Social Interaction of Humanoid Robot Based on Audio-Visual Tracking

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

[1] Kitano Symbiotic Systems Project, ERATO, Japan Science and Technolog Corp.
Mansion 31 Suite 6A, 6-31-15 Jingumae, Shibuya, Tokyo 150-0001 Japan
{*okuno, nakadai, kitano*}*@symbio.jst.go.jp*
[2] Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
[3] Sony Computer Science Laboratories, Inc., Shinagawa, Tokyo 141-0022

Abstract. Social interaction is essential in improving robot human interface. Such behaviors for social interaction may include to pay attention to a new sound source, to move toward it, or to keep face to face with a moving speaker. Some sound-centered behaviors may be difficult to attain, because the mixture of sounds is not well treated or auditory processing is too slow for real-time applications. Recently, Nakadai *et al* have developed real-time auditory and visual multiple-talker tracking technology by associating auditory and visual streams. The system is implemented on a upper-torso humanoid and the real-time talker tracking with 200 msec of delay is attained by distributed processing on four PCs connected by Gigabit Ethernet. Focus-of-attention is programmable and allows a variety of behaviors. This paper demonstrates a receptionist robot by focusing on an associated stream, while a companion robot on an auditory stream.

1 Introduction

Social interaction is essential for humanoid robots, because such robots are getting 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 [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 visually, 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. 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.

In this paper, we focus on audition, or sound input in localizing and tacking talkers. 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 [2, 3, 11, 12]. One of social intelligent behavior is that a robot can attend one conversation at a crowded party and then attend another one. This capability is well known as the *cocktail party effect*.

Some robots are equipped with improved robot-human interface. *AMELLA* [14] 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 emotion in facial or voice expression. *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 [8] 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 [5]. 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).

The reason why the systems developed so far do not support auditory tracking of talkers is that sound input consists of a mixture of sounds. The current technologies concerning sound source separation from a mixture of sounds requires a lot of restriction on an implementation of sound source separation system. In addition, such an implementation usually does not run in real-time in a dynamically changing environment.

Nakadai *et al* developed *real-time* auditory and visual multiple-tracking system [9]. 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 real-time multiple-talke tracking system. Section 3 demonstrates the system behavior of social interaction. Section 4 discusses the observations of the experiments and future work, and Section 5 concludes the paper.

2 Real-time Multiple-Talker Tracking System

2.1 SIG the humanoid

Fig. 1. *SIG* the Humanoid plays as a companion robot

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 visual stereo 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 1) 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 [11].

2.2 Architecture of real-time audio and visual tracking system

The system is designed based on the client/server model (Fig. 2). 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.
4. Association module groups various events into a stream and maintains association and deassociation between streams.

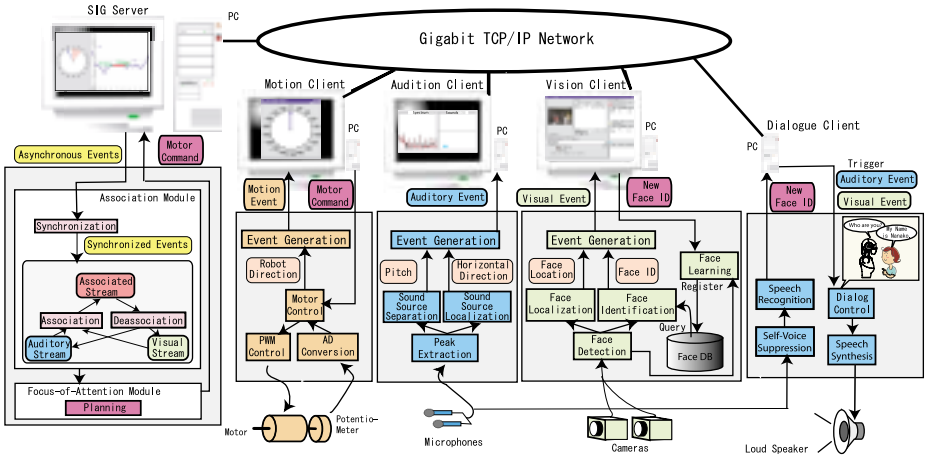


Fig. 2. Hierarchical architecture of real-time audio and visual tracking system

5. Focus-of-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 [7].

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.

Since the system should run in real-time, 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 [9] by introducing stereo vision systems, adding more nodes and Gigabit Ethernet and realizes social interaction system by designing association and focus-of control modules.

2.3 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 [10]. 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 [10]. 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.

2.4 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 [6].

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

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

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.

$$D_l(i) = Median(D(i,j)). \tag{2}$$

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

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

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 creates stereo vision events which consist of distance, azimuth and observation time.

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.

2.5 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.

2.6 Focus-of-Attention and Dialog Control

Focus-of-Attention control is programmable 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 detailed design of each algorithm depends on applications, the focus-of-attention control 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.

3 Design and Experiments of Some Social Interactions

For evaluation of the behavior of *SIG*, one scenario for the receptionist robot and one for the companion robot are designed and executed. The first scenario examines whether an auditory stream triggers Focus-of-Attention to make a plan for *SIG* to turn toward 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.

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.

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

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

Fig. 3. Temporal sequence of snapshots of *SIG*'s interaction as a receptionist robot

3.1 *SIG* as a receptionist robot

The precedence of streams selected by focus-of-attention control as a receptionist robot 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.".

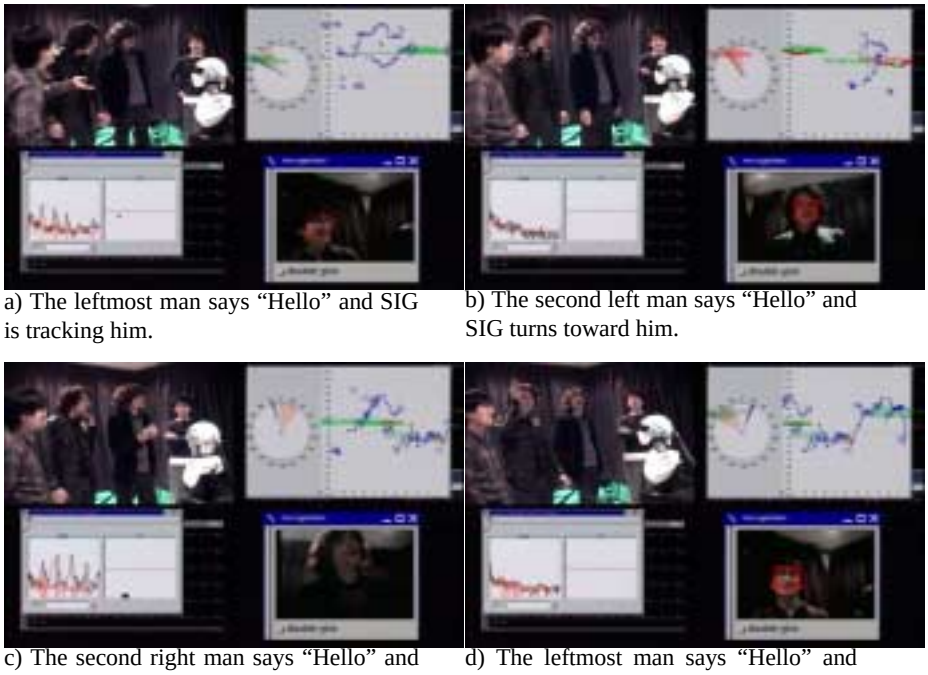
Fig. 3 illustrates four snapshots of this scenario. Fig. 3 a) shows the initial state. The speaker on the stand is the mouth of *SIG*'s. Fig. 3 b) shows when a participant comes to the receptionist, but *SIG* has not noticed him yet, because he is out of *SIG*'s sight. When he speaks to *SIG*, Audition 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 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. In this case, *SIG* asks his name and registers his face and name in the face database.

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.



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

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

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

d) The leftmost man says “Hello” and SIG turns toward him.

Fig. 4. 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).

3.2 *SIG* as a companion robot

The precedence of streams selected by focus-of-attention control as a companion robot is as follows:

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

There is no explicit scenario for evaluating the above control. 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 as shown in Figure 4 a)∼d).

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.

4 Conclusion

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 application of auditory and visual multiple-talker 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. One of important future work is automatic acquisition of social interaction patterns by supervised or unsupervised learning. This capability is quite important to provide a rich collection of social behaviors. Other future work includes applications such as "listening to several things simultaneously" [13], "cocktail party computer", integration of auditory and visual tracking and pose and gesture recognition, and other novel areas.

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. HUANG, J., OHNISHI, N., AND SUGIE, N. Building ears for robots: sound localization and separation. *Artificial Life and Robotics 1*, 4 (1997), 157–163.
6. 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.
7. 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.
8. 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.
9. NAKADAI, K., HIDAI, K., MIZOGUCHI, H., OKUNO, H., 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), MIT Press, pp. 1424–1432.
10. 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.
11. 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.
12. OKUNO, H., NAKADAI, K., LOURENS, T., AND KITANO, H. Sound and visual tracking for humanoid robot. In *Proceedings of Seventeenth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE-2001)* (Jun. 2001), vol. Lecture Notes in Artificial Intelligence 2070, Springer-Verlag, pp. 640–650.
13. OKUNO, H. G., NAKATANI, T., AND KAWABATA, T. Listening to two simultaneous speeches. *Speech Communication 27*, 3-4 (1999), 281–298.
14. 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.

This article was processed using the LaTeX macro package with LLNCS style