2011 IEEE International Conference on Robotics and Automation
Shanghai International Conference Center
May 9-13, 2011, Shanghai, China

Design and Implementation of Selectable Sound Separation on the Texai Telepresence System using HARK

Takeshi Mizumoto, Kazuhiro Nakadai, Takami Yoshida, Ryu Takeda,
Takuma Otsuka, Toru Takahashi and Hiroshi G. Okuno

Abstract— **This paper presents the design and implementation of selectable sound separation functions on the telepresence system "Texai" using the robot audition software "HARK." An operator of Texai can "walk" around a faraway office to attend a meeting or talk with people through video-conference instead of meeting in person. With a normal microphone, the operator has difficulty recognizing the auditory scene of the Texai, e.g., he/she cannot know the number and the locations of sounds. To solve this problem, we design selectable sound separation functions with 8 microphones in two modes, overview and filter modes, and implement them using HARK's sound source localization and separation. The overview mode visualizes the direction-of-arrival of surrounding sounds, while the filter mode provides sounds that originate from the range of directions he/she specifies. The functions enable the operator to be aware of a sound even if it comes from behind the Texai, and to concentrate on a particular sound. The design and implementation was completed in five days due to the portability of HARK. Experimental evaluations with actual and simulated data show that the resulting system localizes sound sources with a tolerance of 5 degrees.**

I. Introduction

Recent globalization of business and improvements of transportation speed has produced the situation where people in different places or in different countries work together. However, communicating with people in distant places is difficult because modalities are limited; for example, phones only use voice, and video-conference systems are limited to a particular room. Such limitations make for less "presence" at the distant place, which leads to misunderstanding. To increase remote presence, a telepresence robot is one of the promising methods for rich communication, thanks to its both mobility and video-conference system. Currently, a wide variety of telepresence robots are available [1].

Current telepresence robots, however, are limited in providing auditory scene awareness. An operator of such a robot is incapable of localizing where sound comes from, and concentrating on particular talker. In other words, the current telepresence robot lacks a capability that provides the so-called "cocktail-party effect" [2]. It shows that humans have the ability to selectively attend to a sound from a particular source, even when it is interfered with by other sounds.

T. Mizumoto, T. Otsuka, R. Takeda, T. Takahashi and H. G. Okuno are with Graduate School of Informatics, Kyoto University, Sakyo, Kyoto 606-8501, Japan. {mizumoto, ohtsuka, rtakeda, tall, okuno}@kuis.kyoto-u.ac.jp

K. Nakadai and T. Yoshida are with Tokyo Institute of Technology, 2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan. K. Nakadai is also with Honda Research Institute Japan Co., Ltd, 8-1 Honcho, Wako-shi, Saitama 351-0114, Japan. nakadai@jp.honda-ri.com, yoshida@cyb.mei.titech.ac.jp

Fig. 1. Three people and one Texai talking around an audition-enhanced Texai in California. In this snapshot, two people are talking together, while the third person is talking to the Texai of which operator is in Illinois.

However, the cocktail-party effect is insufficient from the viewpoint of auditory scene awareness because it gives only a partial aspect of the auditory scene instead of giving an overview.

Auditory scene can be reproduced with high-fidelity by using a user-like dummy head moulded by the subject's head [3]. Since the impulse response of the head is almost the same between human's original head and its dummy head, the acoustic signals captured by the dummy head can be reproduced accurately at the human's head through headphone. Since people can listen to at most two things simultaneously according to psychophysics [4], such a dummy head may not enhance auditory scene awareness.

Auditory scene awareness is enhanced by computational auditory scene analysis (CASA) [5], since it focuses on sound source localization, separation and recognition of separated sounds given a mixture of sounds. The robot audition open-source software "HARK" is designed as *an audition-equivalent of OpenCV* to provide various functions requested by CASA [6]. Kubota et al. [7] designed and implemented a 3-D visualizer called "CASA Visualizer" for HARK outputs. The CASA visualizer displays the direction-of-arrival of sound sources and can replay each separated sound both on-line and off-line. It can also display the subtitles for separated voiced sounds off-line.

The CASA visualizer has three modes based on the visual information seeking mantra, that is, *"overview first, zoom and filter, then details on demand"* [8]. "Overview first" provides the temporal overview of the auditory scene by showing the direction of each sound. "Zoom and filter" provides the presence of sound sources at a particular time.

"Details on demand" provides information about a specific sound source by playing back the relevant sound. To give the operator auditory awareness, we applied HARK to a telepresence system to implement the *selectable sound separation system* on it.

From March 15th to 19th, 2010, we visited the robotics company Willow Garage, Inc., which has been developing a telepresence system named Texai [9], to implement a system which gives an operator auditory awareness. In these five days, we developed a *selectable sound separation* system for **an audition-enhanced Texai** (see Figure 1 for overview). It has two functions:

1) visualizing the existence and the direction of sound around Texai and
2) selecting a directional range to listen to.

Using the first function, an operator of Texai can be aware of a sound even if it comes from behind Texai. Using the second one, the operator can listen to a particular person's sound even if multiple people are talking, by specifying the directional range of interest. Thanks to the portability of HARK, we were able to implement *selectable sound separation* on Texai in only five days. The demonstration video of our system is available on YouTube [1].

This paper is organized as follows: Section II overviews the platform, Texai. Section III describes the selectable sound separation system including the problem, implementation, and overview of HARK. In Section IV, we show our preliminary evaluation of the system and an example of our system, Then, Section VI concludes the paper and discusses about our future work.

II. Overview of Texai and HARK

A. Equipments of Texai

Texai, a telepresence system developed by Willow Garage, Inc., consists mainly of two cameras (a pan-tilt one for looking at a remote place and a wide-angle one for navigation), a stereo microphone and a stereo loudspeaker, a color LCD screen, and two motors for mobility. As shown in Figure 1, people can talk with each other as if they were in the same room. This is achieved because Texai can input and output both audio and visual information.

B. Communication between Texai and remote computer

Figure 2 shows the data flow during a conference through Texai. Using a video-conference software over the Internet, not only motor commands for Texai but also audio and visual information are sent between the Texai and remote computer. Therefore, an operator at a remote computer can use Texai wherever a wireless Internet connection is available.

C. Robot Operating System (ROS)

Texai is controlled with the open-source robot operating system called "ROS" [10] also developed by Willow Garage, Inc. ROS is a meta-operating system for robots, which provides functionality from hardware abstraction to

[1]http://www.youtube.com/watch?v=xpjPun7Owxg

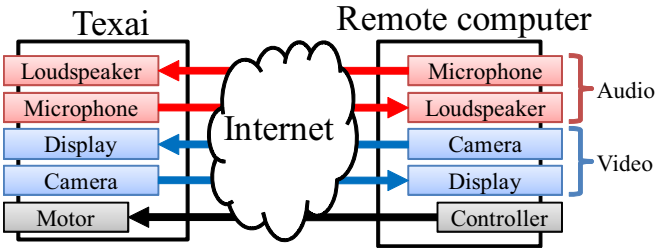


Fig. 2. Data flow of Texai: Audio and Visual information are exchanged between Texai and remote computer through the Internet

message passing between processes. We can easily extend the functions of Texai because ROS is highly modular.

A node and *a topic* are two important keywords to understand ROS. A node is an executable program that communicates with other nodes by sending a topic. A topic is a structure of data defined by ROS users whose structure consists of, for example, strings, integers. When a node *publishes* a topic, it is broadcasted to any node which *subscribes* the topic. Thanks to this structure, each node can concentrate on *publishing* and *subscribing* topics, instead of considering the communication with other nodes like inter process communication.

D. HARK robot audition software

HARK, developed by us, provides various signal processing modules ranging from sound source localization, sound source separation, and recognition of separated sounds on the middleware called "FlowDesigner". We only explain functions needed in implementing an audition-enhanced Texai.

1) Sound source localization: Given the number of sound sources, MUltiple Signal Classification (MUSIC) localizes multiple sound sources robustly in real environments.
2) Sound source separation: Geometrically constrained High-order Decorrelation based Source Separation (GHDSS) [11] is an adaptive frequency-domain blind source separation algorithm. Given the directions of sound source, GHDSS separates corresponding sound sources that originate from the specified direction.

III. Selectable Sound Separation on Texai

A. Problems with Current System and Our Approach

Although Texai achieves one-to-one remote communication, a problem arises when an operator tries to talk with multiple people. It is hard for the Texai operator to:

1) know where a particular sound comes from, and
2) clearly distinguish a particular sound.

As mentioned above, people still have difficulty in recognizing more than two sound sources although people may disambiguate sound source localization by moving their head and focus on a particular sound through the cocktail party effect.

To solve this problem, we implement two functions: (1) visualizing the direction-of-arrival of sounds and (2) sending

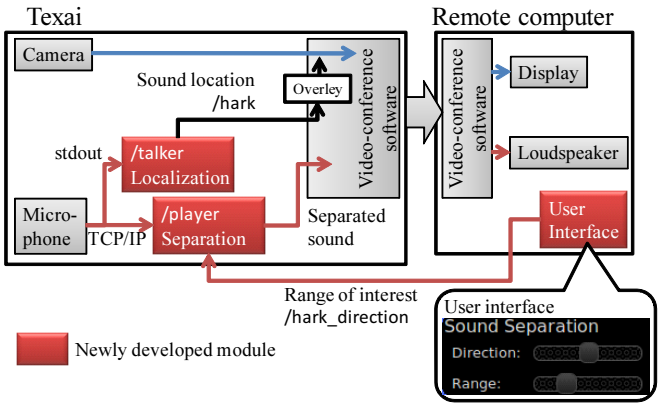


Fig. 3. Block diagram of selectable sound separation on Texai

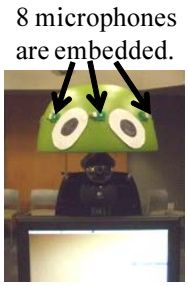


Fig. 4. Head of Audition-enhanced Texai: A bamboo bowl embedded with 8-channel microphone array

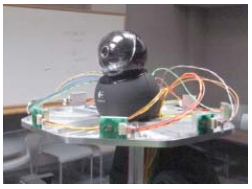


Fig. 5. First version of head: An aliminium disk with 8-channel microphone array on its edge

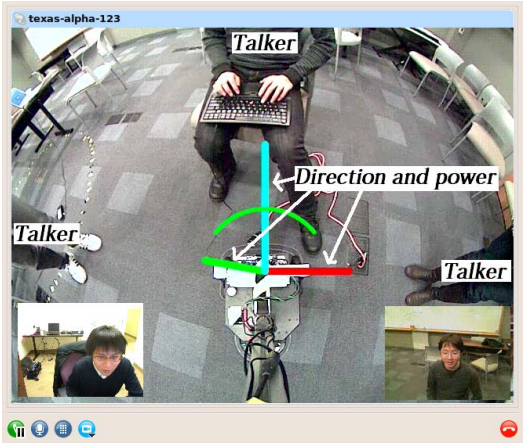


Fig. 6. GUI Interface for Remote Operator: The directions of sound are overlaid with arrows on the video, and the operator specifies the range of directions of sound sources to listen to.

a separated sound to the remote operator. We use HARK to implement the functions of sound source localization and separation from a mixture of sounds.

B. Overview of Selectable Sound Separation

Figure 3 shows a block diagram of our selectable sound separation system based on HARK. The gray boxes are original modules of Texai, the red boxes are newly developed nodes under ROS. We replaced Texai's microphones with a bowl embedded with 8-channel microphone array (see Figure 4) because HARK needs a microphone array processing for sound source localization and separation.

The system works as follows: Through a video camera and microphones, the operator looks at and listens to the remote situation around Texai. When a person talks to Texai, the Localization module detects the direction of the sound, and the /talker node publishes a topic /hark, which consists of time stamp, id, direction-of-arrival, and its power. Then, the video conference subscribes the topic and overlays (superimposes) on the video as shown in Figure 6. The direction and the length of line in the center of Figure 6 denotes the direction and the volume of talker, respectively.

Next, using two slide bars as shown in the right bottom of Figure3, the operator specifies two parameters: (1) the center direction of the range to listen to, and (2) the angular width of the range, as shown in the center of Figure 6. From the parameters, the user interface publishes a topic /hark_direction which consists of the beginning and the ending angles of user's interest. Then, a remote user listens to only the sounds from the specified range.

C. Integration of HARK and Texai

We here describe how we connect the localization and separation programs made with HARK with ROS on Texai. We developed two ROS nodes, talker and player for sound source localization and separation, respectively, as shown in Figure 3. These nodes use two ways for connecting HARK with ROS: talker node steals the standard output (stdout) of HARK, and player connects with HARK through TCP/IP.

The talker runs a sound source localization program made with HARK as a subprocess. Then, talker steals its standard output, analyzes the directions of sounds produced by localization program, and publishes a topic named hark.

The node player and a sound source separation program with HARK run independently. The HARK program sends both a separated sound and corresponding directional information to player through TCP/IP. On the other hand, player subscribes a topic hark_direction, which consists of the beginning and the ending angles of directional range of user's interest, which topic is published from a remote computer. player checks the direction of the separated sound from HARK program. If the direction is within the specified range by hark_direction, it is sent to the remote user through the Video-conference system.

D. Sound source localization and separation with HARK

1) Model of sound signal: We model the signals from sound sources to microphones at first. Suppose that there are M sources and N ($\geq M$) microphones. A spectrum vector of M sources at frequency ω, $\mathbf{s}(\omega)$, is denoted as $[s_1(\omega)\ s_2(\omega)\ \cdots\ s_M(\omega)]^T$, and a spectrum vector of signals captured by the N microphones at frequency ω, $\mathbf{x}(\omega)$, is denoted as $[x_1(\omega)\ x_2(\omega)\ \cdots\ x_N(\omega)]^T$, where T represents a transpose operator. $\mathbf{x}(\omega)$ is, then, calculated as

$$\mathbf{x}(\omega) = \mathbf{H}(\omega)\mathbf{s}(\omega) + \mathbf{N}(\omega), \quad (1)$$

where $\mathbf{H}(\omega)$ is a transfer function (TF) matrix. Each component H_{nm} of the TF matrix represents the TF from the m-th source to the n-th microphone. $\mathbf{N}(\omega)$ denotes a Gaussian noise vector.

2) Sound localization: We are using *MUltiple SIgnal Classification (MUSIC)* based on Standard Eigen Value Decomposition (SEVD) for sound source localization.

a) EVD of observed signal vector: The spatial correlation matrix is defined in each frequency independently as

$$\mathbf{R}(\omega) = \mathrm{E}[\mathbf{x}(\omega)\mathbf{x}^H(\omega)] \tag{2}$$

where $\mathrm{E}[\cdot]$ represents the expectation operator among some frames and H represents the conjugate transpose operator.

The eigenvalue decomposition of $\mathbf{R}(\omega)$ is

$$\mathbf{R}(\omega) = \mathbf{E}(\omega)\mathbf{\Lambda}(\omega)\mathbf{E}^{-1}(\omega). \tag{3}$$

Here, $\mathbf{E}(\omega)$ denotes the eigenvector matrix, the columns of which consist of the eigenvectors of $\mathbf{R}(\omega)$ as $\mathbf{E}(\omega) = [\mathbf{e}_1(\omega)\ \mathbf{e}_2(\omega)\ \cdots\ \mathbf{e}_N(\omega)]$. The matrix $\mathbf{\Lambda}(\omega) = \mathrm{diag}(\lambda_1(\omega), \lambda_2(\omega), \cdots, \lambda_N(\omega))$ represents the eigenvalue matrix in descending order, the diagonal elements of which consist of the eigenvalues of $\mathbf{R}(\omega)$.

Since λ_m represents the power of each sound, λ_i and $\mathbf{e}_i$ where $1 \leq i \leq M$ are the eigenvalues and vectors in terms of the sound sources, and λ_i and $\mathbf{e}_i$ where $M+1 \leq i \leq N$ are those of noise. Since we cannot know the number of sound sources in advance, we have no choice but to use the temporal number of sound sources L in practical use.

b) MUSIC Estimator: The spatial spectrum for localization is defined as

$$P(\omega, \phi) = \frac{|\mathbf{a}_\phi^H(\omega)\mathbf{a}_\phi(\omega)|}{\sum_{n=L+1}^{N} |\mathbf{a}_\phi^H(\omega)\mathbf{e}_m|} \tag{4}$$

where $\mathbf{a}_\phi(\omega) = [a_{\phi,1}(\omega)\ a_{\phi,2}(\omega)\ \cdots\ a_{\phi,N}(\omega)]$ represents a TF that was recorded in advance, and ϕ indicates the index of position. Thus, when the direction of steering vector $\mathbf{a}_\phi(\omega)$ and that of a sound source are the same, $P(\omega, \phi)$ theoretically becomes infinity. Therefore, MUSIC provides easy detectable and reliable peaks and has been used for sound source localization on robots.

Finally, we can integrate the spatial spectrum $P(\omega, \phi)$ from ω_{min} to ω_{max} because we treat a broad-band signal. The criteria $P(\phi)$ is defined with eigenvalues at each frequency to consider the power of frequency components as

$$P(\phi) = \sum_{\omega=\omega_{min}}^{\omega_{max}} \sqrt{\lambda_1(\omega)} P(\omega, \phi) \tag{5}$$

where $\lambda_1(\omega)$ is a maximum eigenvalue at frequency ω.

Note that we must decide three parameters, L, ω_{min} and ω_{max} in advance with this method.

3) Sound source separation: We use sound source separation using high-order information called *Geometrically constrained High-order Decorrelation based Source Separation (GHDSS)* [11].

a) Online GHDSS: Source separation is formulated as

$$\mathbf{y}(\omega) = \mathbf{W}(\omega)\mathbf{x}(\omega), \tag{6}$$

where $\mathbf{W}(\omega)$ is called a *separation matrix*. The separation with the general SSS is defined as finding $\mathbf{W}(\omega)$ which satisfies the condition that output signal $\mathbf{y}(\omega)$ is the same as $\mathbf{s}(\omega)$. Since SSS is done at each frequency independently, we skip denoting ω for readability.

In order to estimate $\mathbf{W}$, GHDSS also introduces two cost functions like GSS, that is, separation sharpness (J_{SS}) and geometric constraints (J_{GC}) defined by

$$J_{SS}(\mathbf{W}) = \|\phi(\mathbf{y})\mathbf{y}^H - \mathrm{diag}[\phi(\mathbf{y})\mathbf{y}^H]\|^2 \tag{7}$$

$$J_{GC}(\mathbf{W}) = \|\mathrm{diag}[\mathbf{WA} - \mathbf{I}]\|^2 \tag{8}$$

where $\|\cdot\|^2$ indicates the Frobenius norm, and $\mathrm{diag}[\cdot]$ is the diagonal operator. The expectation operator is not in eq. (7) because $\mathbf{W}$ is estimated frame-by-frame for realtime estimation. $\mathbf{A}$ is a TF matrix which consists of L TF vectors $\mathbf{a}_\phi$, that is, $\mathbf{A} = [\mathbf{a}_{\phi_1}\mathbf{a}_{\phi_2}\cdots\mathbf{a}_{\phi_L}]$. $\phi(\mathbf{y})$ is a nonlinear function defined as $\phi(\mathbf{y}) = [\phi(y_1), \phi(y_2), \cdots, \phi(y_N)]^T$ and

$$\phi(y_i) = -\frac{\partial}{\partial y_i} \log p(y_i) \tag{9}$$

The function is introduced to consider the higher-order statistics of the signal. There are a variety of definitions for $\phi(y_i)$. In this paper, we selected a hyperbolic-tangent-based function [12]:

$$\phi(y_i) = \tanh(\eta|y_i|)e^{j\cdot\theta(y_i)}, \tag{10}$$

where η is the scaling parameter.

The total cost function $J(\mathbf{W})$ is represented as

$$J(\mathbf{W}) = \alpha J_{SS}(\mathbf{W}) + J_{GC}(\mathbf{W}), \tag{11}$$

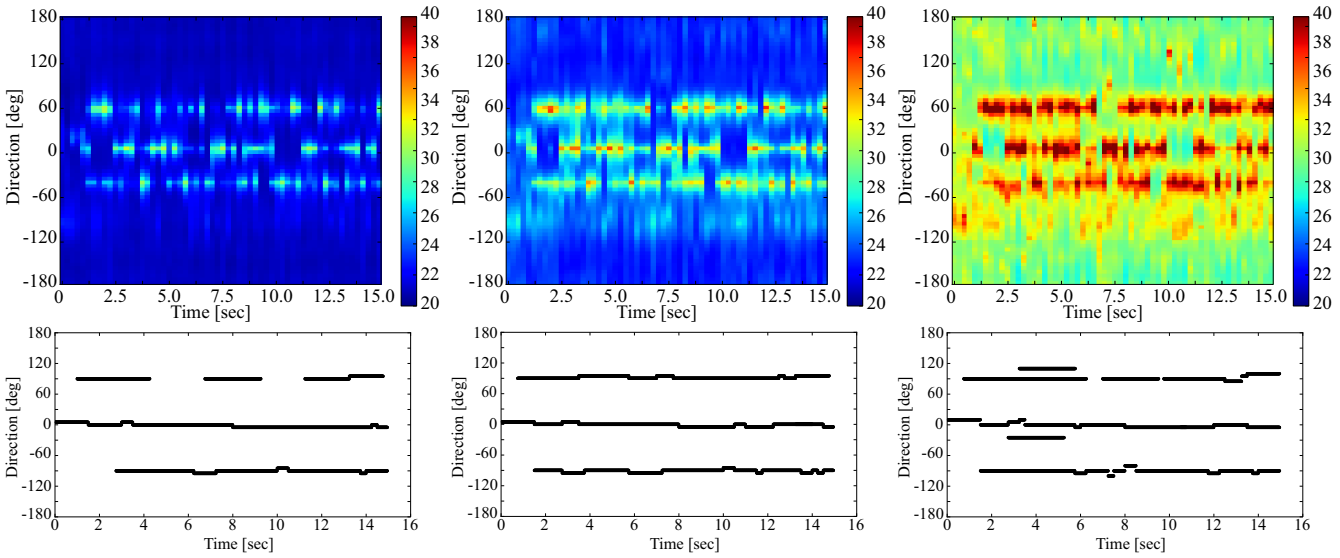
where α means the weight parameter between the costs of separation and geometric constraint.

When a long sequence of $\mathbf{x}$ can be used, we can directly estimate the best $\mathbf{W}$ by minimizing $J(\mathbf{W})$ in an offline manner. However, a robot needs to work in real time, and the best $\mathbf{W}$ is always changing in the real world. Thus, the online GSS adaptively updates $\mathbf{W}$ by using

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \mu_{SS}\mathbf{J'}_{SS}(\mathbf{W}_t) + \mu_{GC}\mathbf{J'}_{GC}(\mathbf{W}_t). \tag{12}$$

where $\mathbf{W}_t$ denotes $\mathbf{W}$ at the current time step t, $\mathbf{J'_{SS}}(\mathbf{W})$ and $\mathbf{J'_{GC}}(\mathbf{W})$ are complex gradients [13] of $\mathbf{J_{SS}}(\mathbf{W})$ and $\mathbf{J_{GC}}(\mathbf{W})$, which decide an update direction of $\mathbf{W}$. μ_{SS} and μ_{GC} are called step-size parameters.

b) Adaptive Step-size control (AS): Adaptive Step-size (AS) [11] is applied to control both μ_{SS} and μ_{GC} optimally. With this method, these step-size parameters become large values when a separation error is high, for example, due to source position changes. These will have small values when the error is small due to the convergence of the separation matrix. Thus, step-size parameters are automatically controlled to be optimal values.



(a) Underestimated: L=1 THRESH=24.6 (b) Exactly-estimated:L=3, THRESH=29.4 (c) Overestimated: L=5 THRESH=38.8

Fig. 7. MUSIC spectrograms and sound localization results of triple talkers are affected by the number of talkers L given in advance; (a) underestimated, (b) exactly-estimated, and (c) overestimated. Three talkers are at -60, 0, 60 degrees, respectively. For each figures, The horizontal and vertical axes denote time and the direction of sounds, respectively. The color denotes the power of MUSIC spectrum defined in Eq. 5 in dB.

By using our AS, Eq. (12) is redefined as

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \mu_{SS}\mathbf{J}'_{SS}(\mathbf{W}_t) - \mu_{GC}\mathbf{J}'_{GC}(\mathbf{W}_t), \quad (13)$$

$$\mu_{SS} = \frac{\|\phi(\mathbf{y})\mathbf{y}^H - \mathrm{diag}[\phi(\mathbf{y})\mathbf{y}^H]\|^2}{8\|\phi(\mathbf{y})\mathbf{y}^H - \mathrm{diag}[\phi(\mathbf{y})\mathbf{y}^H]\tilde{\phi}(\mathbf{y})\mathbf{x}^H\|^2}$$

$$\mu_{GC} = \frac{\|\mathrm{diag}[\mathbf{WA} - \mathbf{I}]\|^2}{8\|\mathrm{diag}[\mathbf{WA} - \mathbf{I}]\mathbf{A}^H\|^2},$$

$$\tilde{\phi}(\mathbf{y}) = [\tilde{\phi}(y_1), \tilde{\phi}(y_2), ..., \tilde{\phi}(y_M)], \quad (14)$$

$$\tilde{\phi}(y_k) = \phi(y_k) + y_k \frac{\partial \phi(y_k)}{\partial y_k}. \quad (15)$$

μ_{SS} and μ_{GC} become large values when a separation error is high, for example, due to source position changes. It will be low when the error is small due to the convergence of the separation matrix. Thus, step-size and weight parameters are controlled optimally at the same time.

IV. Experiments

This section describes the evaluation of the localization performance. Note that we used an actual talker instead of loudspeakers due to a shortage of available equipments. Therefore, the volume of the talkers is different for each trial.

We conduct three experiments: (1) In IV-B, we evaluate the MUSIC spectrum defined in Eq. (5) and the localization performance when the number of talkers L in Eq. 4 is incorrect. As we mentioned in III-D.2.b, it is difficult to give system a correct L because the number of people around the audition-enhanced Texai changes dynamically. Therefore, we investigate the robustness against such a incorrect setting. (2) In IV-C, we evaluate the localization performance by varying the following conditions: the number and the interval of talkers, the rooms, the background noise level, and the distance between talkers and the Texai. (3) In IV-D, we demonstrate how the entire system works using four talkers' actual conversation. Here, we show not only the localization result, but also an example of separated sound.

A. Experimental Conditions

We used Texai with 8 microphones on an off-the-shelf bowl. We conducted all experiments in two rooms called Dining and Cathedral. One of the walls in the Dining is made of glass. Dining is larger than Cathedral. Sounds are recorded using a multi-channel recording system RASP[2]. For the localization, the number of frequency bins are 512, and 172 bins, which the frequency components from 500 Hz to 2800 Hz, are used. The source location is estimated using a MUSIC spectrum which is averaged for 25 frames. Therefore, the localization is executed for each 250 [msec]. The number of sources L, i.e., that of talkers, is determined in advance because it is controllable.

Prior to evaluating our system, we investigated the best performance in simulation using two sets of impulse responses measured in Dining and Cathedral, and two kinds of microphone array shown in Figures 4 and 5. Both microphone arrays use the same MEMS microphones. By this preliminary configuration, we optimized th `THRESH` parameter in a hark module `SourceTracker`. This parameter determines whether the localized sound is noise or not by checking if the power of the sound exceeds the parameter.

B. Experiment 1: Performance under Incorrect Parameters

This experiment investigates how the localization performance is when the given and actual number of talkers, L, is incorrect. Therefore, we need to investigate what happens when the parameter is different from the actual situation. The recording condition is as follows: recorded in Cathedral, three talkers at an interval of 60° with background noise. We

[2]http://jeol-st.com/mt/2007/04/rasp2.html (in Japanese)

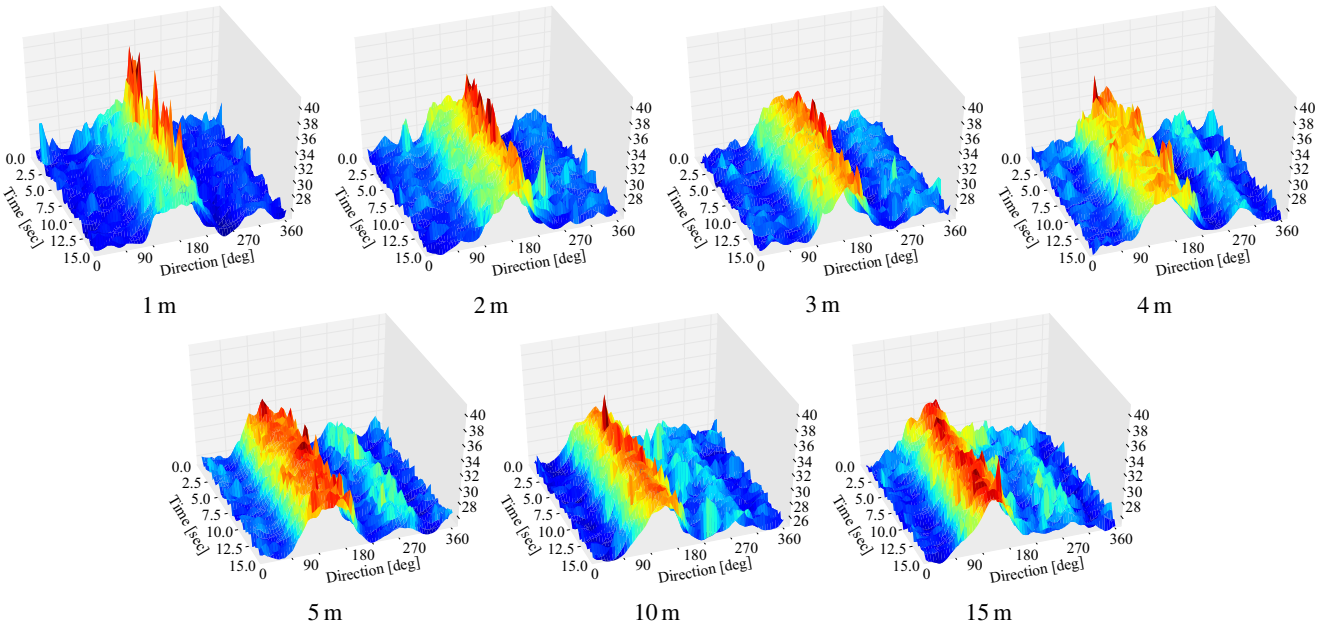


Fig. 8. MUSIC spectrogram of a single talker: As the distance between the talker and Texai becomes longer, the peak becomes smoother. The horizontal axis of each figure denotes the power of MUSIC spectrogram in dB.

localize the mixture of sound in three kinds of parameters, $L = 1$, 3 and 5. $L = 1$ means that the number of talkers is underestimated, $L = 3$ means that the number is exactly estimated, and $L = 5$ means it is overestimated.

Figure 7 shows the result. Figure 7(a), (b), and (c) corresponds to the underestimated, exactly-estimated, and overestimated conditions, respectively. For each conditions, the upper figure is the MUSIC spectrum, and the lower one is the result of sound source localization. The MUSIC spectrums in Figure 7(a) are broken into short pieces although each talker speaks continuously. However, the overestimated condition (c) shows that the not only talkers' voice but also noise are enhanced, as shown in 120° at 10 sec, or -90° from 0 sec to 15 sec. In spite of the sensitivity to the number of talkers, we can modify the parameter of the module for tracking the location, `THRESH` for each conditions. According to the result, we conclude that we maintain a proper the localization performance by modifying the parameter of tracking. even when $M \neq L$.

C. Experiment 2: Localization performance

In this experiment, we evaluate the stability of localization under following conditions: the number and interval of talkers, the rooms, the level of noise, and the distance between talkers and the Texai. Note that the impulse responses for sound source localization are measured every 5 degrees at only the distance of 1 m from Texai, in advance.

Table I shows the standard deviation of the localization, which corresponds to the fluctuation of the localization. We change the distance between the talker and the Texai from 1 m to 5 m. We additionally used the distance of 10 and 15 m in Dining because the room is wide enough. In Dining, the standard deviation is low under with-background-music condition compared with without-background-music condition. This is because the subjects spoke louder with background music than without it.

TABLE I

STANDARD DEVIATION OF LOCALIZATION WITH ONE TALKER

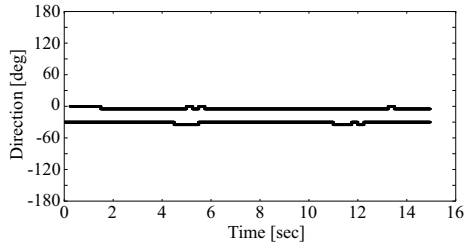
room*	noqise**	1 m	2 m	3 m	4 m	5 m	10 m	15 m
D	w/o	4.78	3.77	4.92	0.00	1.75	0.00	0.00
	w/	0.00	1.82	1.40	0.00	0.00	0.00	5.30
C	w/o	1.09	2.50	1.54	0.64	1.30	–	–
	w/	1.46	0.00	0.00	0.00	0.00	–	[deg]

(*) "D" means Dining, and "C" means Cathedral.
(**) A popular music is used as a background noise

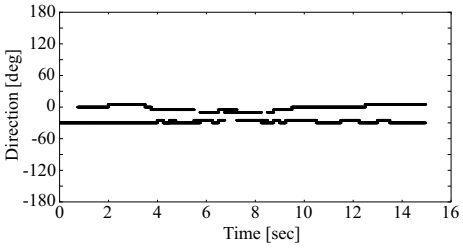
Figure 8 shows the MUSIC spectrogram in various distance in Dining. As shown in the figures, we find clear peaks when the talker stands at a distance of 1 m, but its peaks becomes unclear as the distance increases. The reason is the mismatch between the actual and pre-measured transfer functions from the talker to the Texai. This mismatch becomes severe as the distance increases, which makes the localization difficult.

For conditions of more than one talkers, we do not show the table of standard deviations because the result is similar to Table I. The standard deviations are up to 5° in almost all conditions. This deviations is enough smaller than the interval of talkers. Therefore, the performance is enough high for an remote operator to give information for distinguishing each talker's place around Texai.

Instead of showing the tables, Figures 9 and 10 shows the trajectories of localization with double and triple talkers, respectively, as examples. Figure 9(a) is the successful example, whose trajectories are stable. On the other hand, Figure 9(b) is the example with misestimation. From 6 sec to 8 sec, we find that the trajectory corrupts. Figure 10 (a) and (b) show the similar result to Figure 9.

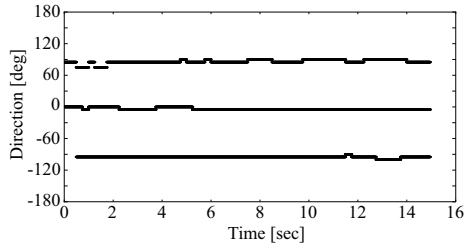


(a) 1 m away from Texai: Successful localization.

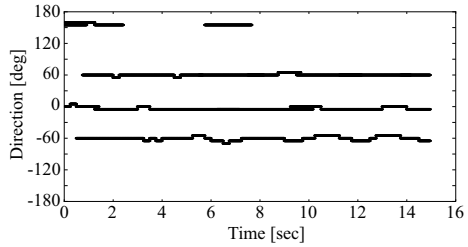


(b) 2 m away from Texai: Estimation fails from 6 to 8 seconds.

Fig. 9. Examples of localization trajectory in double talkers in Cathedral. Their interval is 30°. No background music.



(a) Without background music: Successful localization.



(b) With background music: False localization at 150° and fluctuated localization at -60°

Fig. 10. Examples of localization trajectory in triple talkers in Cathedral. Three talkers are 1 m away from Texai and the intervals of two adjacent talkers are 60°.

D. Example of Texai with selectable sound separation

This section demonstrates how our system works by showing trajectories of localized sounds and spectrograms of separated sounds. The scenario is as follows: the audition-enhanced Texai is at the center of Cathedral, and there are four talkers around the Texai. These talkers speak to the Texai at the same time without walking around, and our system localizes each talkers and separates particular talk. Figure 1 shows the situation of this example.

Figure 11 shows the localization result. Five lines are found in the Figure, each of which lines corresponds to

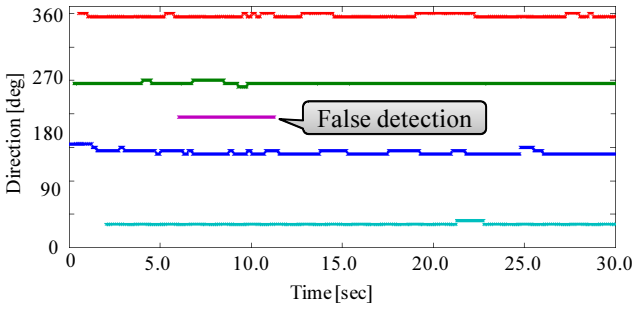


Fig. 11. Trajectories of sound location: The horizontal and vertical axes denote time and talkers's direction, respectively.

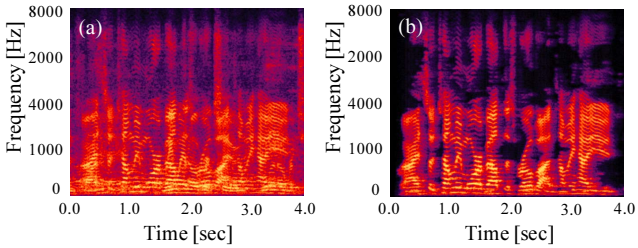


Fig. 12. Spectrograms of the mixed and separated sounds

localized sound. Four long lines at around 45, 180, 270 and 360 degrees successfully localize the talkers. Although each talker stayed at the same place during theier utterances, the localization results fluctuate because of two fluctuations: (1) the talkers' head position while uttering and (2) the criteria $P(\phi)$ shown in eq. (5). The purple line at 225 degrees is the misestimated localization. This misestimation happens because of the reflection of sounds caused by the walls, ceiling and floor of the room, or the spatial aliasing.

Figure 12 shows spectrograms of the mixture of and separated sounds. Figure 12(a) shows the mixture of four talkers' sounds. For an operator, it is extremely difficult for the Texai operator to tell what each talker is uttering because the speech signal is totally interfered with the others' speech. Figure 12(b) is a separated speech from 270°. This function enables the Texai operator to understand what the talker is talking.

V. Discussion

The MUSIC spectrum theoretically has a sharp peak when a sound exists, but the peak becomes smooth because of reverberation or existence of noise. Moreover, the performance degrades when the power of noise is more than that of sound sources to be detectded. Note that such problem did not alise in the experiments because we assumed that such noise do not exist. To solve such a problem, generalized eigenvalue decomposition based MUSIC [14], which uses a covariance matrix of noise for whitening. The MUSIC described in this paper is a special version that the covariance matrix is a unit matrix. Our group have developed the method in real-time and dynamic environment [15].

VI. Conclusion

This paper presented the audition-enhanced telepresence system Texai, modified with the selectable sound separation function using HARK. We developed a sound location visualization system with separated sound play for a remote Texai operator. We also installed an eight-microphone array on a salad bowl in Texai. Evaluation of our system shows that the resulting system is capable of localizing the surrounding sounds at a tolerance of 5 degrees, although the performance degrades when the talkers are close together. The implementation time was only in five days, which means that HARK speeds up the development time of auditory awareness functionality to robots.

We have two future works. (1) Detecting the sounds from MUSIC spectrum (Eq. 5) is currently based on comparing with a given threshold. We, therefore, need to tune the threshold when the number of talkers or the rooms change. More sophisticated sound location estimation from the shape of the spectrum is needed. (2) More precise evaluation is needed. Because of time and cost constraints, we need to concentrate on developing the system on Texai, and we evaluated our system in only a preliminary way. For example, the use of multiple loudspeakers fixes the talkers' volume and position, and a usability test comparing the current Texai with the audition-enhanced Texai is an important feature to be evaluated.

Acknowledgements

We thank Aki Oyama, Rob Wheeler and Curt Meyers from Willow Garage, Inc. for their helpful advice and cooperation, Masatoshi Yoshida for his assistance on data analysis, and Angelica Lim and Louis-Kenzo Cahier for their valuable comments on earlier drafts. This work was partially supported by a Grant-in-Aid for Scientific Research (S) (No. 19100003) from MEXT, Japan, and the Global COE Program at Kyoto University from JSPS, Japan. Part of this work was done while the authors were visiting Willow Garage, Inc.

References

[1] E. Guizzo. When my avatar went to work. *IEEE Spectrum*, pages 24–29, 48, 50, Sep. 2010.

[2] E. C. Cherry. Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.*, 25(5):975–979, 1953.

[3] I. Toshima and S. Aoki. Effect of head movement on sound localization in an acoustical telepresence robot: Telehead. In *Proc. of IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 872–877, 2006.

[4] M. Kashino and T. Hirahara. One, two, many – judging the number of concurrent talkers. *J. Acoust. Soc. Am.*, 99(4):Pt.2, 2596, 1996.

[5] D. Rosenthal and H.G. Okuno, editors. *Computational Auditory Scene Analysis*. Lawrence Erlbaum Associates, Mahwah, New Jersey, 1998.

[6] K. Nakadai, H.G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino. Design and implementation of robot audition system "HARK". *Advanced Robotics*, 24:739–761, 2009.

[7] Y. Kubota, M. Yoshida, K. Komatani, T. Ogata, and H.G. Okuno. Design and implementation of 3D auditory scene visualizer towards auditory awareness with face tracking. In *Proc. of IEEE Intl. Symp. on Multimedia (ISM2008)*, pages 468–476, 2008.

[8] B. Shneiderman. *Designing the User Interface (3rd Ed)*. Addison-Wesley, New York, 1998.

[9] Willow Garage, Inc. Texas robot. "http://www.willowgarage.com/blog/2009/10/26/texas-robot", Oct. 2009.

[10] S. Cousins, B. Gerkey, K. Conley, and W. Garage. Sharing software with ros. *IEEE Robotics & Automation Magazine*, 17(2):12–14, 2010.

[11] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino. Blind source separation with prameter-free adaptive step-size method for robot audition. *IEEE trans. Audio, Speech, and Language Processing*, 18(6):1476–1484, 2010.

[12] H. Sawada, R. Mukai, S. Araki, and S. Makino. Polar coordinate based nonlinear function for frequency-domain blind source separation. In *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1001–1004, 2002.

[13] D.H. Brandwood. A complex gradient operator and its application in adaptive array theory. *IEEE Proc.*, 130(1):251–276, 1983.

[14] R. Roy and T. Kailath. ESPRIT- estimation of signal parameters via rotational invariance techniques. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 37(7):984–995, 1989.

[15] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino. Intelligent sound source localization for dynamic environments. In *Proc. of IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 664–669, 2009.