Proceedings of the 2001 IEEE/RSJ
International Conference on Intelligent Robots and Systems
Maui, Hawaii, USA, Oct. 29 - Nov. 03, 2001

Human-Robot Interaction through Real-Time Auditory and Visual Multiple-Talker Tracking

Hiroshi G. Okuno[†§], Kazuhiro Nakadai[†], Ken-ichi Hidai[†],
Hiroshi Mizoguchi[‡], and Hiroaki Kitano [†¶]

† Kitano Symbiotic Systems Project, ERATO, Japan Science and Technology Corp.
Mansion 31 Suite 6A, 6-31-15 Jingumae, Shibuya-ku, Tokyo 150-0001, Japan
§ Department of Intelligence Science and Technology, Kyoto University, Kyoto 606-8501, Japan
‡ Department of Information and Computer Science, Saitama University, Saitama 338-8570, Japan
¶ Sony Computer Science Laboratories, Inc., Tokyo 141-0022, Japan
{okuno, nakadai, hidai, kitano}@symbio.jst.go.jp, hm@me.ics.saitama-u.ac.jp

Abstract

Sound is essential to enhance visual experience and human robot interaction, but usually most research and development efforts are made mainly towards sound generation, speech synthesis and speech recognition. The reason why only a little attention has been paid on auditory scene analysis is that real-time perception of a mixture of sounds is difficult. Recently, Nakadai et al have developed real-time auditory and visual multiple-talker tracking technology. In this paper, this technology is applied to human-robot interaction including a receptionist robot and a companion robot at a party. The system includes face identification, speech recognition, focus-of-attention control, and sensorimotor task in tracking multiple talkers. The system is implemented on a upper-torso humanoid and the talker tracking is attained by distributed processing on three nodes connected by 100Base-TX network. The delay of tracking is 200 msec. Focus-of-attention is controlled by associating auditory and visual streams with using the sound source direction and talker position as a clue. Once an association is established, the humanoid keeps its face to the direction of the associated talker.

1 Introduction

In the 21st century, autonomous robots are more common in social and home environments, such as a pet robot at living room, a service robot at office, or a robot serving people at a party. The robot shall identify people in the room, pay attention to their voice and look at them to identify visually, and associate voice and visual images, so that highly robust event identification can be accomplished. These are minimum requirements for social interaction [4].

Sound has been recently recognized as essential in order to enhance visual experience and human computer interaction, and thus not a few contributions have been done by academia and industries at IROS and its related conferences [2, 3, 15]. This is because people can dance with sounds but not with images [6]. Handel also pointed out the importance of selective attention by writing *"We hear and see things and events that are important to us as individuals, not sound waves or light rays"* [6].

Sound, however, has not been utilized so much as input media except speech recognition. There are at least two reasons for this tendency:

1. **Handling of a mixture of sounds** — We hear a mixture of sounds, not a sound of single sound source. Automatic speech recognition (ASR) assumes that the input is a voiced speech and this assumption holds as long as a microphone is set close to the mouth of a speaker. Of course, speech recognition community develops *robust* ASR to make this assumption hold on wider fields [7].

2. **Real-time processing** — Some studies with computational auditory scene analysis (CASA) to understand a mixture of sounds has been done [18]. However, one of its critical problems in applying CASA techniques to a real-world system is a lack of real-time processing.

Usually, people hear a mixture of sounds, and people with normal hearing can separate sounds from the mixture and focus on a particular voice or sound in a noisy environment. This capability is known as the *cocktail party*

0-7803-6612-3/01/$10.00 © 2001 IEEE

effect [5]. Real-time processing is essential to incorporate cocktail party effect into a robot.

Nakadai *et al* developed *real-time* auditory and visual multiple-tracking system [16]. The key idea of their work is to integrate auditory and visual information to track several things simultaneously. In this paper, we apply the real-time auditory and visual multiple-tracking system to a receptionist robot and a companion robot of a party in order to demonstrate the feasibility of a cocktail party robot. The system is composed of face identification, speech separation, automatic speech recognition, speech synthesis, dialog control as well as the auditory and visual tracking.

The rest of the paper is organized as follows: Section 2 describes the design of the whole system, Section 3 describes the details of each subsystem, in particular, real-time auditory and visual multiple-tracking system. Section 4 demonstrates the system behavior. Section 5 discusses the observations of the experiments and future work, and Section 6 concludes the paper.

2 Task of Speaker Tracking

Real-time object tracking is applied to some applications that will run under auditorily and visually noisy environments. For example, at a party, many people are talking and moving. In this situation, strong reverbaration (echoes) occurs and speeches are interfered by other sounds or talks. Not only reverbaration, but also lighting of illuminating conditions change dynamically, and people are often occluded by other people and reappear.

2.1 Robots at a party

To design the system for such a noisy environment and prove its feasibility, we take "a robot at a party" as an example. Its task is the following two cases:

1) Receptionist robot

At the entrance of a party room, a robot interacts with a participant as a receptionist. It talks to the participant according to whether it knows the participant.

If the robot knows a participant, the task is very simple; it will confirm the name of the participant by asking "*Hello. Are you XXX-san?*". If it does not know a participant, it asks the name and then registers the participant's face with his/her name to the face database.

During this conversation, the robot should look at the participant and should not turn to any direction during the conversation.

2) Companion robot

In the party room, a robot plays a role of a passive companion. It does not speaks to a participant, but sees and listens to people. It identifies people's face and the position and turns its body to face the speaker.

The issue is to design and develop the tracking system which localizes the speakers and participants in real-time by integrating face identification and localization, sound source localization, and its motor-control.

In this situation, we don't make the robot to interact with the participants, because we believe that a silent companion is more suitable for the party and such attitude is more socially acceptable. Please note that the robot knows all the participants, because they registered at the receptionist desk.

2.2 SIG the humanoid

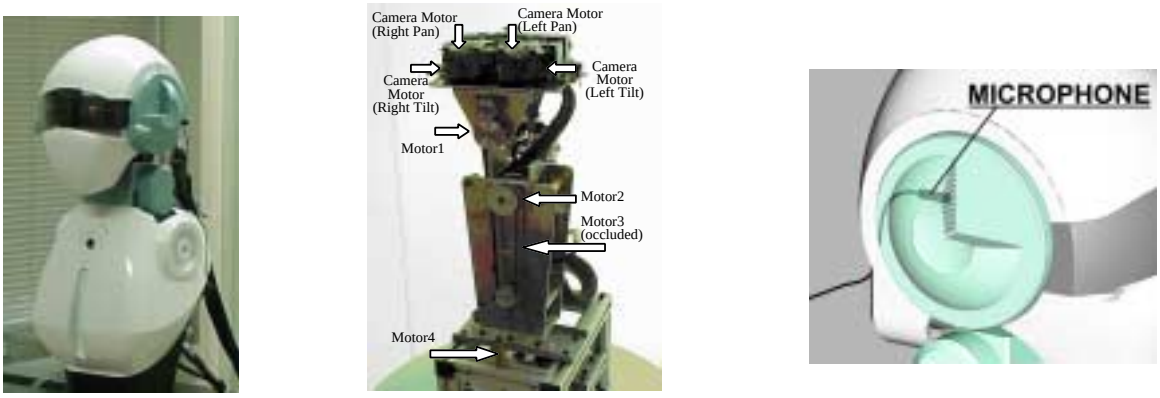
As a testbed of integration of perceptual information to control motor of high degree of freedom (DOF), we designed a humanoid robot (hereafter, referred as *SIG*) with the following components [12]:

- 4 DOFs of body driven by 4 DC motors — Its mechanical structure is shown in Figure 1b. Each DC motor has a potentiometer to measure the direction.
- A pair of CCD cameras of Sony EVI-G20 for visual stereo input — Each camera has 3 DOFs, that is, pan, tilt and zoom. Focus is automatically adjusted. The offset of camera position can be obtained from each camera (Figure 1b).
- Two pairs of omni-directional microphones (Sony ECM-77S) (Figure 1c). One pair of microphones are installed at the ear position of the head to collect sounds from the external world. Each microphone is shielded by the cover to prevent from capturing internal noises. The other pair of microphones is to collect sounds within a cover.
- A cover of the body (Figure 1a) reduces sounds to be emitted to external environments, which is expected to reduce the complexity of sound processing.
  This cover, made of FRP, is designed by our professional designer for making human robot interaction smoother as well [15].

3 System Description

Fig. 2 depicts the logical structure of the system based on client/server model. Each server or client executes the following modules:

1. Audition — extracts auditory events by pitch extraction, sound source separation and localization, and



a) Cover made of FRP b) Mechanical structure c) A microphone for the external world

Figure 1: *SIG* the Humanoid: Its cover, mechanical structure, and a microphone

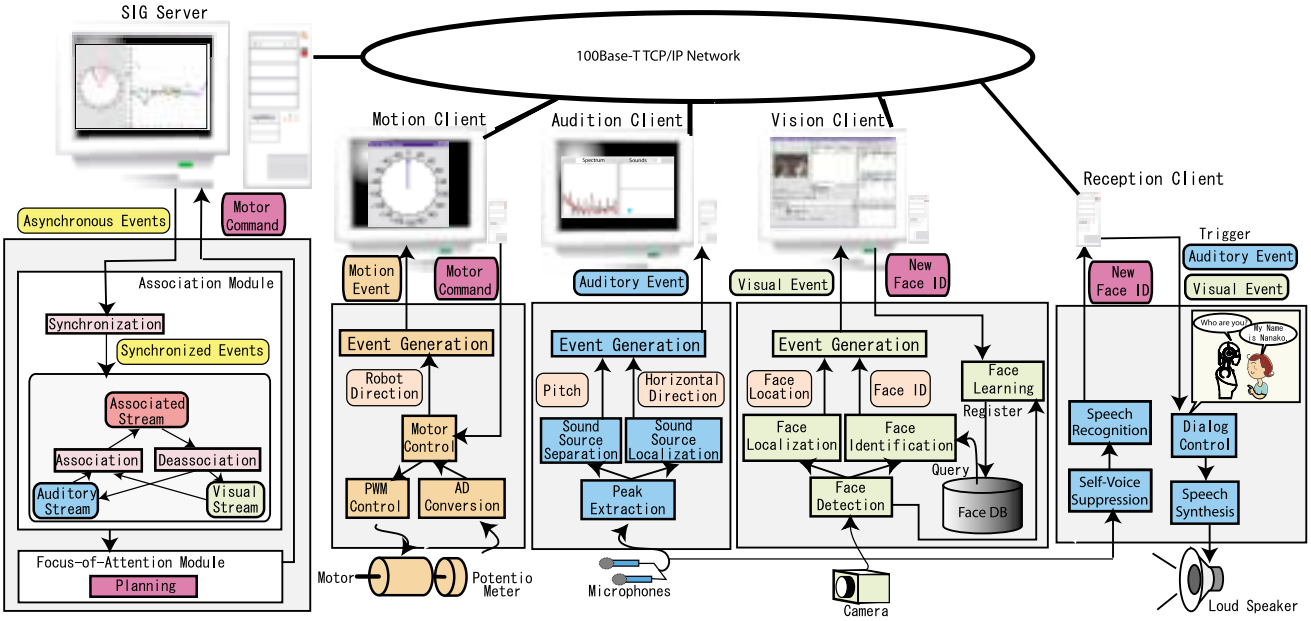


Figure 2: Logical organization of the system composed of SIG server, Motion, Audition, Vision, and Reception clients from left to right.

sends those events to Association,

2. Vision — extracts visual events by face extraction, identification and localization, and then sends visual events to Association,

3. Motor Control — generates PWM (Pulse Width Modulation) signals to DC motors and sends motor events to Association,

4. Association — integrates various events to create streams,

5. Focus-of-Attention — makes a plan of motor control,

6. Dialog Control — communicates with people by speech synthesis and speech recognition,

7. Face Database — maintains the face database, and

8. Viewer — instruments various streams and data.

Instrumentation is implemented distributedly on each node. SIG server displays the radar chart of objects and the stream chart. Motion client displays the radar chart of the body direction. Audition client displays the spectrogram of input

sound and pitch (frequency) vs sound source direction chart. Vision client displays the image of the camera and the status of face identification and tracking.

Among these subsystems, Focus-of-Attention, and Dialog Control are newly developed for the system reported in this paper. We use the automatic speech recognition system, "Julian" developed by Kyoto University [11].

Since the system should run in real-time, the above clients are physically distributed to three Linux nodes connected by TCP/IP over 100Base-TX network and run asynchronously. Vision is placed on a node of Pentium-III 733 MHz, Audition is on a node of Pentium-III 600 MHz, and the rest modules are on a node of Pentium-III 450 MHz.

3.1 Active Audition Module

To understand sound in general, not restricted to a specific sound, a mixture of sound should be analyzed. There are lots of techniques for CASA developed so far, but only the real-time active audition proposed by Nakadai *et al* runs in real-time [16]. Therefore, we use this system as the base of the receptionist and companion robots.

To localize sound sources with two microphones, first a set of peaks are extracted for left and right channels, respectively. Then, the same or similar peaks of left and right channels are identified as a pair and each pair is used to calculate interaural phase difference (IPD) and interaural intensity difference (IID). IPD is calculated from frequencies of less than 1500 Hz, while IID is from frequency of more than 1500 Hz.

Since auditory and visual tracking involves motor movements, which cause motor and mechanical noises, audition should suppress or at least reduce such noises. In human robot interaction, when a robot is talking, it should suppress its own speeches. Nakadai *et al* presented the *active audition* for humanoids to improve sound source tracking by integrating audition, vision, and motor controls [14]. We also use their heuristics to reduce internal burst noises caused by motor movements.

From IPD and IID, the epipolar geometry is used to obtain the direction of sound source [14]. The key ideas of their real-time active audition system are twofold; one is to exploit the property of the harmonic structure (fundamental frequency, $F0$, and its overtones) to find a more accurate pair of peaks in left and right channels. The other is to search the sound source direction by combining the belief factors of IPD and IID based on Dempster-Shafer theory.

Finally, audition module sends an auditory event consisting of pitch ($F0$) and a list of 20-best direction (θ) with reliability for each harmonics.

3.2 Vision: Face identification Module

Since the visual processing detects several faces, extracts, identifies and tracks each face simultaneously, the size, direction and brightness of each face changes frequently. The key idea of this task is the combination of skin-color extraction, correlation based matching, and multiple scale images generation [8].

The face identification module (see Fig. 2) projects each extracted face into the discrimination space, and calculates its distance d to each registered face. Since this distance depends on the degree (L, the number of registered faces) of discrimination space, it is converted to a parameter-independent probability P_v as follows.

$$P_v = \int_{\frac{d^2}{2}}^{\infty} e^{-t}\, t^{\frac{L}{2}-1} dt \qquad (1)$$

The discrimination matrix is created in advance or on demand by using a set of variation of the face with an ID (name). This analysis is done by using Online Linear Discriminant Analysis [9].

The face localization module converts a face position in 2-D image plane into 3-D world space. Suppose that a face is $w \times w$ pixels located in (x, y) in the image plane, whose width and height are X and Y, respectively (see screen shots shown in Fig. 4). Then the face position in the world space is obtained as a set of azimuth θ, elevation ϕ, and distance r as follows:

$$r = \frac{C_1}{w},\ \theta = \sin^{-1}\left(\frac{x - \frac{X}{2}}{C_2\, r}\right),\ \phi = \sin^{-1}\left(\frac{\frac{Y}{2} - y}{C_2\, r}\right)$$

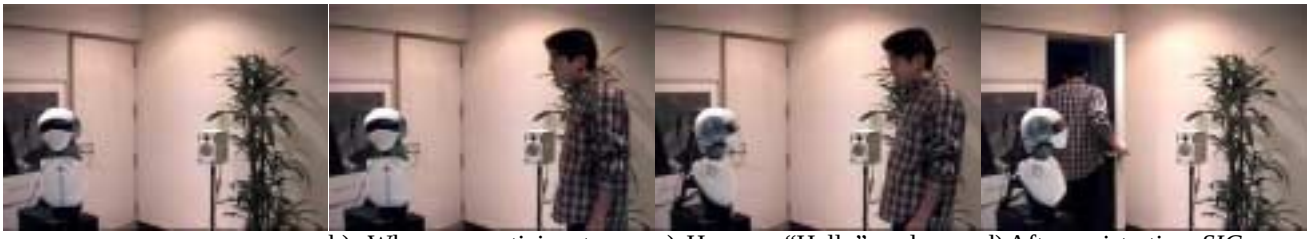
where C_1 and C_2 are constants defined by the size of the image plane and the image angle of the camera.

Finally, vision module sends a visual event consisting of a list of 5-best Face ID (Name) with its reliability and position (distance r, azimuth θ and elevation ϕ) for each face.

3.3 Stream Formation and Association

Association synchronizes the results (events) given by other modules. It forms an auditory, visual or associated stream by their proximity. Events are stored in the short-term memory only for 2 seconds. Synchronization process runs with the delay of 200 msec, which is the largest delay of the system, that is, vision module.

An auditory event is connected to the nearest auditory stream within $\pm 10^\circ$ and with common or harmonic pitch. A visual event is connected to the nearest visual stream within 40 cm and with common face ID. In either case, if there are plural candidates, the most reliable one is



a) Initial state

b) When a participant comes, *SIG* turns toward him.

c) He says "Hello" and introduces himself to *SIG*.

d) After registration, *SIG* registers his faces and name.

Figure 3: Temporal sequence of snapshots of *SIG*'s interaction as a receptionist robot

selected. If any appropriate stream is found, such an event becomes a new stream. In case that no event is connected to an existing stream, such a stream remains alive for up to 500 msec. After 500 msec of keep-alive state, the stream terminates.

An auditory and a visual streams are associated if their direction difference is within $\pm 10^\circ$ and this situation continues for more than 50% of the 1 sec period.

If either auditory or visual event has not been found for more than 3 sec, such an associated stream is deassociated and only existing auditory or visual stream remains. If the auditory and visual direction difference has been more than 30° for 3 sec, such an associated stream is deassociated to two separate streams.

3.4 Focus-of-Attention and Dialog Control

Focus-of-Attention Control is based on continuity and triggering. By continuity, the system tries to keep the same status, while by triggering, the system tries to track the most interesting object. Since the details of each algorithm depend on the applications, the focus-of-attention algorithm for a receptionist and companion robot is described in the next section.

Dialog control is a mixed architecture of bottom-up and top-down control. By bottom-up, the most plausible stream means the one that has the highest belief factors. By top-down, the plausibility is defined by the applications. For a receptionist robot, the continuity of the current focus-of-attention has the highest priority. For a companion robot, on the contrary, the stream that are associated the most recently is focused.

4 Experiments and Performance Evaluation

The width, length and height of the room of experiment is about 3 m, 3 m, and 2 m, respectively. The room has 6 down-lights embedded o n the ceiling.

For evaluation of the behavior of *SIG*, one scenario for the receptionist robot and one for the companion robot are executed. The first scenario examines whether an auditory stream triggers Focus-of-Attention to make a plan for *SIG* to turn to a speaker, and whether *SIG* can ignore the sound it generates by itself.

The second scenario examines how many people *SIG* can discriminate by integrating auditory and visual streams.

4.1 *SIG* as a receptionist robot

The rough scenario is specified as follows: (1) A participant comes to the receptionist robot, whose face has been registered in the face database. (2) He says Hello to *SIG*. (3) *SIG* replies "Hello. Are you XXX-san?" (4) He says yes. (5) *SIG* says "XXX-san, Welcome to the party. Please enter the room.".

By "rough", we mean that the scenario may not be executed literally, since there are a lot of possibilities; for example, the speaker's voice is so thin that speech recognition fails and *SIG* repeats the question.

Fig. 3 illustrates four snapshots of this scenario. Fig. 3 a) shows the initial state. The speaker on the stand is the mouth of *SIG*'s. Fig. 3 b) shows when a participant comes to the receptionist, but *SIG* has not noticed him yet, because he is out of *SIG*'s sight. When he speaks to *SIG*, Audition generates an auditory event with sound source direction, and sends it to Association, which creates an auditory stream. This stream triggers Focus-of-Attention to make a plan that *SIG* should turn to him. Fig. 3 c) shows the result of the turning. In addition, Audition gives the input to Speech Recognition, which gives the result of speech recognition to Dialog Control. It generates a synthesized speech. Although Audition notices that it hears the sound, *SIG* will not change the attention, because association of his face and speech keeps *SIG*'s attention on him. Finally, he enters the room while *SIG* tracks his walking.

This scenario shows that *SIG* takes two interesting behaviors. One is voice-triggered tracking shown in Fig. 3 c). The other is that *SIG* does not pay attention to its own

a) The leftmost man is being tracked.

b) He starts talking.

c) Associated stream is created.

d) The next right man is being tracked.

e) Face and voice are associated.

f) He stops talking.

g) The second right man starts talking.

h) The leftmost man starts talking.

Figure 4: Temporal sequence of snapshots for a companion robot: scene, radar and sequence chart, spectrogram and pitch-vs-direction chart, and the image of the camera.

speech. This is attained naturally by the current association algorithm, because this algorithm is designed by taking into account the fact that conversation is proceeded by alternate initiatives.

The variant of this scenario is also used to check whether the system works well. (1') A participant comes to the receptionist robot, whose face has not been registered in the face database. (2) He says Hello to *SIG*. (3) *SIG* replies "Hello. Could you give me your name?" (4) He says his name. (5) *SIG* says "XXX-san, Welcome to the party. Please enter the room.". After giving his name to the system, Face Database module is invoked.

4.2 *SIG* as a companion robot

There is no explicit scenario. Four speakers actually talks spontaneously in attendance of *SIG*. Then *SIG* tracks some speaker and then changes focus-of-attention to others.

The observed behavior is evaluating by consulting the internal states of *SIG*, that is, auditory and visual localization shown in the radar chart, auditory, visual, and associated streams shown in the stream chart, and peak extraction.

The top-left image in each snapshot shows the scene of this experiment recorded by a video camera.

The top-right image consists of the radar chart (left) and the stream chart (right) updated in real-time. The former shows the environment recognized by *SIG* at the moment of the snapshot. A pink sector indicates a visual field of *SIG*. Because of using the absolute coordinate, the pink sector rotates as *SIG* turns. A green point with a label is the direction and the face ID of a visual stream. A blue sector is the direction of an auditory stream. Green, blue and red lines indicate the direction of visual, auditory and associated stream, respectively. Blue and green *thin* lines indicate auditory and visual streams, respectively. Blue, green and red *thick* lines indicate associated streams with only auditory, only visual, and both information, respectively.

The bottom-left image shows the auditory viewer consisting of the power spectrum and auditory event viewer. The latter shows an auditory event as a filled circle with its pitch in X axis and its direction in Y axis.

The bottom-right image shows the visual viewer captured by the *SIG*'s left eye. A detected face is displayed with a red rectangle. The current system does not use a stereo vision.

The temporal sequence of *SIG*'s recognition and actions are summerized below:

Fig. 4a): *SIG* detects the leftmost man's face as a visual event of a red rectangle in the visual viewer. The visual stream shown as a green thin line in the stream chart is created by this event.

Fig. 4b): An auditory event is detected as a set of harmonic peaks in the spectrum, which is shown as a small circle int he auditory event viewer. This auditory event is also given to an auditory stream shown as a blue sector in the radar chart. This indicates that someone starts speaking.

Fig. 4c): Red lines are generated in the radar and stream charts. This indicates that an association is made because a visual stream and an auditory one have common direction for more than a constant time. By this association, *SIG* fixes focus-of-attention on him.

Fig. 4d): The associated stream is deassociated, because he stops speaking for a moment. By the deassociation, focus-of-attention of *SIG* is released. Then the rightmost man has spoken something, and *SIG* tried to turn to him. But he stoped speaking before *SIG* finishes turning. Then *SIG* stops turning, and is intersted in the second left man because his face is detected accidently in the visual viewer.

Fig. 4e): The second left man starts speaking and an associated stream on him is created. Thus, *SIG* fixes focus-of-attention on him.

Fig. 4f): He stops speaking, but *SIG* does not move, because it still detects his face and thus the association stream will remain for a few seconds.

Fig. 4g): After the associated stream is deassociated, the third left man starts speaking. Then *SIG* turns to him. In this case, *SIG* fails to detect his face, and thus no associated stream is created.

Fig. 4h): The leftmost man starts speaking again. Since there is no associated stream, *SIG* turns to him by an auditory trigger.

5 Discussions and Related Work

5.1 Observations

As a receptionist robot, once an association is established, *SIG* keeps its face fixed to the direction of the speaker of the associated stream. Therefore, even when *SIG* utters via a loud speaker on the left, *SIG* does not pay an attention to the sound source, that is, its own speech. This phenomena of focus-of-attention results in an automatic suppression of self-generated sounds. Of course, this kind of suppression is observed by another benchmark which contains the situation that *SIG* and the human speaker utter at the same time.

As a companion robot, *SIG* pays attention to a speaker appropriately. *SIG* also tracks the same person well when two moving talkers cross and their faces are out of sight of *SIG*. These results prove that the proposed system succeeds in real-time sensorimotor tasks of tracking with face identification. The current system has attained a passive companion. To design and develop an active companion may be important future work.

5.2 Related Work

Some robots are equipped with improved robot-human interface. *AMELLA* [19] can recognize pose and motion gestures, and some robots have microphones as ears for sound source localization or sound source separation. However, they have attained little in auditory tracking. Instead a microphone is attached close to the mouse of a speaker. For example, *Kismet* of MIT AI Lab can recognize speeches by speech-recognition system and express various kinds of sensation. *Kismet* has a pair of omni-directional microphones outside the simplified pinnae [2]. Since it is designed for one-to-one communication and its research focuses on social interaction based on visual attention, the auditory tracking has not been implemented so far. The adopted a simple and easy approach that a microphone for speech recognition is attached to the speaker.

Hadaly of Waseda University [13] can localize the speaker as well as recognize speeches by speech-recognition system. *Hadaly* uses a microphone array for sound source localization, but the microphone array is mounted in the body and its absolute position is fixed during head movements. Sound source separation is not exploited and a microphone for speech recognition is attached to the speaker

Jijo-2 [1] can recognize a phrase command by speech-recognition system. *Jijo-2* uses its microphone for speech recognition, but when it first stops, listens to a speaker, and recognize what he/she says. That is, *Jijo-2* lacks the capability of active audition.

Huang *et al* developed a robot that had three microphones [10]. Three microphones were installed vertically on the top of the robot, composing a regular triangle. Comparing the input power of microphones, two microphones that have more power than the other are selected and the sound source direction is calculated. By selecting two microphones from three, they solved the problem that two microphones cannot determine the place of sound source in front or backward. By identifying the direction of sound source from a mixture of an original sound and its echoes, the robot turns the body towards the sound source. Their demonstration is only turning the face triggered by a hand clapping not by continuous sounds. It could not track a moving sound source (talker).

6 Conclusion and Future Work

In this paper, we demonstrate that auditory and visual mutliple-object tracking subsystem can augment the functionality of human robot interaction. Although a simple scheme of behavior is implemented, human robot interaction is drastically improved by real-time multiple-person tracking. We can pleasantly spend an hour with *SIG* as a companion robot even if its attitude is quite passive.

Since the application of auditory and visual multiple-object tracking is not restricted to robots or humanoids, auditory capability can be transferred to software agents or systems. As discussed in the introduction section, auditory information should not be ignored in computer graphics

or human computer interaction. By integrating audition and vision, more cross-modal perception can be attained. Future work includes applications such as "listening to several things simultaneously" [17], "cocktail party computer", integration of auditory and visual tracking and pose and gesture recognition, and other novel areas. Since the event-level communication is less expensive than the low-level data representation, say signals itself, auditory and visual multiple-object tracking can be applied to tele-existence or virtual reality.

Acknowledgments

We thank our colleagues of Symbiotic Intelligence Group, Kitano Symbiotic Systems Project, Tatsuya Matsui, and former colloeague, Dr. Tino Lourens, for their discussions. We also thank Prof. Tatsuya Kawahara of Kyoto University for allowing us to use "Julian" automatic speech recognition system.

References

[1] ASOH, H., HAYAMIZU, S., HARA, I., MOTOMURA, Y., AKAHO, S., AND MATSUI, T. Socially embedded learning of the office-conversant mobile robot *jijo-2*. In *Proceedings of 15th International Joint Conference on Artificial Intelligence (IJCAI-97)* (1997), vol. 1, AAAI, pp. 880–885.

[2] BREAZEAL, C., AND SCASSELLATI, B. A context-dependent attention system for a social robot. In *Proceedints of the Sixteenth International Joint Conference on Atificial Intelligence (IJCAI-99)* (1999), pp. 1146–1151.

[3] BROOKS, R., BREAZEAL, C., MARJANOVIE, M., SCASSELLATI, B., AND WILLIAMSON, M. The cog project: Building a humanoid robot. In *Computation for metaphors, analogy, and agents* (1999), C. Nehaniv, Ed., Spriver-Verlag, pp. 52–87.

[4] BROOKS, R. A., BREAZEAL, C., IRIE, R., KEMP, C. C., MARJANOVIC, M., SCASSELLATI, B., AND WILLIAMSON, M. M. Alternative essences of intelligence. In *Proceedings of 15th National Conference on Artificial Intelligence (AAAI-98)* (1998), AAAI, pp. 961–968.

[5] CHERRY, E. C. Some experiments on the recognition of speech, with one and with two ears. *Journal of Acoustic Society of America 25* (1953), 975–979.

[6] HANDEL, S. *Listening*. The MIT Press, MA., 1989.

[7] HANSEN, J., MAMMONE, R., AND YOUNG, S. Editorial for the special issue on robust speech processing. *IEEE Transactions on Speech and Audio Processing 2*, 4 (1994), 549–550.

[8] HIDAI, K., MIZOGUCHI, H., HIRAOKA, K., TANAKA, M., SHIGEHARA, T., AND MISHIMA, T. Robust face detection against brightness fluctuation and size variation. In *Proceedings of IEEE/RAS International Conference on Intelligent Robots and Systems (IROS-2000)* (2000), IEEE, pp. 1397–1384.

[9] HIRAOKA, K., HAMAHIRA, M., HIDAI, K., MIZOGUCHI, H., MISHIMA, T., AND YOSHIZAWA, S. Fast algorithm for online linear discriminant analysis. In *Proceedings of ITC-2000* (2000), IEEK/IEICE, pp. 274–277.

[10] HUANG, J., OHNISHI, N., AND SUGIE, N. Building ears for robots: sound localization and separation. *Artificial Life and Robotics 1*, 4 (1997), 157–163.

[11] KAWAHARA, T., LEE, A., KOBAYASHI, T., TAKEDA, K., MINEMATSU, N., ITOU, K., ITO, A., YAMAMOTO, M., YAMADA, A., UTSURO, T., AND SHIKANO, K. Japanese dictation toolkit – 1997 version –. *Journal of Acoustic Society Japan (E) 20*, 3 (1999), 233–239.

[12] KITANO, H., OKUNO, H. G., NAKADAI, K., FERMIN, I., SABISH, T., NAKAGAWA, Y., AND MATSUI, T. Designing a humanoid head for robocup challenge. In *Proceedings of the Fourth International Conference on Autonomous Agents (Agents 2000)* (2000), ACM.

[13] MATSUSAKA, Y., TOJO, T., KUOTA, S., FURUKAWA, K., TAMIYA, D., HAYATA, K., NAKANO, Y., AND KOBAYASHI, T. Multi-person conversation via multi-modal interface — a robot who communicates with multi-user. In *Proceedings of 6th European Conference on Speech Communication Technology (EUROSPEECH-99)* (1999), ESCA, pp. 1723–1726.

[14] NAKADAI, K., LOURENS, T., OKUNO, H. G., AND KITANO, H. Active audition for humanoid. In *Proceedings of 17th National Conference on Artificial Intelligence (AAAI-2000)* (2000), AAAI, pp. 832–839.

[15] NAKADAI, K., MATSUI, T., OKUNO, H. G., AND KITANO, H. Active audition system and humanoid exterior design. In *Proceedings of IEEE/RAS International Conference on Intelligent Robots and Systems (IROS-2000)* (2000), IEEE, pp. 1453–1461.

[16] NAKADAI, K. HIDAI, K., MIZOGUCHI, H., OKUNO, H. G., AND KITANO, H. Real-time auditory and visual multiple-object tracking for robots. In *Proceedints of the Seventeenth International Joint Conference on Atificial Intelligence (IJCAI-01)* (2001), *to appear*.

[17] OKUNO, H. G., NAKATANI, T., AND KAWABATA, T. Listening to two simultaneous speeches. *Speech Communication 27*, 3-4 (1999), 281–298.

[18] ROSENTHAL, D., AND OKUNO, H. G., Eds. *Computational Auditory Scene Analysis*. Lawrence Erlbaum Associates, Mahwah, New Jersey, 1998.

[19] WALDHERR, S., THRUN, S., ROMERO, R., AND MARGARITIS, D. Template-based recoginition of pose and motion gestures on a mobile robot. In *Proceedings of 15th National Conference on Artificial Intelligence (AAAI-98)* (1998), AAAI, pp. 977–982.