009. IROS 2009. IEEE, pp 664–669
29. Elmenreich W (2002) An introduction to sensor fusion. Tech. report, Vienna University of Technology, Austria 502:1–28
30. Eickeler S, Kosmala A, Rigoll G (1998) Hidden markov model based continuous online gesture recognition. In: Proceedings. Four-

teenth International Conference on Pattern Recognition (Cat. No. 98EX170), 1998. Proceedings, vol 2. IEEE, pp 1206–1208
31. Fine S, Singer Y, Tishby N (1998) The hierarchical hidden Markov model: analysis and applications. Machi Learn 32(1):41
32. Stephens M (1969) Tests for the von Mises distribution. Biometrika 56(1):149
33. OptiTrack (2016). http://optitrack.com