From: AAAI Technical Report WS-02-05. Compilation copyright © 2002, AAAI (www.aaai.org). All rights reserved.

Non-Verbal Eliza-like Human Behaviors in Human-Robot Interaction through Real-Time Auditory and Visual Multiple-Talker Tracking

Hiroshi G. Okuno †,‡**, Kazuhiro Nakadai** ‡**, and Hiroaki Kitano** ‡,∗
† Graduate School of Informatics, Kyoto University, Sakyo, Kyoto 606-8501, Japan, email: okuno@nue.org
‡ Kitano Symbiotic Systems Project, ERATO, Japan Science and Technology Inc.,
M-31 #6A, 6-31-15 Jingumae, Shibuya, Tokyo 150-0001, Japan, email: nakadai@symbio.jst.go.jp
∗ Sony Computer Science Labratories, Shinagawa, Tokyo 141-0022, Japan, email: kitano@csl.sony.co.jp

Abstract

We are studying how to create social physical agents, i.e., humanoids that perform actions empowered by real-time audio-visual tracking of multiple talkers. Social skills require complex perceptual and motor capabilities as well as communicating ones. It is critical to identify primary features in designing building blocks for social skills, because performance of social interaction is usually evaluated as a whole system but not as each component. We investigate the minimum functionalities for social interaction, supposed that a humanoid is equipped with auditory and visual perception and simple motor control but without sound output. Real-time audio-visual multiple-talker tracking system is implemented on the humanoid, SIG, by using sound source localization, stereo vision, face recognition, and motor control. It extracts either auditory or visual streams and associates them by the proximity in localization. Socially-oriented attention control makes the best use of personality variations classified by the Interpersonal Theory of psychology. It also provides task-oriented functions with decaying factor of belief for each stream. We demonstrate that the resulting behavior of SIG invites the users' participation in interaction and encourages the users to explore its behaviors. These demonstrations show that SIG behaves like a physical non-verbal Eliza.

Introduction

Social interaction is essential for (humanoid) robots, because they are getting more common in social and home environments, such as a pet robot in a living room, a service robot at office, or a robot serving people at a party (Brooks *et al.* 1998). Social skills of such robots require robust complex perceptual abilities; for example, it identifies people in the room, pays attention to their voice and looks at them to identify, and associates voice and visual images. Intelligent behavior of social interaction should emerge from rich channels of input sensors; vision, audition, tactile, and others.

Perception of various kinds of sensory inputs should be *active* in the sense that we hear and see things and events that are important to us as individuals, not sound waves or light rays (Handel 1989). In other words, selective attention of sensors represented as looking versus seeing or listening versus hearing plays an important role in social interaction. Other important factors in social interaction are recognition and synthesis of emotion in face expression and voice tones (Breazeal & Scassellati 1999; Breazeal 2001).

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 AAAI, and its related conferences (Nakatani, Okuno, & Kawabata 1994; 1995; Okuno, Nakatani, & Kawabata 1996; 1997; Nakatani & Okuno 1998; Breazeal & Scassellati 1999; Brooks *et al.* 1999; Nakagawa, Okuno, & Kitano 1999; Nakadai *et al.* 2000a; 2000b; 2001; Nakadai, Okuno, & Kitano 2002).

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 (Hansen, Mammone, & Young 1994).
2. **Real-time processing** — Some studies with computational auditory scene analysis (CASA) to understand a mixture of sounds has been done (Rosenthal & Okuno 1998). 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 effect* (Cherry 1953). Real-time processing is essential to incorporate cocktail party effect into a robot. Handel 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"* (Handel 1989).

Nakadai *et al* developed *real-time* auditory and visual multiple-tracking system (Nakadai *et al.* 2001). 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 non-verbal cocktail party robot.

Some robots realize social interaction, in particular, in visual and dialogue processing. Ono *et al.* use the robot named *Robovie* to make common attention between human and robot by using gestures (Ono, Imai, & Ishiguro 2000). Breazeal incorporates the capabilities of recognition and synthesis of emotion in face expression and voice tones into the robot named *Kismet* (Breazeal & Scassellati 1999; Breazeal 2001). Waldherr *et al.* makes the robot named *AMELLA* that can recognize pose and motion gestures (Waldherr *et al.* 1998). Matsusaka *et al.* built the robot named *Hadaly* that can localize the talker as well as recognize speeches by speech-recognition system so that it can interact with multiple people (Matsusaka *et al.* 1999). Nakadai *et al* developed *real-time* auditory and visual multiple-tracking system for the upper-torso humanoid named *SIG* (Nakadai *et al.* 2001).

The contribution of this paper on Cognitve Robotics is that we claim that robot's behaviors depends heavily on hierarchical perceptual capabilities as well as high level cognitive functions for reasoning and planning under dynamic, incompletely known, and unpredictable environments. We will demonstrate that *SIG*'s non-vebal actions based on real-time audio-visual multiple-speaker tracking cause users' complex behaviors during the interaction.

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 reverberations (echoes) occur and speeches are interfered by other sounds or talks. Not only reverberations, but also lighting of illuminating conditions change dynamically, and people are often occluded by other people and reappear.

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. 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 speak 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.

SIG the humanoid

As a test bed 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 (Kitano *et al.* 2000):

a) Cover made of FRP c) A microphone installed in an ear

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

- 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 (Nakadai *et al.* 2000b).

System Description

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

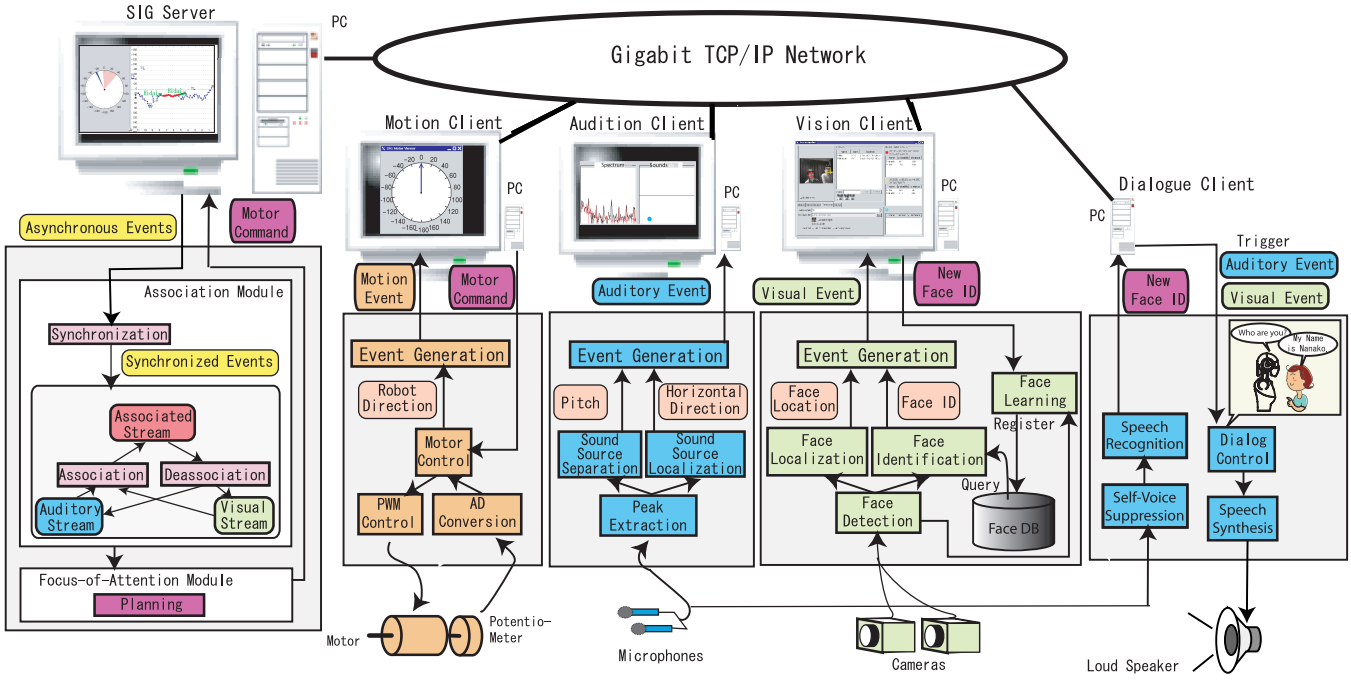


Figure 2: Logical organization of the system composed of SIG server, Motion, Audition, Vision, and Reception clients from left to right.

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 distributed 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.

Since the system should run in real-time, the above clients are physically distributed to four Linux nodes (Pentium-III 1 GHz) connected by TCP/IP over Gigabit Ethernet and 100Base-TX network and run asynchronously.

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 (Nakadai *et al.* 2001). 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 (Nakadai *et al.* 2000a). 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 sound source direction (Nakadai *et al.* 2000a). The ideas are twofold; one is to exploit 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 the belief factors of IPD and IID combined by 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.

Vision: Face identification Module

Since the visual processing detects several faces, extracts, identifies and tracks each face simultaneously, the size, di-

rection 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 (Hidai *et al.* 2000).

The face identification module (see Figure 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 (Hiraoka *et al.* 2000).

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 Figure 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.

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 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 disassociated 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 disassociated to two separate streams.

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. Attention and Dialog control have two modes, socially-oriented and task-oriented. In this paper, a receptionist robot adopts task-oriented, while a companion robot adopts socially-oriented focus-of-attention control. Dialog control uses the automatic speech recognition system, "Julian" developed by Kyoto University (Kawahara *et al.* 1999) and speech synthesis system.

Experiments and Performance Evaluation

The width, length and height of the room of experiment are about 3 m, 3 m, and 2 m, respectively. The room has 6 downlights embedded on the ceiling.

SIG as a receptionist robot

Since the control of a receptionist robot is task-oriented and should focus on the user, the precedence of streams selected by focus-of-attention control is specified from higher to lower as follows:

associated stream $\succ$ *auditory stream* $\succ$ *visual stream.*

One scenario to evaluate the above control is specified as follows: (1) A known participant comes to the receptionist robot. His face has been registered in the face database. (2) He says Hello to *SIG*. (3) *SIG* replies "Hello. You are XXX-san, aren't you?" (4) He says "yes". (5) *SIG* says "XXX-san, Welcome to the party. Please enter the room.".

Figure 3 illustrates four snapshots of this scenario. Figure 3 a) shows the initial state. The loud speaker on the stand is the mouth of *SIG*'s. Figure 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. Figure 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 Figure 3 c). The other is that *SIG* does not pay attention to its own speech. This is attained naturally by the current association algorithm, because this algorithm is designed by taking into account the fact that conversation is conducted 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

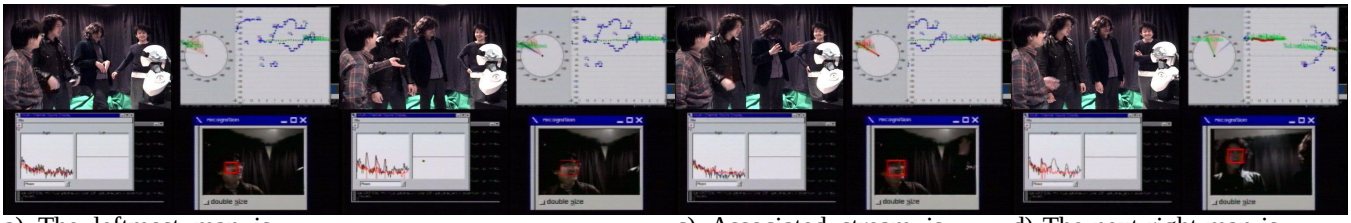
a) Initial state

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



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.

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.

SIG as a companion robot

Since the control of a companion robot is socially-oriented and should pay attention to a new auditory or visual event, the precedence of streams selected by focus-of-attention control is specified from higher to lower as follows:

$$\textit{auditory stream} \succ \textit{associated stream} \succ \textit{visual stream}$$

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 shows that the design of companion robot works well and pays its attention to a new talker. The current system has attained a passive non-verbal companion.

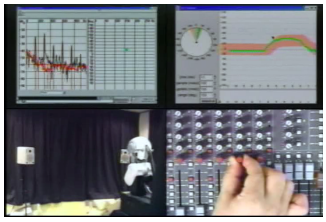


Figure 5: *SIG* follows according to balance control

SIG as simple sound tracker

The current implementation of *SIG* can track a sound source successfully without using visual information. The experiment was performed by changing the balance control of two loud speakers. The sound source is a pure tone of 500 Hz. *SIG* tracks the sound source by changing the balance control as is shown in Figure 5.

Consider that a man is standing out of sight of *SIG* and the sound source controlled by the balance control is moving toward him. When *SIG* finds him, the auditory stream and visual stream are associated and *SIG* stops to keep facing him. Since the sound source is moving, *SIG* suddenly starts moving to track the sound source by dissociating two streams. We are currently investigating the details of this association/dissociation mechanism of streams.

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 phenomenon 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.

SIG's personality in selective attention

In addition, association/dissociation of streams shows a very interesting phenomena of displaying *SIG*'s interest. We are investing this mechanism from the viewpoint of *personality* in selective attention. The *Five-Factor Model* is often used in analyzing the personality of media including software agents (Reeves & Nass 1996). The *big five* dimensions of personality are *Dominance/Submissiveness*, *Friendliness*, *Conscientiousness*, *Emotional Stability*, and *Openness*. Although these five dimensions generally define an agent's basic personality, they are not appropriate to define humanoid's one, because the latter three dimensions cannot be applied to current capabilities of humanoids.

We use the *Interpersonal Theory* instead for defining personality in selective attention. It deals with people's characteristic interaction patterns, varying along the *Dominance/Submissiveness* and *Friendness/Hostility*. The variation is represented by the *interpersonal circumplex*, which is a circular model of the interpersonal domain of personality.

SIG as a non-verbal Eliza

As socially-oriented attention control, interesting human behaviors are observed. The mechanism of associating auditory and visual streams and that of socially-oriented attention control are explained in advance to the user.

1. Some people walk around talking with their hand covering *SIG*'s eyes in order to confirm the performance of auditory tracking.
2. Some people creep on the floor with talking in order to confirm the performance of auditory tracking.
3. Some people play hide-and-seek games with *SIG*.
4. Some people play sounds from a pair of loud speakers with changing the balance control of pre-amplifier in order to confirm the performance of auditory tracking.
5. When one person reads loud a book and then another person starts to read loud a book, *SIG* turns its head to the second talker for a short time and then is back to the first talker and keeps its attention on him/her.

Above observations remind us of Eliza (Weizenbaum 1966), although *SIG* does not say anything except a receptionist robot. When the user says something to *SIG*, it turns to him/her, which invites the participation of the user into interaction. *SIG* also invites exploration of the principles of its functioning, that is, the user is drawn in to see how *SIG* will respond to variations in behavior. Since *SIG* takes only passive behaviors, it does not arouse higher expectations of verisimilitude that it can deliver on.

Needless to say, there are lots of works remaining to validate the proposed approach for personality of artifacts. We are currently working to incorporate active social interaction by developing the capability of listening to simultaneous speeches.

Conclusions and Future Works

In this paper, we demonstrate that auditory and visual multiple-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" (Okuno, Nakatani, & Kawabata 1999), "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.

Although "*Every one knows what attention is.*" (James 1890), but we don't know the details of *SIG*'s behaviors.

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.

References

Breazeal, C., and Scassellati, B. 1999. A context-dependent attention system for a social robot. In *Proceedints of 16th International Joint Conference on Atificial Intelligence (IJCAI-99)*, 1146–1151.

Breazeal, C. 2001. Emotive qualities in robot speech. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2001)*, 1389–1394.

Brooks, R. A.; Breazeal, C.; Irie, R.; Kemp, C. C.; Marjanovic, M.; Scassellati, B.; and Williamson, M. M. 1998. Alternative essences of intelligence. In *Proceedings of 15th National Conference on Artificial Intelligence (AAAI-98)*, 961–968.

Brooks, R.; Breazeal, C.; Marjanovie, M.; Scassellati, B.; and Williamson, M. 1999. The cog project: Building a humanoid robot. In Nehaniv, C., ed., *Computation for metaphors, analogy, and agents*, 52–87. Spriver-Verlag.

Cherry, E. C. 1953. Some experiments on the recognition of speech, with one and with two ears. *Journal of Acoustic Society of America* 25:975–979.

Handel, S. 1989. *Listening*. MA.: The MIT Press.

Hansen, J.; Mammone, R.; and Young, S. 1994. Editorial for the special issue on robust speech processing. *IEEE Transactions on Speech and Audio Processing* 2(4):549–550.

Hidai, K.; Mizoguchi, H.; Hiraoka, K.; Tanaka, M.; Shigehara, T.; and Mishima, T. 2000. Robust face detection against brightness fluctuation and size variation. In *Proceedings of IEEE/RAS International Conference on Intelligent Robots and Systems (IROS 2000)*, 1397–1384.

Hiraoka, K.; Hamahira, M.; Hidai, K.; Mizoguchi, H.; Mishima, T.; and Yoshizawa, S. 2000. Fast algorithm for online linear discriminant analysis. In *Proceedings of ITC-2000*, 274–277. IEEK/IEICE.

James, W. 1890. *The Principles of Psychology*. NY.: Dover, 1950.

Kawahara, T.; Lee, A.; Kobayashi, T.; Takeda, K.; Minematsu, N.; Itou, K.; Ito, A.; Yamamoto, M.; Yamada, A.; Utsuro, T.; and Shikano, K. 1999. Japanese dictation toolkit – 1997 version –. *Journal of Acoustic Society Japan (E)* 20(3):233–239.

Kitano, H.; Okuno, H. G.; Nakadai, K.; Fermin, I.; Sabish, T.; Nakagawa, Y.; and Matsui, T. 2000. Designing a humanoid head for robocup challenge. In *Proceedings of the Fourth International Conference on Autonomous Agents (Agents 2000)*, 17–18. ACM.

Matsusaka, Y.; Tojo, T.; Kuota, S.; Furukawa, K.; Tamiya, D.; Hayata, K.; Nakano, Y.; and Kobayashi, T. 1999. 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)*, 1723–1726.

Nakadai, K.; Lourens, T.; Okuno, H. G.; and Kitano, H. 2000a. Active audition for humanoid. In *Proceedings of 17th National Conference on Artificial Intelligence (AAAI-2000)*, 832–839.

Nakadai, K.; Matsui, T.; Okuno, H. G.; and Kitano, H. 2000b. Active audition system and humanoid exterior design. In *Proceedings of IEEE/RAS International Conference on Intelligent Robots and Systems (IROS 2000)*, 1453–1461.

Nakadai, K. Hidai, K.; Mizoguchi, H.; Okuno, H. G.; and Kitano, H. 2001. Real-time auditory and visual multiple-object tracking for robots. In *Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI-01)*, 1425–1432.

Nakadai, K.; Okuno, H. G.; and Kitano, H. 2002. Exploiting auditory fovea in humanoid-human interaction. In *Proceedings of 18th National Conference on Artificial Intelligence (AAAI-2002)*.

Nakagawa, Y.; Okuno, H. G.; and Kitano, H. 1999. Using vision to improve sound source separation. In *Proceedings of 16th National Conference on Artificial Intelligence (AAAI-99)*, 768–775.

Nakatani, T., and Okuno, H. G. 1998. Sound ontology for computational auditory scene analysis. In *Proceedings of 15th National Conference on Artificial Intelligence (AAAI-98)*, 1004–1010.

Nakatani, T.; Okuno, H. G.; and Kawabata, T. 1994. Auditory stream segregation in auditory scene analysis with a multi-agent system. In *Proceedings of 12th National Conference on Artificial Intelligence (AAAI-94)*, 100–107.

Nakatani, T.; Okuno, H. G.; and Kawabata, T. 1995. Residue-driven architecture for computational auditory scene analysis. In *Proceedings of 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, 165–172.

Okuno, H. G.; Nakatani, T.; and Kawabata, T. 1996. Interfacing sound stream segregation to speech recognition systems — preliminary results of listening to several things at the same time. In *Proceedings of 13th National Conference on Artificial Intelligence (AAAI-96)*, 1082–1089.

Okuno, H. G.; Nakatani, T.; and Kawabata, T. 1997. Understanding three simultaneous speakers. In *Proceedings of 15th International Joint Conf. on Artificial Intelligence (IJCAI-97)*, 30–35.

Okuno, H. G.; Nakatani, T.; and Kawabata, T. 1999. Listening to two simultaneous speeches. *Speech Communication* 27(3-4):281–298.

Ono, T.; Imai, M.; and Ishiguro, H. 2000. A model of embodied communications with gestures between humans and robots. In *Proceedings of Twenty-third Annual Meeting of the Cognitive Science Society (CogSci2001)*, 732–737. AAAI.

Reeves, B., and Nass, C. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge, UK: Cambridge University Press.

Rosenthal, D., and Okuno, H. G., eds. 1998. *Computational Auditory Scene Analysis*. NJ.: Lawrence Erlbaum Associates.

Waldherr, S.; Thrun, S.; Romero, R.; and Margaritis, D. 1998. Template-based recoginition of pose and motion gestures on a mobile robot. In *Proceedings of 15th National Conference on Artificial Intelligence (AAAI-98)*, 977–982.

Weizenbaum, J. 1966. Eliza – a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1):36–45.