Realizing Audio-Visually triggered ELIZA-like non-verbal Behaviors

Hiroshi G. Okuno[1,2], Kazuhiro Nakadai[2], and Hiroaki Kitano[2,3]

[1] Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
okuno@nue.org http://winnie.kuis.kyoto-u.ac.jp/∼okuno/
[2] Kitano Symbiotic Systems Project, ERATO, Japan Science and Technolog Corp.
Mansion 31 Suite 6A, 6-31-15 Jingumae, Shibuya, Tokyo 150-0001 Japan
nakadai@symbio.jst.go.jp http://www.symbio.jst.go.jp/∼nakadai/
[3] Sony Computer Science Laboratories, Inc., Shinagawa, Tokyo 141-0022
kitano@csl.sony.co.jp http://www.symbio.jst.go.jp/∼kitano/

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 not with 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 audio and visual streams 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 funcitons 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 *SIG*'s behaviors. These demonstrations show that *SIG* behaves like a physical non-verbal Eliza.

1 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 [4]. 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 [7]. 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 [3, 2].

Selectivity and *capacity limitation* are two main factors in attention control [19]. A humanoid does some perception intentionally based on selectivity [23]. It also has some limitation in the number of sensors or processing capabilities, and thus only a limited number of sensory information is processed. Since selectivity and capacity limitation are the flip side of the same coin, only selectivity is argued in this paper. Selective attention of auditory processing called the *cocktail party effect* was reported by Cherry in 1953 [6]. At a crowded party, one can attend to one conversation and then change to another one. But the questions are to what one pays one's attention and how one changes one's attention.

Personality in selective attention consists in answers of these questions. Reeves and Nass uses the *Five-Factor Model* in analyzing the personality of media including software agents [20]. 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.

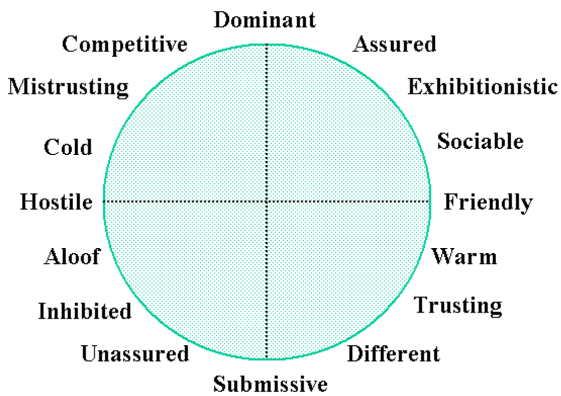

Fig. 1. Interpersonal Circumplex: variation of personality

Fig. 2. *SIG* the Humanoid

We use the *Interpersonal Theory* instead for defining personality in selective attention. It deals with people's characteristic interaction patterns, as is shown in Figure 1, 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 [11].

Physically embodied agents, or humanoid robots have no explicit personality as far as we know. Usually personality is emphasized in language generation, whether verbal or textual. Although the most important human communication means is language, non-verbal sensorimotor based behavior is non-the-less important. In this paper, we use personality to define attention control and report some observations of non-verbal interactions between humanoid and human.

1.1 Related Work

Personality for software agents are studied extensively. Bates and his group propose *believable agents* that can express emotion clearly in appropriately timed manner [1]. Loyall and Bates built engaging characters that allow the viewer to suspend disbelief long enough to interact in interesting ways with the character, or to be engaged by the character's interactions with another computer character [12]. Their system uses language, verbal or in text form. Cassell developed conversational agents that integrate face and gesture [5]. She also argues that implementation of conversational agents should be based on actual study of human-human interaction. Hayes-Roth organizes the Virtual Theater project, which studies the creation of intelligent, automated characters that can act either in well-defined stories or in improvisational environments [8].

Personality for robots are also investigated to widen communication channels in human-robot interaction, although most work do not mention personality explicitly. Ono *et al.* use the robot called *Robovie* to make common attention between human and robot by using gestures [18]. Breazeal incorporates the capabilities of recognition and synthesis of emotion in face expression and voice tones into the robot called *Kismet* [3, 2]. Waldherr *et al.* makes the robot called *AMELLA* that can recognize pose and motion gestures [21]. Matsusaka *et al.* built the robot called *Hadaly* that can localize the talker as well as recognize speeches by speech-recognition system so that it can interact with multiple people [13]. Nakadai *et al* developed *real-time* auditory and visual multiple-tracking system for the upper-torso humanoid called *SIG* [14]. They extended the system to attain in-face interaction by incorporating *auditory fovea* that is the azimuth dependency in resolution of sound source localization [17].

Physically embodied agents, or humanoid robots have no explicit personality as far as we know. Usually personality is emphasized in language generation, whether verbal or not. Although the most important human communication means is language, non-verbal sensori-motor based behavior is non-the-less important. In this paper, we use personality to define attention control and report some observations of non-verbal interactions between humanoid and human.

2 Humanoid Hardware

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:

- 4 DOFs of body driven by 4 DC motors — Each DC motor has a potentiometer to measure the direction.
- A pair of CCD cameras of Sony EVI-G20 for stereo vision input.
- Two pairs of omni-directional microphones (Sony ECM-77S). 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 2) 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 [16].

3 Perceptual Systems in Real-time Multiple-Talker Tracking

The real-time multiple-talker tracking system is designed based on the client/server model (Figure 3). Each server or client executes the following logical modules:

1. Audition client extracts auditory events by pitch extraction, sound source separation and localization, and sends those events to Association.
2. Vision client uses a pair of cameras, extracts visual events by face extraction, identification and localization, and then sends visual events to Association.
3. Motor client generates PWM (Pulse Width Modulation) signals to DC motors and sends motor events to Association.

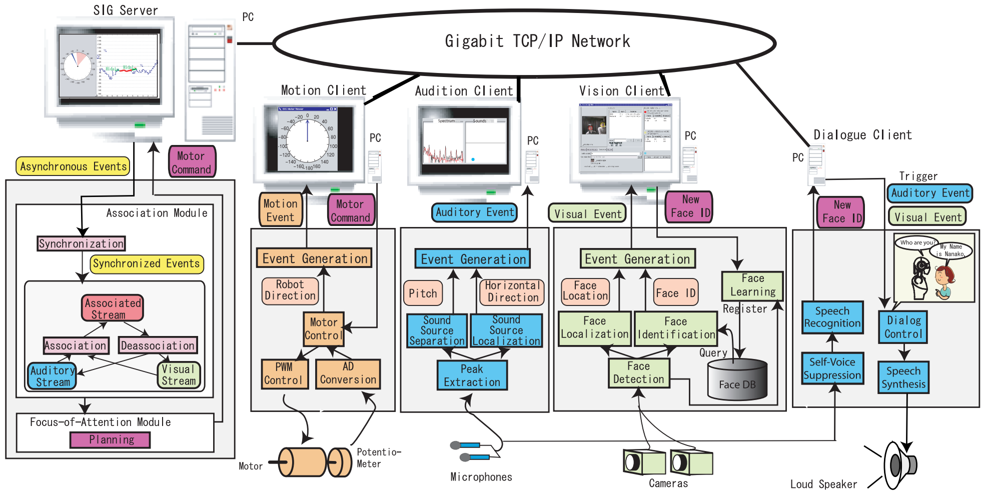

Fig. 3. Hierarchical architecture of real-time audio and visual tracking system

4. Association module groups various events into a stream and maintains association and deassociation between streams.
5. Attention module selects some stream on which it should focus its attention and makes a plan of motor control.
6. Dialog client communicates with people according to its attention by speech synthesis and speech recognition. We use “Julian” automatic speech recognition system [10].

The status of each modules is displayed 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.

To attain real-time tracking, the above modules are physically distributed to five Linux nodes connected by TCP/IP over Gigabit Ethernet TCP/IP network and run asynchronously. The system is implemented by distributed processing of five nodes with Pentium-IV 1.8 GHz. Each node serves Vision, Audition, Motion and Dialogue clients, and SIG server. The whole system upgrades the real-time multiple-talker tracking system [14] by introducing stereo vision systems, adding more nodes and Gigabit Ethernet and realizes social interaction system by designing association and attention control modules.

3.1 Active audition module

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 [15]. 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 [15]. 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 Face recognition and identification module

Vision extracts lengthwise objects such as persons from a disparity map to localize them by using a pair of cameras. First a disparity map is generated by an intensity based area-correlation technique. This is processed in real-time on a PC by a recursive correlation technique and optimization peculiar to Intel architecture [9].

In addition, left and right images are calibrated by affine transformation in advance. An object is extracted from a 2-D disparity map by assuming that a human body is lengthwise. A 2-D disparity map is defined by

$$DM_{2D} = \{D(i,j) | i = 1, 2, \cdots W, j = 1, 2, \cdots H\} \quad (1)$$

where W and H are width and height, respectively and D is a disparity value.

As a first step to extract lengthwise objects, the median of DM_{2D} along the direction of height shown as Eq. (2) is extracted.

$$D_l(i) = Median(D(i,j)). \quad (2)$$

A 1-D disparity map DM_{1D} as a sequence of $D_l(i)$ is created.

$$DM_{1D} = \{D_l(i) | i = 1, 2, \cdots W\} \quad (3)$$

Next, a lengthwise object such as a human body is extracted by segmentation of a region with similar disparity in DM_{1D}. This achieves robust body extraction so that only the torso can be extracted when the human extends his arm. Then, for object localization, epipolar geometry is applied to the center of gravity of the extracted region. 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 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.

4 Attention System

Attention control focuses on one of auditory, visual, or associated streams. It has two modes; socially-oriented and task-oriented. Selective attention is performed according to personality. To define personality, the interpersonal circumplex of the Interpersonal Theory is used. With two mutually independent axes, dormant and friendly, variations of personality are *Dominant, Assured, Exhibitionistic, Sociable, Friendly, Warm, Trusting, Different, Submissive, Unassured, Inhibited, Aloof, Hostile, Cold, Mistrusting,* and *Competitive* (Figure 1) [11].

Since these variations are represented as a circle (circumplex), each variation of personality is represented as a point, (r, θ), inside the interpersonal circumplex, where $0 \leq r \leq 1$ and $0 \leq \theta \leq 2\pi$. Therefore, the value of *Friendly/Hostile* axis and that of *Dominant/Submissive* axis are represented as $r \cos\theta$ and $r \sin\theta$, respectively. Each variation occupies a pie of $\pi/8$. For example, *Friendly* is specified as a pie section of $-\frac{\pi}{16} \sim \frac{\pi}{16}$, and *Dominant* as that of $\frac{3\pi}{16} \sim \frac{5\pi}{16}$.

4.1 Socially-Oriented Attention Control

In this paper, we focus on *passive immediate sensorial attention* in the sense that the stimulus is a sense-impression and no derived ways is considered. We believe that passive attention control with non-verbal interaction is complementary to verbal interaction.

To what the system attend is called "*interested*". The total amount of interest in the system keeps the same and a newly focused stream takes all the amount of *interest* in winner-take-all competition between streams. attention control module selects the stream of the largest *interest*. Three mental factors are defined based on personality.

1. *interest* in a new stream — When a new stream is generated, other streams lose *interest* multiplied by the value of r. Then the new stream gets the total amount of lost *interest*. This means that a robot of large r changes its focus to a new stream. When an auditory stream and a visual one are associated, the sum of *interest* of each stream is given to the associated one.
2. decay of *interest* — The *interest* of a focused stream is reduced at the rate of e^{-kT} every minute, where k is $\{$1.5 - "the value of *Dominant/Submissive*"$\}$/3. The lost *interest* is distributed to other streams.
3. decay of belief — Disappeared stream still remains in the system, because a unseen talker resumes to talk after a short time of silence. If disappeared stream is deleted immediately, the continuity of stream is difficult to maintain. The value of *Friendly/Hostile* is used as the decay factor of belief.

As an example of socially-oriented control, we implement a companion robot. It 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:

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

4.2 Task-Oriented Attention Control

Task-oriented attention control forces Attention to behave according to a specific script. In this paper, we implement a simple receptionist robot. It should focus on the user, the precedence of streams selected by focus-of-attention control is specified from higher to lower as follows:

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

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.".

5 Experiments and Observation

Experiments was done with a small room in a normal residential apartment. 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 on the ceiling.

5.1 Task-oriented interaction: *SIG* as a receptionist robot

a) When a participant comes and says "Hello", *SIG* turns toward him.

b) *SIG* asks his name and he introduces himself to it.

Fig. 4. Temporal sequence of snapshots of *SIG*'s interaction as a receptionist robot

Figure 4 depicts two snapshots of this script. Figure 4 a) shows the initial state. The loud speaker on the stand is the mouth of *SIG*'s. 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 Attention to make a plan that *SIG* should turn to him, and *SIG* does it (Figure 4 b)).

This experiment demonstrates *SIG*'s two interesting behaviors. One is voice-triggered tracking, and the other is that *SIG* does not pay attention to its own speech. As a receptionist robot, once an association is established, *SIG* keeps its face fixed to the direction of the talker 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.

a) The leftmost man says “Hello” and SIG is tracking him.

b) The second right man says “Hello” and SIG turns toward him.

Fig. 5. Temporal sequence of snapshots for a companion robot: scene (upper-left), radar and sequence chart (upper-right), spectrogram and pitch-vs-direction chart (lower-left), and face-tracking chart (lower-right).

5.2 Socially-orineted interaction: *SIG* as a companion robot

When four talkers actually talks spontaneously in attendance of *SIG*, *SIG* tracks some talker and then changes focus-of-attention to others. The observed behavior is evaluated by checking 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 as shown in Figure 5 a)∼b).

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

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 companion. To design and develop an active companion may be important future work.

5.3 Observation: *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 convering *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 performacne 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. Whe one person reads loud a book and then another person starts to read loud a book, *SIG* with *Dominant* personality 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. On the contrary, *SIG* with *Submissive* personality often turns its head to each talker. In either case, the value of r is set to 1.

Above observations remind us of Eliza [22], 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 work remaining to validate the proposed approach for personality of artifacts. We are currently working to incorporate active social interaction by developing the capability of listneing to simultaneous speeches.

6 Conclusions

In this paper, we demonstrate that auditory and visual multiple-talker tracking subsystem can improve social aspects of human robot interaction. Although a simple scheme of behavior is implemented, human robot interaction is drastically improved by real-time multiple-talker tracking system. We can pleasantly spend an hour with *SIG* as a companion robot even if its behavior is quite passive.

Since the Interpersonal Theory research community provides software for analysing circumplex correlation matrices, we have plan to gather the data of user interaction to evaluate whether the presented architecture of selective attention based on personality realizes the target variation of personality. In this persuit may lead to a general theory of personality for software agents and humanoid robots.

Acknowledgments We thank our colleagues of Symbiotic Intelligence Group, Kitano Symbiotic Systems Project, in particular, Kenichi Hidai (currently with Sony DCL), for their discussions. We also thank Prof. Tatsuya Kawahara of Kyoto University for his supports in using “Julius” automatic speech recognition system.

References

1. BATES, J. The role of emotion in believable agents. *Communications of the ACM 37*, 7 (1994), 122–125.
2. BREAZEAL, C. Emotive qualities in robot speech. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2001)* (2001), IEEE, pp. 1389–1394.
3. BREAZEAL, C., AND SCASSELLATI, B. A context-dependent attention system for a social robot. In *Proceedints of 16th International Joint Conference on Atificial Intelligence (IJCAI-99)* (1999), pp. 1146–1151.
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. CASSELL, J. More than just another pretty face: Embodied conversational interface agents. *Communications of the ACM 43*, 4 (2000), 70–78.
6. 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.
7. HANDEL, S. *Listening*. The MIT Press, MA., 1989.
8. HAYES-ROTH, B., BALL, G., LISETTI, C., PICARD, R., AND STERN, A. Affect and emotion in the user interface. In *Proceedings of 1998 International Conference on Intelligent User Interfaces* (1998), ACM, pp. 91–96.
9. KAGAMI, S., OKADA, K., INABA, M., AND INOUE, H. Real-time 3d optical flow generation system. In *Proc. of International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI'99)* (1999), pp. 237–242.
10. 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.
11. KIESLER, D. The 1982 interpersonal circle: A taxonomy for complementarity in human transactions. *Psychological Review 90* (1993), 185–214.
12. LOYALL, A. B., AND BATES, J. Personality-rich believable agents that use language. In *Proceedings of the First International Conference on Autonomous Agents* (1997), pp. 106–113.
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., HIDAI, K., MIZOGUCHI, H., OKUNO, H. G., AND KITANO, H. Real-time auditory and visual multiple-object tracking for robots. In *Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI-01)* (2001), vol. II, IJCAI, pp. 1425–1432.
15. 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.
16. 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.
17. NAKADAI, K., OKUNO, H. G., AND KITANO, H. Exploiting auditory fovea in humanoid-human interaction. In *Proceedings of 18th National Conference on Artificial Intelligence (AAAI-2002)* (2002), AAAI, *(in print)*.
18. ONO, T., IMAI, M., AND ISHIGURO, H. A model of embodied communications with gestures between humans and robots. In *Proceedings of 23rd Annual Meeting of the Cognitive Science Society (CogSci2001)* (2000), AAAI, pp. 732–737.
19. PASHLER, H. *The Psychology of Attention*. The MIT Press, MA., 1997.
20. REEVES, B., AND NASS, C. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, Cambridge, UK, 1996.
21. 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.
22. WEIZENBAUM, J. Eliza – a computer program for the study of natural language communication between man and machine. *Communications of the ACM 9*, 1 (1966), 36–45.
23. WOLFE, J., CAVE, K. R., AND FRANZEL, S. Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance 15*, 3 (1989), 419–433.