Framework of Distributed Audition

T. Ikeda T. Ishida H. Ishiguro
Dept. of Adaptive Machine Systems
Graduate School of Engineering, Osaka University
2-1 Yamadaoka, Suita Osaka, 565-0871, JAPAN
Email ikeda@ed.ams.eng.osaka-u.ac.jp

Abstract

In this paper, we propose distributed audition for natural man-machine interface, for robots that act in the environment, and for continuous personal identification. The distributed audition system, consisting of a network of microphones and speakers, monitors the environment, maintains the environment models, and provide information to agents in the environment. The distributed audition system can calibrate locations and parameters in self-organizing manner by producing sounds and observing them. Concepts and fundamental problems of distributed audition are discussed. This paper also provides a prototype system and an experimental result of auto calibration.

1 Introduction

In this paper, we propose distributed audition consisting of microphones and speakers placed in pairs and connected with a computer network. We call a pair of a microphone and a speaker as audition agent. In a distributed audition system, many audition agents are installed in the environment in distributed manner and communicate using computer network and sound to realize intelligent environment. There are three motivations to construct distributed audition systems.

The first is to realize a man-machine interface based on human behavior that doesn't depend on locations of users. Recognition of many kinds of human behavior is key technology to create natural computer interfaces and to design robots that behave in our daily environment. A considerable number of studies have been performed on gesture and face recognition using cameras and sound recognition using microphones. However, the appropriate distance between sensors and targets are assumed in these researches. That is because the recognition system works well only when targets are observed in good resolution and signal-to-noise ratio. So users always have to be conscious of the location of sensors and keep adequate position so that the recognition works well.

Recently, sensor networks with many kinds of sensors in the environment attract increasing attention. By using the network of sensors in the environment, we can realize the interface that can recognize human behavior independent of human locations [1][2]. In camera networks, the system can observe human behavior everywhere in the environment from a few cameras and the system can acquire enough information to recognize the behavior. In the case of the interfaces using sound, we can realize more natural interface by installing many microphones and speakers in the environment. The system recognizes the voice and sound made by human everywhere in the environment and users don't have to be conscious of locations of sensors.

The second is to realize a natural robot interface. The required function for robots that behave in our daily environment is to communicate with humans using their own sensors and actuators, and controlling the attention to select viewing points according to various events is necessary to continue communication [3]. However, it requires a complex plan for robot to behave and observe simultaneously by selecting appropriate viewing points. Furthermore, signal-to-noise ratio of observed signal become lower since the motion makes noise[4] and makes the images from camera unstable.

To cope with this problem, a promising approach is to install a network of sensors in the environment and construct perceptual information infrastructure (PII) [5]. Now robots are not restricted to its body and make fully use of information from PII to communicate with people. For example, to listen to the sign from people in the environment using sound, robots can use not only the microphone on their body but also the microphones in the environment. Background noise that depends on each location can be reduced using developed techniques for fixed microphones (e.g. spectral subtraction [6]). When the system informs a person in the environment using sound, the speaker nearest to the person can be used instead of broadcasting. Furthermore, observation from different point of view is sometimes essential to accomplish the task.

The third is to realize more secure personal identification. Currently, personal identification is performed at the specific devices and it is obvious when the identification process is performed. Generally, it is more secure if personal identification process is hidden and anyone doesn't know when the identification process has performed. By using sensor network in the environment, we can realize the identification system based on behavior and sound of daily human activities, and the system that performs identification at any time.

In this paper, we propose distributed audition system that many audition agents are installed in the environment to realize intelligent environment. In chapter 2, the concept of distributed audition and fundamental problems of distributed audition are shown. In chapter 3, the problem of position estimation of sound sources, microphones, and speakers is described. In chapter 4, a prototype system of distributed audition is shown.

2 Concepts and Fundamental Problems of Distributed Audition

2.1 Concept of Distributed Audition

Location-free interface

Distributed audition provides a man-machine communication interface that is independent of the human location by installing many audition agents in distributed manner in the environment.

Perceptual information infrastructure

Distributed audition models local acoustic environment, maintains the dynamic environment models, and provides information for the agent in the environment. The robot in a distributed audition system is free from the constraint that sensors are put on his body and can get information for the task from the perceptual information infrastructure.

Anytime identification

Distributed audition constructs personal sound models of daily life by observing human in the environment for a long period. Then the personal identification is performed at any time and is not restricted to the specific position and time.

Depends on the situation, it is hard to realize these objectives only by sound. By combining other kinds of sensors (cameras, touch sensors, and so on), we can build more reliable and robust system.

2.2 Related Works

Many microphones and speakers are used in the research area of microphone array and speaker array [7][8]. Distributed audition and these arrays differ in many points.

First, in distributed audition, one microphone and speaker are placed in combined manner. It enables to the auto calibration by generating and recognizing sounds for the calibration.

Secondly, spatial extent of microphone and speakers are essential in both approach, however, the size of array is small compared to the distance from the array to the target. In distributed audition, microphones and speakers are installed everywhere in the environment in widely distributed manner.

Thirdly, microphones and speakers in the array are placed at well-designed location (usually at even intervals). In distributed audition, microphones and speakers are placed at the any locations in the environment and the relation between them is calibrated manually or in self-organizing manner.

2.3 Fundamental Problems in Distributed Audition

In distributed audition, the followings are fundamental problems to be tackled.

1. Estimation of sound sources location in the environment

Robust estimation of locations of sound sources in the environment based on the location of microphones and observed signals.

2. Calibration of distributed audition

When we construct a perceptual information infrastructure in the specific place, measuring the locations and parameters is required. By producing sound from speakers and observing by microphones, the relative locations of microphones and speakers, microphone gain, and speaker gain can be computed automatically.

3. Behavior recognition based on sounds of daily life

By making statistics of sounds of daily life in the environment, we can construct models to recognize human behavior. Observation by sensor distributed in the environment is useful since human behaviors highly depend on the location. By combining with camera network, better human behavior recognition system can be built.

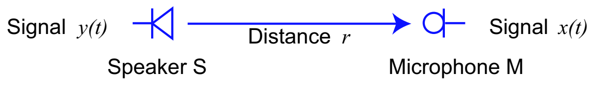
4. Personal identification in the environment by combining many kinds of sensors

By modeling the human daily behavior with cameras, microphones, and touch sensors, the personal model for identification is created. We can perform personal identification continuously by applying this model.

3 Measuring Locations of Microphones and Sound sources

In this section, location estimation of the sound source and speakers are considered. Here, the number of the sound source that makes sound at the same time is limited to one. There are a number of estimation methods according to which quantities are known or unknown and whether a microphone and a speaker are placed in pair or not.

Parameters and variables are defined as Fig. 1. We assume (1) and (2) on playing and recording the sound signal.



r	Distance between speaker S and microphone M
$y(t)$	Signal played at speaker S
$x(t)$	Signal observed at microphone M
x, y	Short time average of signal $x(t)$ and $y(t)$
P	Power of signal emitted from speaker S
H	Speaker gain
G	Microphone gain

Figure 1: Explanatory notes

$$P = Hy^2 \tag{1}$$

$$x^2 = G\frac{P}{r^2} = GH\frac{y^2}{r^2} \tag{2}$$

3.1 Energy Based Method

Here we discuss the estimation method based on the average energy of the signal observed by microphones. Estimation methods are summarized in Table 1. In the table, $\surd$ indicates the objective quantity and O indicates the required quantity to estimate the objective quantity in the same row. Due to the limitation of space, not all estimation algorithm are described here.

Algorithm: Measuring the Location of a Sound Source

Table 1: Energy based method

1. Location of microphones and speakers (sound sources) are different

Type	Location of microphones	Microphone gain	Location of speakers	Speaker gain
1-a	O	O	$\surd$	O
1-b	O	O*	$\surd$	
1-c	O		$\surd$	O*
2-a	$\surd$	O	O	O
2-b	$\surd$		O	O*
2-c	$\surd$	O*	O	
3-a	O	$\surd$	O	O
3-b	O	$\surd^*$	O	
4-a	O	O	O	$\surd$
4-b	O		O	$\surd^*$

2. A microphone and a speaker are placed in pair (audition agent)

Type	Location of audition agents	Microphone gain	Speaker gain
5-a	$\surd$	O	O
5-b	$\surd^*$	$\surd^*$	O*
5-c	$\surd^*$	O*	$\surd^*$

$\surd$ indicates the objective quantity and O indicates the required quantity to estimate the objective quantity in the same row. $\surd^*$ and O* indicates not absolute value but relative ratio of the quantity.

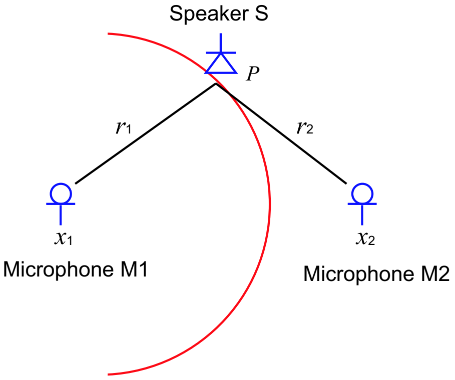


Figure 2: Measureing the location of a microphone

The first example shows estimation of the location of a sound source when the locations of microphones known (type 1-b in Table 1).

Let P be unknown power of the sound emitted from the sound source S and x_j be observed signal and G_j be gain of microphne M$_j$ respectively. Then,

$$x_1^2 = G_1 \frac{P}{r_1^2},$$

$$x_2^2 = G_2 \frac{P}{r_2^2},$$

$$\left(\frac{r_1}{r_2}\right)^2 = \frac{G_1}{G_2}\left(\frac{x_2}{x_1}\right)^2. \qquad (3)$$

From (3), we can compute the ratio of the distance r_1/r_2 from the knowledge of G_1/G_2. It means location of the sound source is limited on a circle. By observing the sound source from other pairs of microphones repeatedly, the location of the sound source is estimated.

Algorithm: Mutual Estimation of Locations of Audition Agents (power-based)

The second example shows mutual location estimation of microphones and speakers when a microphone and a speaker are placed in pair (type 5-b and 5-c in Table 1).

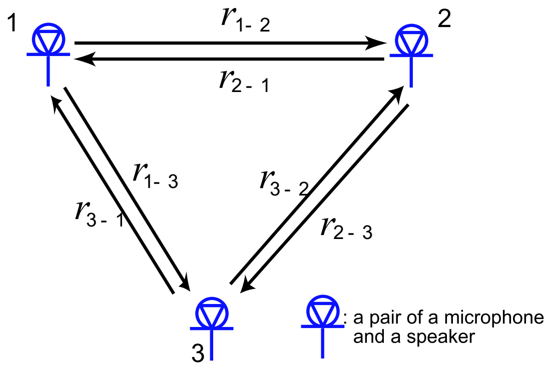


Figure 3: Mutual estimation of locations of audition agents (power-based)

Let y_i be the signal played at speaker i, H_i be gain of the speaker i. Let G_j be gain of microphone j and $x_{i\to j}$ be the observed signal at microphone j emitted from speaker i. Let $r_{i\to j}$ be distance between speaker i and microphone j. Then,

$$x_{1\to 2}^2 = G_2 \frac{H_1\, y_1^2}{r_{1\to 2}^2},$$

$$x_{2\to 1}^2 = G_1 \frac{H_2\, y_2^2}{r_{2\to 1}^2}.$$

Then,

$$\frac{x_{1\to 2}^2}{x_{2\to 1}^2} = \frac{G_2\, H_1\, y_1^2}{G_1\, H_2\, y_2^2},$$

As a result,

$$\frac{G_2}{G_1} = \frac{H_2}{H_1}\frac{x_{1\to 2}^2\, y_2^2}{x_{2\to 1}^2\, y_1^2}. \qquad (4)$$

From (4), the ratio of microphone gain G_2/G_1 can be computed from the ratio of speaker gain H_2/H_1, and vise versa. Namely,

$$\frac{G_2}{G_1} \rightleftharpoons \frac{H_2}{H_1}. \qquad (5)$$

After knowing the ratio of microphone gain and the ratio of speaker gain, the ratio between each pair of a microphone and a speaker can be computed by making use of the third pair of a microphone and a speaker.

$$\frac{r_{3\to 2}^2}{r_{3\to 1}^2} = \frac{G_2}{G_1}\frac{x_{3\to 1}^2}{x_{3\to 2}^2}.$$

$$\frac{r_{2\to 3}^2}{r_{1\to 3}^2} = \frac{H_2}{H_1}\frac{y_2^2\, x_{1\to 3}^2}{y_1^2\, x_{2\to 3}^2}$$

Table 2: Impulse based method

1. Location of microphones and speakers (sound sources) are different

Type	Location of microphones	Location of speakers
6-a	O	√
6-b	√	O

2. A microphone and a speaker are placed in pair (audition agent)

Type	Location of audition agents
7-a	√

3.2 Impulse Based Method

The distance between sound source and a microphone can be computed based on the time difference of arrival time when an impulse signal is used. In this case, estimation does not depend on the gain of microphones and speakers. Estimation methods are summarized in Table 2.

Algorithm: Mutual Estimation of Location of Audition Agents (impulse-based)

The last example shows location estimation of microphones and speakers without any knowledge but a microphone and a speaker is installed in combined manner (type 7-1).

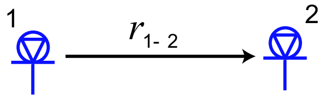


Figure 4: Mutual estimation of locations of audition agents (impulse-based)

The time difference between the making impluse and observation of impuluse can be computed. So, each distance between audition agents is estimated.

$$r_{1\to 2} = c * \Delta$$

where c is the speed of the sound, Δ is the time difference.

4 A Prototype of Distributed Audition System

A prototype of distributed audition system and estimation of location based on impulse method is shown. Overview of the system is shown in Fig. 5.

Number of audition agents are four and are described as A, B, C, and D. We make an impulse noise (Fig. 6) five times from the each audition agent and observe from all microphones.

Fig. 7 shows the average and the standard deviation of 10 times observation of the distance between audition agents. For example, the distance between A and B is measured 10 times that includes 5 times from A to B and 5 times from B to A. Fig. 8 shows the location estimation result based on the average distance in Fig. 7.

5 Conclusion

In this paper, we propose distributed audition for natural man-machine interface, for robots that behave in the environment, and for continuous personal identification. The concept of distributed audition is installing pairs of a microphone and a speaker in the environment in distributed manner. Unlike microphone array and speaker array, microphones and speakers are placed widely spread in the environment. Fundamental problem of the distributed audition includes location estimation of sound sources, location estimation of microphones and speakers in the system, behavior recognition and personal identification combined with

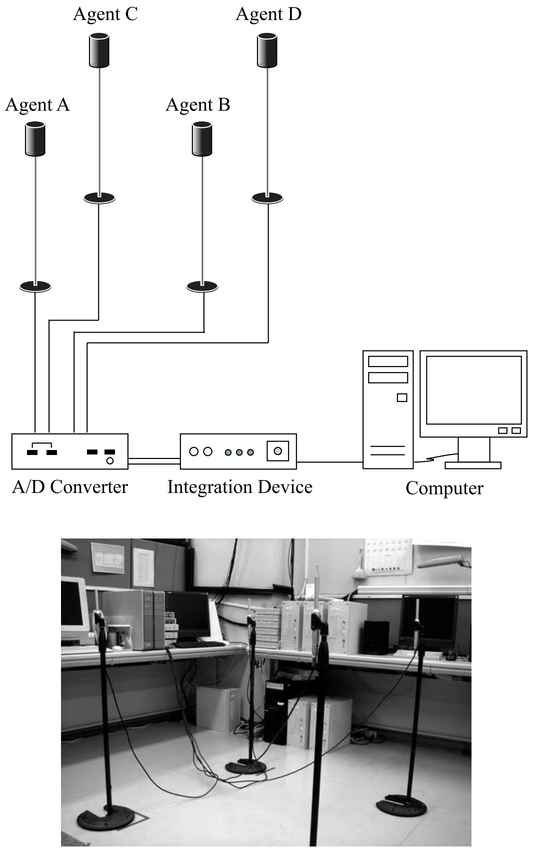


Figure 5: An overvew of the prototype system

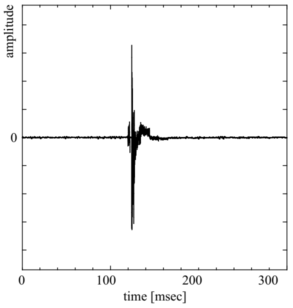


Figure 6: An example of an impluse noise

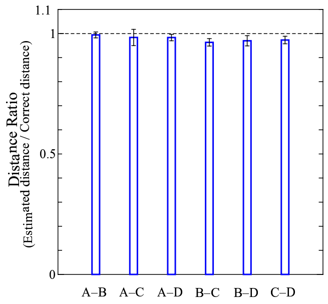
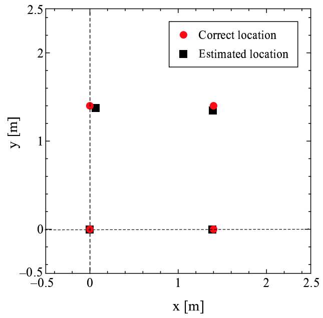


Figure 7: Estimated distance between microphones



	Correct location [m]		Estimated location [m]		
	x	y	x	y	z
A	0.00	0.00	0.00	0.00	0.00
B	1.40	0.00	1.39	0.00	0.00
C	0.00	1.40	0.07	1.38	0.00
D	1.40	1.40	1.39	1.35	0.16

Figure 8: Estimated locations of audition agents

other kinds of sensors. In this paper, summaries of mutual location estimation algorithm of microphones and speakers are shown, and a few algorithms are described. A prototype system of distributed audition and an experimental result of location estimation based on the impulse method are shown.

Our next step includes developing location estimation algorithms in real environmnt and behavior recognition systems combined with other kinds of sensors. Now we are developing the distributed audition system with eight microphone in the office environment with walls that shield and reflect the noise (Fig. 9). In this environment, the problem is that not all computed distances between audition agents are correct. We are going to include relaxation method to estimate correct positions of audition agents.

References

[1] H. Ishiguro and T. Nishimura, “VAMBAM: View and motion-based aspect models for distributed omnidirectional vision systems,” in *Proc. International Joint Conference Artificial Intelligence*, 2001, pp. 1375–1380.

[2] T. Matsuyama and N. Ukita, “Real-time multi-target tracking by a cooperative distributed vision system,” *Proc. IEEE*, vol. 90, no. 7, pp. 1136–1150, 2002.

[3] D. H. Ballard, “Reference frames for animate vision,” in *Proc. International Joint Conference on Artificial Intelligence (IJCAI’89)*, 1989, pp. 1635–1641.

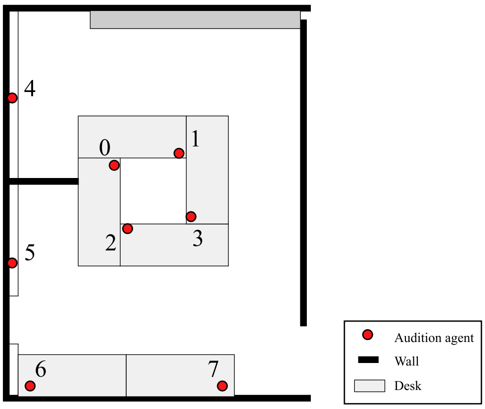


Figure 9: An overvew of the prototype in the office environment

[4] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, “Active audition for humanoid,” in *Proc. the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, 2000, pp. 832–839.

[5] H. Ishiguro, “Distributed vision system: A perceptual information infrastr ucture for robot navigation,” in *Proc. International Joint Conference Artificial Intelligence*, 1997, pp. 36–41.

[6] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 113–120, 4 1979.

[7] T. Yamada, S. Nakamura, and K. Shikano, “Robust speech recognition with speaker localization by a microphone array,” in *Proc. International Conference on Spoken Language Processing (ICSLP ’96)*, vol. 3, 1996, pp. 1317–1320.

[8] H. Mizoguchi et al., “Virtual earphone: Integration of beam forming by speaker array and realtime visual face tracking,” *Key Engineering Material*, vol. 243-244, pp. 117–122, 2003.