

# Humanoid with Interaction Ability Using Vision and Speech Information

Junichi Ido\*, Ryuichi Nisimura\*\*, Yoshio Matsumoto\*  
and Tsukasa Ogasawara\*

*\*Nara Institute of science and technology,*

*\*\*Wakayama university*

*Japan*

## 1. Introduction

Recently, there are many research on harmonized human robot interaction in the real environment (Honda; Suzuki et al., 2002). Speech recognition is useful for human-robot communication, and there are many robots that have such interface (Asoh et al., 2001) (Matsusaka et al., 2001). Some interface use non-verbal information such as facial expression and gaze, which are also seen as important for interaction. We have developed a reception guidance humanoid robot “ASKA”, which can interact with humans using verbal and non-verbal information such as gaze direction, head pose and lip motion (Ido et al., 2003).



Fig. 1. Humanoid robot HRP-2 interacting with a user using vision and speech information

In this paper, we introduce a humanoid robot system for human-robot communication research. Fig.1 shows the overview of our humanoid robot HRP-2. This humanoid robot system with interaction ability was developed at NAIST (Nara Institute of Science and Technology) under the collaboration of Robotics and Speech Laboratories. It is used as a research platform to develop an intelligent real-world interface using various information

technologies studied in our institute. The following functions are implemented for human-robot interaction:

1. Speech recognition
2. Voice synthesizing
3. Facial information measurement
4. Portrait drawing
5. Gesture recognition

The dialogue system which use a large vocabulary continuous speech recognition, and the eye contact system, which use a facial information measurement system, are the unique features of this robot. The rest of this paper is organized as follows: First, the design concepts are discussed in Section 2. The hardware and software system configuration are described separately in Section 3. In Section 4, the voice interface implemented in this system is explained. The interaction modules using visual information are explained in Section 5. In section 6, are explained the demonstration and the experiment in which a person interacts with the humanoid. Finally, we summarize our research and future works in Section 7.

## **2. Design concept**

Our robot system has been designed based on two concepts: (1) as a research platform for various information technologies, and (2) as an achievement of human-robot interaction.

### **2.1 Research platform**

The main objective in the early stage of its development was to build a research platform for various information technologies using a robot. The software architecture was designed based on this concept. Fig.4 shows a simple configuration in which each module communicates its own status and sensory information to the server. Each module runs independently and can start and stop at an arbitrary timing. This modularity enables rapid development and easy maintenance of the modules.

### **2.2 Human-robot interaction**

The information utilized for face-to-face communication is classified in two major categories, "verbal" and "non-verbal" information. Although the primary information in communication is the former, the latter, such as facial direction, gaze and gesture, is recently emphasized as a mean of natural human-robot interaction. We focus on face direction and gesture information in this research and try to achieve more natural interaction by combining them with speech information. In the next section, we describe how the software and the hardware of our system are constructed. The typical scenario of the interaction is also described.

## **3. System configuration**

### **3.1 Hardware configuration**

The system is composed of a humanoid body, stereo cameras, hand-held microphones, a speaker and several PCs as shown in Fig.2. HRP-2 (KAWADA Industries, Inc.) is used as the humanoid body. A stereo camera system with four IEEE1394 cameras (Flea, Point Grey Research Inc.), eight tiny microphones and an 8ch A/D board (TD-BD-8CSUSB, Tokyo Electron Device Ltd.) are installed in the head of HPR-2. Eight built-in microphones attached to the head are connected to the on-board vision PC via A/D board, and 8ch speech signals

can be captured simultaneously. Additionally, a hand-held microphone can be connected to an external PC for speech recognition. Switching between these two microphone systems is achieved by software. The use of the hand-held microphone enables the interaction in places where the background noise is large to such an extent that recognition using the built-in microphone fails. Two external PCs are used besides the PC built into the robot. One of them is used for the speech recognition and speech synthesis, and the other is used as the terminal PC of the robot.

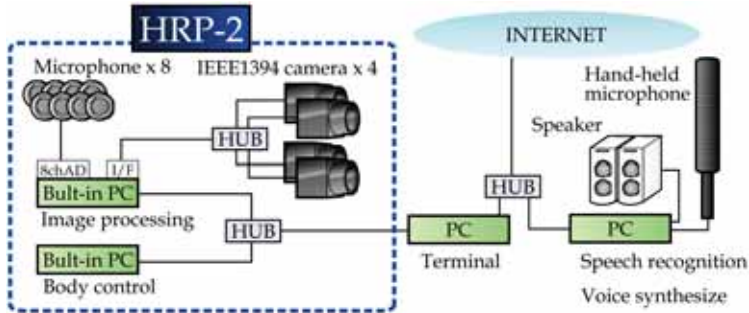


Fig. 2. Hardware configuration

A special chair, as shown in Fig. 3, was built in order for HRP-2 to sit down drawing the experiment. Regrettably, HRP-2 cannot seat itself because it has to be bolted to the chair for stability as shown Fig. 3.



Fig. 3. HRP-2 fixed on a chair

### 3.2 Software configuration

The basic software of this system consists of six modules:

- Speech Recognition Module
- Speech Synthesizing Module

- Face Tracking Module
- Gesture Recognition Module
- Portrait Drawing Module
- Body Gesture Controller Module

In our system, the Face Tracking Module, the Gesture Recognition Module and the Portrait Drawing Module are all used as Visual Measurement Modules, and they are connected to the vision subserver. The speech recognition module has an independent interface called “adintool” to record, split, send and receive speech data. These interfaces enable to select the speech input with no influence on the other modules.

These modules run on the distributed PCs and communicate with a server program by socket communication over TCP/IP protocols as shown in Fig. 4. This is a simple implementation of the blackboard system (Nii, 1986). The server collects all the information (sensory information and status of execution) from all the client modules. Each client module can access the server to obtain any information in order to decide what actions to take. Each module runs independently and can start and stop at an arbitrary timing. This modularity enables the rapid development and easy maintenance of the modules.

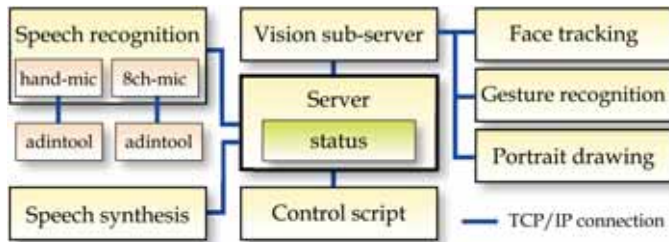


Fig. 4. Software configuration

### 3.3 Interaction scenario

HRP-2 sitting opposite to a user across a table can detect face and gaze directions of the user and recognize the question asked by the user. The typical scenario of the interaction between a user and the humanoid is as follows:

1. The humanoid detects the direction of the user's face.
2. When the face direction of the user is detected to be facing the humanoid, the user is regarded as having an intention to talk to the humanoid. The user can then talk with gestures to humanoid.
3. The humanoid recognizes the question and makes a response with voice and gesture or carries out an ordered task.

The speech dialogue system of the humanoid can answer the following questions:

- Office and laboratory locations
- Extension telephone numbers of staffs
- Locations of university facilities
- Today's weather report, news and current time
- Greetings

In addition to these questions, commands such as “pass the objects” or “draw a portrait” can be recognized and carried out. The training corpus used for speech recognition is described in the following section.

The motions comprising the gesture responses are defined beforehand using a dedicated software, "Motion Creator" (Nakaoka et al., 2004). These responses are linked to its corresponding sentences manually.

#### 4. Voice interface using speech and noise recognition

The voice interface of our system was developed to contain two parallel sound recognition methods to be able to have flexible interactions with users. We implemented a spoken dialogue routine based on a continuous speech recognition technology with a large vocabulary dictionary for accepting users' various utterances. We also introduced a nonstationary noise recognition program based on likelihood measurements using Gaussian Mixture Models (GMMs). It realizes not only rejection mechanisms of environmental noises, but also novel human-robot interaction schemes by discerning unintended user's voices such as laughter, coughing, and so on. This section explains about the speech recognition and the noise recognition.

##### 4.1 Speech recognition

The continuous speech recognition has accomplished remarkable performance. However, sufficient accuracy when recognizing natural spontaneous utterances has not been attained yet. To obtain higher accuracy, we needed to organize task-suitable statistical models beforehand.

Our speech recognition engine, "Julius" (Lee et al., 2001) requires a language model and an acoustic model as statistical knowledge.

For an acoustic model, we use the speaker-independent PTM (Lee et al., 2000) triphone HMM (Hidden Markov Model). The model can deal with an appearance probability of phonemes with considering context dependent co-articulations consisting of the current phoneme and its left and right phonemes.

An acoustic model for the HRP-2 was trained from the following data using HTK (Hidden Markov Model Toolkit) (Young et al., 2002):

- **Dialog** Natural users' utterances in using actual dialogue system (24,809 utterances).
- **JNAS** Reading style speech by speakers, extracted from the JNAS (Japanese Newspaper Article Sentences) (Itou et al., 1999) database (40,086 utterances).

**Dialog** data are actual human-machine dialogue data extracted from utterance logs collected by a long-term field test of our spoken dialogue system "Takemaru-kun System" (Nisimura et al., 2005), which has been deployed in a public city office since November 2002 and operated every business day. We have obtained over 300,000 recorded inputs as of February 2005. The accuracy improvement of natural utterance recognition can be obtained efficiently by using these actual conversation data. We can also say that the built model can obtain better performance for recognition of child voices because the **Dialog** data contains many voices uttered by children. See (Nisimura et al., 2004) for details.

The training data of the acoustic model included the **JNAS** data due to the necessities of holding a large amount of speech data in building the model.

We adopted a word trigram model as the language model, which is one of the major statistical methods in modeling appearance probabilities of a sequence of words (Jelinek, 1990). There are two well-known task description approaches in continuous speech recognition: (1) finite state network grammar, and (2) word trigram language model. Finite state network grammar is usually adopted for small restricted tasks. By using a statistical

method instead of a network grammar, some utterances even in out-of-domain task are correctly recognized. Utterances including various expression styles can also be recognized more flexibly than with the network grammar based recognition. In order to train the model, a training corpus consisting of the following texts was prepared:

- **Dialog** Transcribed utterances collected by the field testing Takemaru-kun system (15,433 sentences).
- **Web Texts** extracted from web pages (826,278 sentences).
- **Chat Texts** extracted from Internet Relay Chat (IRC) logs (2,720,134 sentences).
- **TV Texts** of request utterances in operating a television through a spoken dialogue interface (4,256 sentences).

We produced a vocabulary dictionary which includes 41,443 words, each appearing 20 or more times in the corpus. Then, language model tools provided from IPA Japanese free dictation program project (Itou et al., 2000) was used to build a baseline model. Finally, a task-dependent network grammar was adapted to the model. We wrote a finite state network grammar for the HRP-2 task, which included 350 words. Adaptation was performed by strengthening the trigram probabilities in the baseline model on the basis of word-pair constraints in the written grammar. This method enables more accurate recognition of in-task utterances while keeping the acceptability of statistical model against unexpected utterances.

Class	# of training data
Adult voice	7,497
Child voice	7,503
Laughter	849
Coughing	321
Beating by hand	101
Beating by soft hammer	104
Background noise	5,000
Other noise	6,380

Sampling rate/bit	16 kHz, 16 bit
Window width/shift	25/19 msec
Parameter	MFCC (12 dim.), $\Delta$ MFCC, $\Delta$ Power
Mixtures of Gaussian	64

Table 1. Training Conditions of GMMs

#### 4.2 Noise recognition

We introduced noise recognition programs to the HRP-2 to realize a novel human-robot interaction that mediates unintended sound inputs, such as coughing, laughing, and other impulsive noises. Although noises have been deleted as needless inputs in a general dialogue system (Lee et al., 2004), the proposed system can continue to dialogue with humans while recognizing a noise category.

We investigated sound verification to determine whether the inputted voice was intended by comparison of acoustic likelihood given by GMMs. GMMs have proven to be powerful for text-independent speech verification technique. Although conventional speech verification studies have only focused on environmental noises, our previous studies found that GMMs can also discriminate more utterance-like inputs.

Table 1 shows the training conditions of GMMs, where training data were recorded through a microphone used when performing a spoken dialogue for the HRP-2. When laughter or coughing is recognized, the response corresponding to the recognition result is returned to

the user. To realize the identification of the voice and non-voice, adult and child's voices were included in the training data. If the input is identified as voice, the system executes a normal spoken dialogue routine. "Beating by hand" and "Beating by soft hammer" indicate impulsive noises when a user beats the head of HRP-2 by hand or by a soft hammer. The system will use the identification result of beatings for dealing with mischief from users when the robot is installed in a house.

8-class GMMs with 64 Gaussian mixtures were made from each class training data. As for an acoustic parameter, we adopted the mel frequency cepstral coefficients (MFCC), which is a major parameter when analyzing human voices for speech recognitions. The class of GMM that has the highest acoustic likelihood against parameters of input sound is chosen as an output.

### 4.3 Dialogue strategy

The spoken dialogue strategy of our system was designed based on a simple principle. Candidate responses to a user's question are prepared beforehand. Selection of a suitable response among the candidates is performed by keyword or key-phrase matching mechanism. We defined keywords for each candidate. After recording the user's voice, the number of keywords matched with recognized text is totaled for all prepared candidates. The system will choose the most matched candidate as a response. In this procedure, the N-best output is used as the speech recognition result that complements recognition errors.

## 5. Interaction system using visual information

Our system obtains gray-scale images from the stereo camera system installed in the head, and the following three vision based functions were implemented; facial information measurement, pointing gesture recognition and portrait drawing. These functions are described in the following sections.

### 5.1 Facial information measurement

The face and gaze information provides important information showing intentions and interests of a human. In a previous study, it is shown that humans tend to be conscious of an object at the time of utterance (Kendon, 1967). Facial module is based on a facial measurement system (Matsumoto et al., 2000) and sends measured parameters such as the pose and the position of the head and the gaze direction to the server via network. This module tracks the face using a 3D facial model of the user, and measures various facial information such as head position and orientation, gaze direction, blinks and lip motions. Fig. 5 illustrates the 3D facial model, which consists of template images of facial features and their 3D coordinates. The position and orientation of the head is calculated by fitting the 3D facial model to the set of 3D measurements of the facial features based on the following equation:

$$E = \sum_{i=0}^{N-1} w_i (Rx_i + t - y_i)^T (Rx_i + t - y_i)$$

where  $E$  is the fitting error,  $N$  is the number of facial features,  $x_i$  is the position vector of each feature in the 3D facial model,  $y_i$  is the measured 3D position of the corresponding feature obtained from the current image, and  $w_i$  is the reliability of the measurements.  $T$  and  $R$  are the translation vector and the rotation matrix to be estimated. The problem of achieving the best fitting boils down to finding a set of  $T$  and  $R$  which minimizes the fitting error  $E$ , and can be solved by Steepest Descent Method.

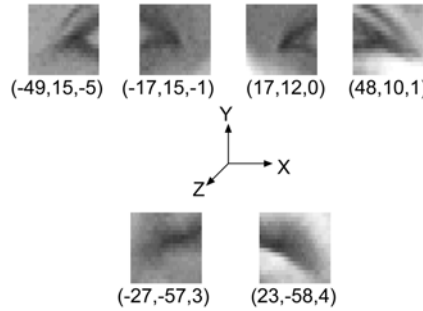


Fig. 5. 3D facial model

How to estimate the gaze direction is illustrated in Fig. 6. As the position and orientation of the head is estimated, the position and the size of the eyeball in the image can be estimated assuming that it is located at a fixed position inside the head. If the position of the iris (or the pupil) can be detected in the image, the relative (vertical and horizontal) position of the iris and the center of the eyeball in the image produces the 3D direction of the gaze. Fig. 7 shows how the facial information is measured. In this figure, rectangles indicate feature areas in a face utilized for tracking, and the two lines indicate gaze direction. The main purpose of this measurement is to detect valid speech period. The speech input is recognized only after the user turns his face to HRP-2. Therefore, the robot does not recognize utterances directed to other people. We regard this function as a simple implementation of “eye contact.”

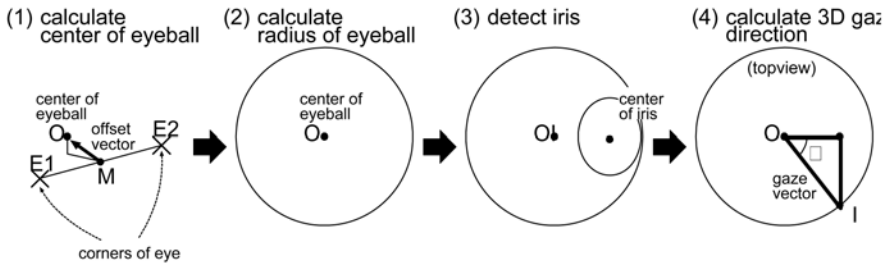


Fig. 6. Modeling of gaze

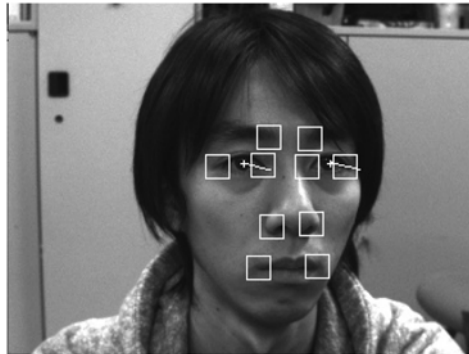


Fig. 7. Facial information measurement

### 5.2 Point gesture recognition

The gesture recognition module which recognizes simple pointing gesture is described. Gestures such as motion of the head help attain clearer communication. Furthermore, gestures which points to directions are important when considering guiding tasks. If the robot can recognize only speech it is difficult to make natural communication because demonstrative pronouns are often used in such a situation. Pointing gesture recognition module used depth information generated by correlation based on SAD (Sum of Absolute Difference). Fig. 8 is an example of the disparity map. The process of recognizing the pointing gesture is as follows:

1. The disparity map is generated after correcting for lens distortion.
2. The pointing direction is detected on the supposition that the closest part of the user to the robot in the disparity map is the user's hand.

The recognition of the pointing gesture enables HRP-2 to respond, even if a user gives questions with demonstrative pronoun. For example, HRP-2 can choose and pass a proper newspaper to the user when it is asked "Pass me that newspaper" with a pointing gesture.

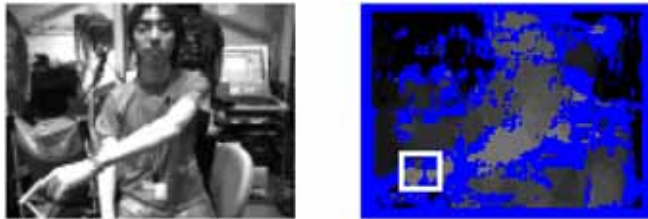
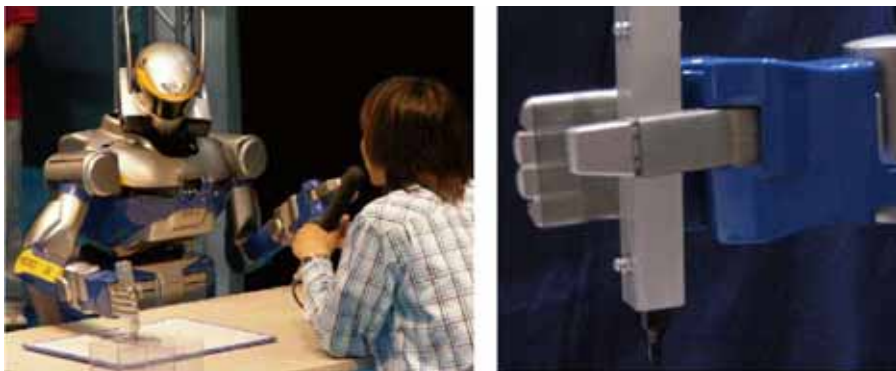


Fig. 8. Pointing gesture recognition based on depth image

### 5.3 Portrait drawing

The third module using vision information, the portrait drawing module, is described here. This module literally provides the functionality of drawing a portrait as shown in Fig. 9. This function was implemented to show that HRP-2 can perform skillful tasks with its motion in addition to communicating with a user in the demonstration. From a technical viewpoint, portrait drawing requires the segmentation of the face region from the background. The procedures to draw the portrait of a user are as follows:



(a) HRP-2 while drawing portrait

(b) Pen holder

Fig. 9. HRP-2 drawing portrait of a user

1. When the module detects using the foregoing facial information that the user has turned his face to HRP-2, a still image of the user is captured.
2. A canny edge image, a depth mask image and an ellipsoid mask image are generated.
3. The face region is extracted from the edge image using the mask data.
4. The face image is converted to a sequence of points by chain method.
5. The sequence of points are sorted and thinned.
6. HRP-2 draws a portrait using the generated data.

When the user requests a portrait to be drawn by HRP-2, it asks the user to pass the pen and to turn the user's face toward the robot. Then it captures an image as shown in Fig.10 (A). An edge image using Canny's algorithm (Fig. 10 (B)) is generated by the appropriate face image. A depth mask image (Fig. 10 (C)), and an ellipsoid mask image (Fig. 10 (D)) are generated by two kinds of data, the stereo image pair and the measurement value of face position. Fig. 10 (E) shows the facial part extracted from the whole edge image using the masks. The portrait is drawn on an actual white-board by HRP-2 using the sequence data generated. Inverse kinematics for eight degrees of freedom is solved under the condition that the pose of the pen is kept vertical. After the sequence of hand positions is determined, the hand moves interpolating these points. Fig. 10 (F) shows the resulting image drawn on the whiteboard. When HRP-2 actually draws a portrait, it uses a felt-tip pen with a holder that was designed to help it grasp and absorb the position errors of the hand by a built-in spring (Fig. 9 (b)).

## 6. Experiment

### 6.1 Facial information while talking with the robot

We implemented the dialog system using simple "eye contact." However, it is not clear whether a user looks at the robot's face when talking with the humanoid robot. Therefore we conducted an interaction experiment in a public exhibition to answer this question. To investigate the facial information of users while talking with the humanoid robot, images from the camera attached to its head were stored for subsequent analysis. Some examples of these images are shown in Fig.11. Subjects were fifteen visitors who consist of five men, five women and five children. After we gave a simplified explanation about the robot, users talked with the robot freely. In this experiment, the robot sitting opposite to a user across a table always respond to the user's utterances without face and gaze information. Users spoke to the robot for about 14 seconds on average, which included about 10 sentences. Users sometimes spoke to other people such as staffs or their accompanying persons. The total time for talking to others beside the robot averages about 4 seconds per person.

We analyzed how often users turned their face and gaze to the robot's face when speaking to the robot. As a result, users gazed at the robot at a rate of 69% and turn their face to it at the rate of 95% on average when speaking to the humanoid robot.

This result shows that people tend to look at the face when they talk to a humanoid robot. It also indicates that our dialog system using eye contact works well regardless of user's age or gender.

### 6.2 Interaction experiment with two users

In order to investigate accuracy of our dialogue system using "eye contact", we experimented on an interaction with two people. Two users sat opposite to the robot across a table. One of them talked to the other and to the robot based on the scenario given before. The built-in microphones were used for speech recognition. The number of subjects was 10 pairs. The scenario was composed of 7 "person-to-person" sentences, 4 "person-to-robot" sentences and 4 responses from the robot.

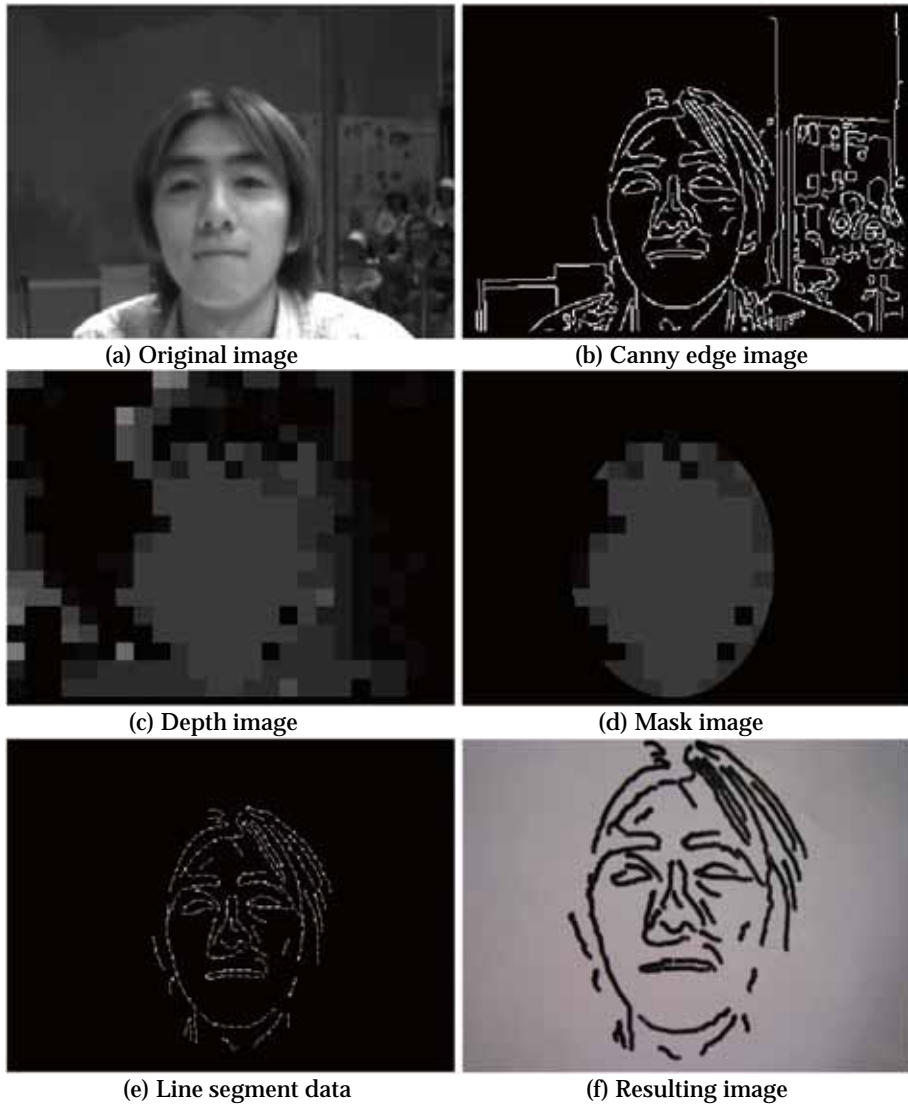


Fig. 10. Portrait drawing process



Fig. 11. Visitors talking to the humanoid robot

Fig.12 shows the ratio of false responses to the conversations between the users. Without using eye contact, this ratio was 75.7[%] on average, and it dropped down to 4.3 [%] when the robot utilized eye contact information effectively. This result shows that the utilization of eye contact information improved the accuracy of response.

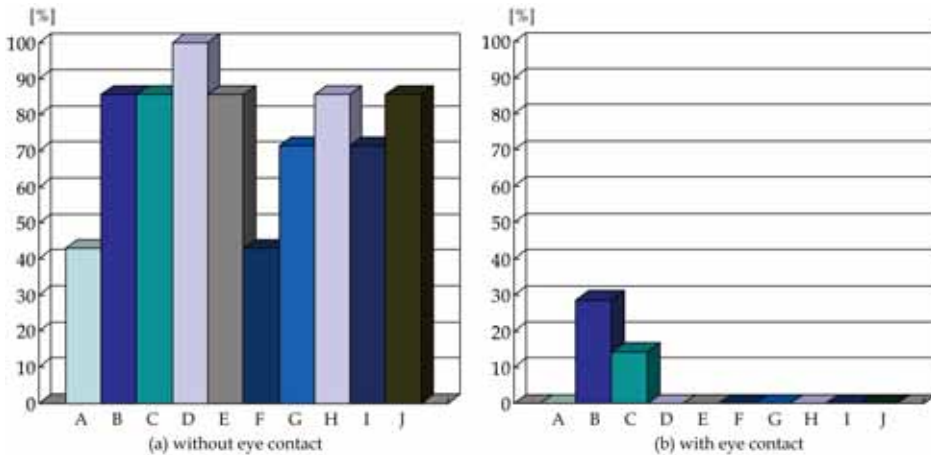


Fig. 12. Ratio of false response

### 6.3 Demonstration

We verified the usefulness of our system in a real environment through a demonstration in the Prototype Robot Exhibition at Aichi World EXPO 2005 as shown in Fig.13. The exhibition area was so noisy, full of audience and with simultaneously held demonstrations that hand-held microphone was utilized in the demonstration. The nonstationary noise recognition system was utilized for recognizing users' coughing to start talking about cold in the demo scenario. When connected to the Internet, HRP-2 can answer questions on weather information, news headlines and so on.

The uncontrollable lighting condition was also crucial for image processing. However, since our method does not rely on skin color detection which is known to be sensitive to lighting condition, the face measurement and gesture recognition was robust enough in such an environment. HRP-2 was also able to draw a portrait by extracting the face of the user from cluttered background. Our demonstration was successfully carried-out for two weeks without problems.

## 7. Conclusion

The HRP-2 is a speech-oriented humanoid robot system which realizes natural multi-modal interaction between human and robot. This system has a vision and a speech dialogue system to communicate with visitors. The voice interface that has two aspects was implemented on the HRP-2 to realize flexible interactions with users. One is the spoken dialogue routine based on a continuous speech recognition technology with a large vocabulary dictionary, and the other is a non-stationary noise recognition system. We also implemented the face measurement function in order for the humanoid to realize "eye contact" with the user. In addition, the pointing gesture recognition function was

implemented based on depth-map generation. By integrating speech information and gesture information, HRP-2 can recognize questions that include a demonstrative pronoun. The feasibility of the system was demonstrated at EXPO 2005. Some issues and demands have been gradually clarified by the demonstration and the experiment.

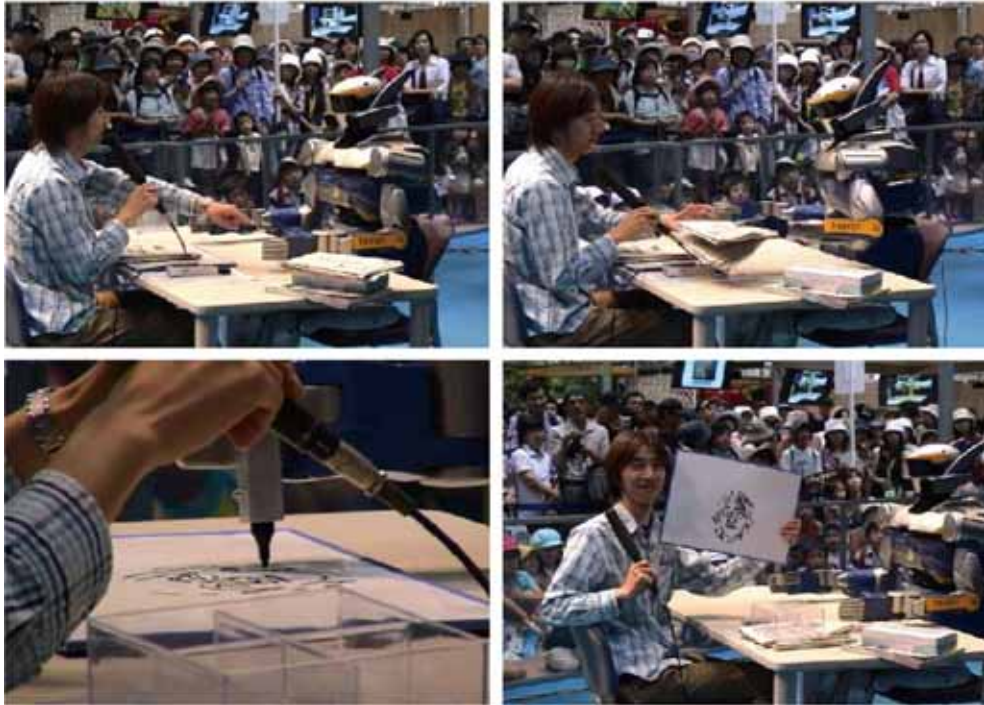


Fig. 13. Demonstration in the Aichi EXPO 2005

The future work in vision and speech involve several aspects. Since the current system doesn't fully make use of the microphone array, there is a room for improvement in this regard. For example, the realization of Blind Source Separation (BSS) using multiple microphones will enable dialogue with multiple users simultaneously. The strengthening of noise robustness and improvements of the dialogue system will be also necessary. The improvement of the number of the recognizable gestures is also an important issue for a more natural interaction.

## 8. References

- Asoh, H. Motomura, Y. Asano, F. Hara, I. Hayamizu, S. Itou, K. Kurita, T. Matsui, T. Vlassis, N. Bunschoten, R. Kroese, B. (2001). Jijo-2: An office robot that communicates and learns, *IEEE Intelligent Systems*, vol. 16, no. 5, pp. 46-55
- Honda, Asimo robot, <http://world.honda.com/ASIMO/>
- Ido, J. Myouga, Y. Matsumoto, Y. & Ogasawara, T. (2003). Interaction of receptionist ASKA using vision and speech information, in *IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 335-340

- Itou, K. Shikano, K. Kawahara, T. Takeda, K. Yamada, A. Itou, A. Utsuro, T. Kobayashi, T. Minematsu, N. Yamamoto, M. Sagayama, S. & Lee, A. (2000). Ipa Japanese dictation free software project, in *Proc. International Conference on Language Resources and Evaluation*, pp. 1343–1349
- Itou, K. Yamamoto, M. Takeda, K. Takezawa, T. Matsuoka, T. Kobayashi, T. Shikano, K. & Itahashi, S. (1999). Inas: Japanese speech corpus for large vocabulary continuous speech recognition research, *The Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206
- Jelinek, F. (1990). Self-organized language modeling for speech recognition, *Language Processing for Speech Recognition*, pp. 450–506
- Kendon, A. (1967). Some functions of gaze-direction in social interaction, *Acta Psychologica*, vol. 26, pp. 22–63
- Lee, A. Kawahara, T. & Shikano, K. (2001). Julius—an open source real-time large vocabulary recognition engine, in *Proc. ISCAEUROSPEECH2001*, pp. 1691–1694
- Lee, A. Kawahara, T. Takeda, K. & Shikano, K. (2000). A new phonetic tiedmixture model for efficient decoding, in *Proc. ICASSP2000*, pp. 1269–1272
- Lee, A. Nakamura, K. Nisimura, R. Hiroshi, S. & Shikano, K. (2004). Noise robust real world spoken dialogue system using gmm based rejection of unintended inputs, in *Proc. INTERSPEECH2004*, vol. 1, pp. 173–176
- Matsusaka, Y. Fujie, S. & Kobayashi, T. (2001). Modeling of conversational strategy for the robot participating in the group conversation, in *Proc. ISCA-EUROSPEECH2001*, pp. 2173–2176
- Matsumoto Y. & Zelinsky, A. (2000). An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement, in *Proc. of Fourth Int. Conf. on Automatic Face and Gesture Recognition*, pp. 499–505
- Nakaoka, S. Nakazawa, A. & Ikeuchi, K. (2004). An efficient method for composing whole body motions of a humanoid robot, in *In Proc. of the Tenth International Conference on VIRTUAL SYSTEMS and MULTIMEDIA*, pp. 1142–1151
- Nii, H. P. (1986). The blackboard model of problem solving and the evolution of blackboard architectures, *AI Magazine*, pp. 38–53
- Nisimura, R. Lee, A. Saruwatari, H. & Shikano, K. (2004). Public speech-oriented guidance system with adult and child discrimination capability, in *Proc. ICASSP2004*, vol. 1, pp. 433–436
- Nisimura, R. Lee, A. Yamada, M. & Shikano, K. (2005). Operating a public spoken guidance system in real environment, in *Proc. INTERSPEECH2005*
- Suzuki, K. Hikiji R. & Hashimoto, S. (2002). Development of an autonomous humanoid robot, iSHA, for harmonized human-machine environment, *Journal of Robotics and Mechatronics*, vol. 14, no. 5, pp. 324–332
- Young, S. Evermann, G. Hain, T. Kershaw, D. Moore, G. Odell, J. Ollason, D. Povey, D. Valtchev, V. & Woodland, P. (2002). The HTK book (for HTK version 3.2.1). <http://htk.eng.cam.ac.uk/>