International Sumposium on Simulation, visualization and Auralization for Acoustic Research and Education

ASVA97

2-4 April 1997 Tokyo, Japan

SOUND TECHNOLOGIES AND HUMANOID RESEARCH

Shuji Hashimoto, Yoshio Yamasaki and Katsuhiko Shirai

Humanoid Research laboratory,
Advanced Research Institute for Science and Engineering,
Waseda University

ABSTRACT

The recent results on sound technologies concerning with Humanoid Robot is described. Topics include a multimedia sensing system to track a speaking person using both image and sound processing and a dialogue system to integrate auditory and visual infomation. These are implemented in the actual robot system. The techniques for visualization and reproduction of sound field based on the multi-microphone measurement and the wave-front synthesis are also described, which is important to make robot understand the environment.

1. INTRODUTCTION

As one of the industry-university collaboration projects, the Advanced Research Center for Science and Engineering in Waseda University has been executing the Project "Humanoid" with financial supports from companies and governmental organizations. "Humanoid" is an anthropomorphic robot to work in interaction with humans within the human living environment.

This paper describes the recent results on sound technologies concerning with Humanoid Robot. First, we introduce the Humanoid Project briefly in the next section. Next, three topics are reported. The first one is a multimedia sensing system to find and track a speaking person using both image and sound processing. The second one is a dialogue analysis to realize a natural conversation between robot and human, where auditory verbal infomation and visual nonverbal infomation are integrated. The last one is a visualization and reproduction of sound field based on the multi-microphone measurement and the wave-front synthesis, which is important to make robot understand the environment.

2. HUMANOID PROJECT

Many people expect that as we move toward the 21st century, more and more robots will be developed and broadly utilized in primary and tertiary industries. With the arrival of an aged society right ahead of us, the expectation is high for personal use robots to accomplish such objectives as supports for housework within homes and supports for providing care for the elderly and handicapped peoples.

Since these robots will be operated within a residential environment built primarily for humans, it is necessary for the robots to have configurations and functions that are suited for residential environments. Moreover, the robots will be engaged in their operations in collaboration with the users. Therefore, smooth communication with humans will be indispensable. This means that the physical behavioral space and the information related mental space of a robot should have a high compatibility and permeability with those of the human. Especially, emotion is one of the most important factors that has crucial influence on the success or failure of communication in our human community. Therefore, if the robothad the same "mind" (intelligence, emotion, and will) as human, it would be far easier for robot to perfom the cooperative works with human. Moreover, robots to work with human have to under stand the environment with multi-modal sensing ability just as humando.

Thus the future robot must be a sensor-complex with high intelligence and flexible mechanical body. Various sensory infomation will be organically integrated and the robot actively works with humans using various communication channels such as voice, facial expressions, and motion of the head, eyes, arms and legs. The robot will try to understand the will of an human partner while iterating bi-directional communication. In this process, an active intelligence will be created on the robot to allow it to smoothly perfom collaborative behavior with the human.

Under these considerations, Humanoid Project in Waseda University started in 1992. The project has been planed to continue by ten years until the beginning of the next century. As the first phase of the project, basic researches on both the configurations and functions of robot are going on at present. The research topics include more than 30sub-themes in the fields of electrical engineering, information engineering and mechanical engineering. The current participating members include 10 faculty members at Waseda University most of whom were engaged in the "WABOT" anthropomorphic robot project in 1972, 4 research associates within Waseda University, 9 advisory members from outside Waseda University (including 1 Italian and 3 Americans) and about 80 graduate students at Waseda University.

3. SOUND LOCALIZATION

3.1 Composition of the audio-visual processing

We constructed the audio-visual processing system to grasp the surrounding conditions around robot from image and sound information. The main purpose of the system is to detect the human who is speaking to the robot[1].

The moving area of the image and the sound intensity is measured every 1/10 second. Then by examining the synchronization of between sound and image motion, the system detects a speaking person in the scene. As a system to be installed in a robot, it must have a real-time processing ability. Therefore, the system was developed using special hardware with simplified soft ware algorithm. Figure 1 shows the composition of the audio-visual

system. A monochrome CCD camera is equipped on the robot head. The camera image is converted into mosaic image by the image processing hardware. The sound signals from two condenser microphones installed on right and left sides of robot head are pre-processed for localization the sound source by the sound processing hardware. PC/AT compatible PC, processes the individual data from the hardware processors in an integrated manner and sends instruction to the robot head controller.

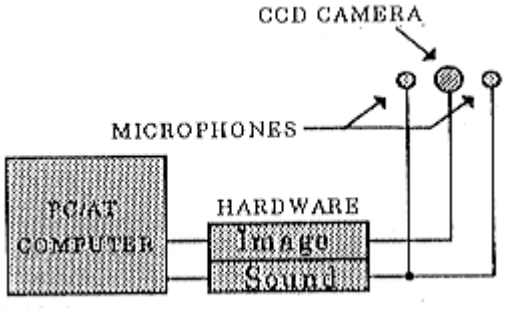

Fig.1 Audio-visual localization system

3.2 Visual infomation processing

After AD conversion of the image signals with 256 × 256resolution, the image processing hardware converts the image into four reduced images up to 16 × 16resolution. By using the mosaic images at with different resolutions, it is possible to increase the efficiency of analyzing objects in the scene by taking advantage of the pyramid structure of the images. The mosaic images of the rate of 60 frames per second are averaged for every six frames to obtain the smoothed images of 10 frames per second. Then the frame differences are calculated to detect the moving area. The upper part of the gravity center of the moving area is regarded as the candidate position of the target moving body(the face of a human).

3.3 Audio infomation processing

Two sound data from the left and right microphones are digitized at 12.7kHz sampling rate after filtering with LPF and HPF. Then the cross correlation functions are calculated for every 1/60 second that is the same interval as the image capturing. Next, correlation functions are averaged for six frames to obtain the average gravity center around the peak which is the phase differences of the two sound. The direction of the sound source is obtained by using knowledge about the distance between two microphones and the sound velocity. As the calculation of the correlation function is performed by the hardware, both the robust operation under the noisy environment and real time processing are simultaneously realized.

3.4 Integration of audio-Visual infomation

Infomation about the image and audio data are obtained every 1
situations are classified into the following two cases for the image data.

a) image case 1:Moving area exists in the scene.
b) image case 2:Moving area does not exist.

As for the audio data, the situations are classified into three cases;

a) sound case 1:Sound is not detected.
b) sound case 2:Sound is detected outside of the range of camera view

c) sound case 3:Sound is detected in the direction within the range of camera view.
The six combinations of these cases provide the frame types of the scene as shown in Figure2.

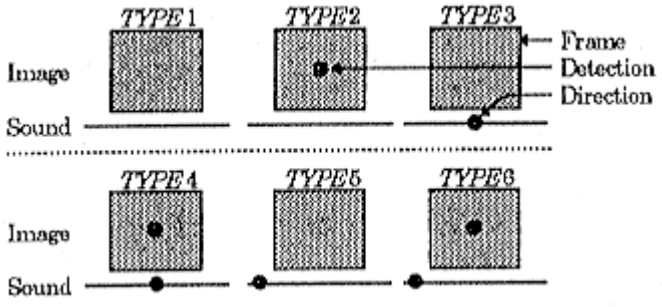

Fig.2 Frame types of the scene

In the experiments, the robot was controlled based on the frame type and synchronization between sound and image motion to show a human-like behavior[2]. The typical action patterns of the robot are summarized as follows:

(1) If the robot has detected the synchronization, it tracks the moving area in the scene.
(2) Even if the robot detects the sound in other direction, it does not turn to that location. However, if the sound is loud enough, it does.
(3) If the robot detected the sound more than five second later after the detection of synchronization, it turns to the sound location.
(4) After the robot has turned to the sound location, if it detects the moving area but no sound, it tracks the moving area.
(5) If the robot cannot detect sounds for three second, it turns to the previous location again.

4. ANALYSIS OF CONVERSATION

4.1 Analysis of spoken dialogue

Dialogue is an interactive communication of information mainly based on speech. We can have natural conversation with out actually seeing each other for example considering conversation by telephone. However, in practical conversation, visual information such as gesture, facial expression, and head movement clearly makes the progress of conversation much smoother and more natural. Therefore, in the more natural human interface that can use multiple modalities, visual infomation becomes as important as auditory infomation.

We are not only researching elementary technologies to realize human interface such as speech recognition, speech synthesis, but also trying to clarify the mechanism of dialogue management methods from both auditory and visual aspects. In particular, we are dealing with the following topics (Figure3).

(1) Modeling of dialogue[3].
Constructing a transition model of utterances.
(2) Rhythm of dialogue[3].
Analyzing pause and speed of dialogues.
(3) Nonverbal infomation in dialogue[4].
Analyzing nonverbal information such as gaze and gesture to display an effective feedback from the system.

(4) Head movements in dialogue[5].
In the following section, we will focus on the visual aspects of spoken dialogue.

4.2 Analysis of head movements

Unlike spoken language, nonverbal behavior has no standard rule, and social and individual differences are significant. Among these human movements in a conversation, the head movement is rather clear and can be easily classified. Therefore head movement is one of the visual information that we can obtain a statistical result by dealing with a certain amount of data, even if there are individual differences. To study the role of head movements in spoken dialogue, we analyzed conversations under two conditions. One is face to face conversation with a visualcontact, and the other is conversation through telephone line with no visual contact.

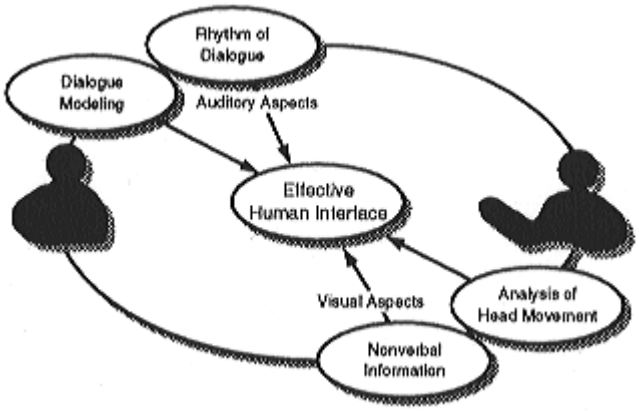

Fig.3 Intefrated analysis of spoken dialigue

In our daily conversation, head movements such as nodding occurs even under the situation that we cannot see each other. Therefore, we can say head movements are not necessarily used to communicate information to one'spartner. On the other hand, there exist head movements that are more communicative actions intended to give some sort of information to his partner even without utterance. This implies the importance of visual information that could potentially affect the dialogue without voice infomation.

Considering head movement not only as a signal but also an interactive action, we can recognize the effect of visual information to the dialogue. In constructing an effective man machine interface, in conjunction with speech information, the use of visual information will become important. And the head movement in conversation is a useful element to study the role of cach utterance and the dynamic structure of the dialogue.

4.3 Nonverbal information

We also considered gaze information, especially when turn taking occurs, as an important visual informationin our conversation. The application of nonverbal information such as gaze is essential to realize a robot which can communicate with humans, as humans can. From recent analysis of human conversation, it was found that gaze information plays an important role in the progress of conversation, especially in turn taking. We constructed a spoken dialogue system simulating the actions of a humanoid robot, and applied this gaze information to it. We made experiments with this system to evaluate the effectiveness of gaze information. The results of this

experiment show that gaze information was indeed effective, especially when the user interrupted the system's action.

5. VISUALIZATION AND REPRODUCTION OF SOUND FIELD

5.1 Four Microphone Measurement

The four microphone measurementis a method to grasp spatial information of sound fields from impulse responses measured at closely located four points. These four points must not place on the same plane, usually we set them on the origin and the three points on the rectangular coordinate axess 5cm distance from the origin [6].

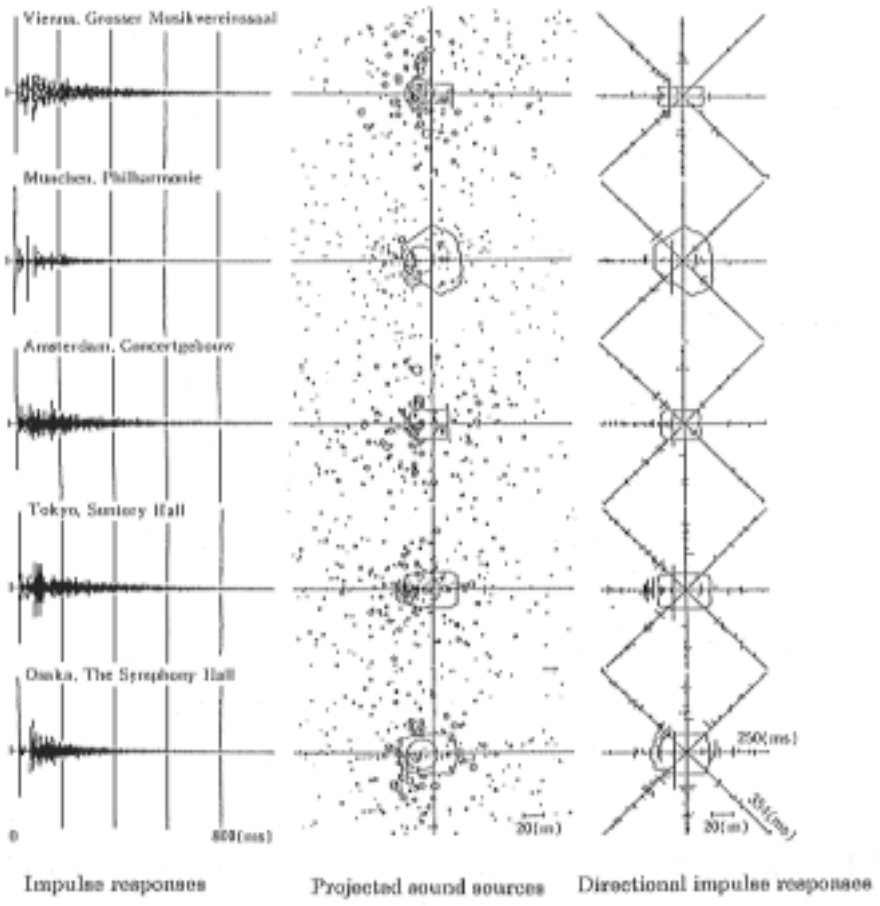

Fig.4 Results of the four microphone measurement

Three dimensional coordinates of direct and reflected sound sources are calculated by correlation technique or intensity technique. Figure 4 shows the results of the four microphone measurements in various concert halls; impulse responses, sound sources distribution projected to floor plane and the directional impulse responses obtained from it. In sound sources distribution, the center of each circle is the projected coordinate of direct or reflected sound source, the area of each circle represents the power of the corresponding source, and the cross point of the two orthogonal lines is the observation point, the outlines of the concert hall are also shown.

By convolution of the signal and each directional impulse response, the sound field reproduction is realized.

5.2 Wave-front synthesis

Based on the Kirchhoff-Helmholtz integral, an arbitrary sound field within an enclosed sace can be determined by the normal component of particle velocities and the sound pressures on the surface of the space. The only restriction is that there are no sound sources within this space.

However, realization of this theory numerous number of loudspeakers are needed, for example up to 17kHz loudspeakers should be placed at 1cm intervals. Therefore we are obliged to reduce the number of loudspcakers. Figure 5 shows the wave front synthesis function W proposed by A.J.Berkhout, D.de Vries and P.Vogel [7][8] which reduce discrete points on surface $z = 0$ to $l (n > l)$ on surface $z = l$ is introduced.

An experiment was done in an echoic chamber with 261 ludspeakers. Figure 6(a) shows the distribution of the sound sources measured in Aichi Art Concert Hall, Fig.6(b) shows the distribution of the sound sources in the reconstructed sound field.

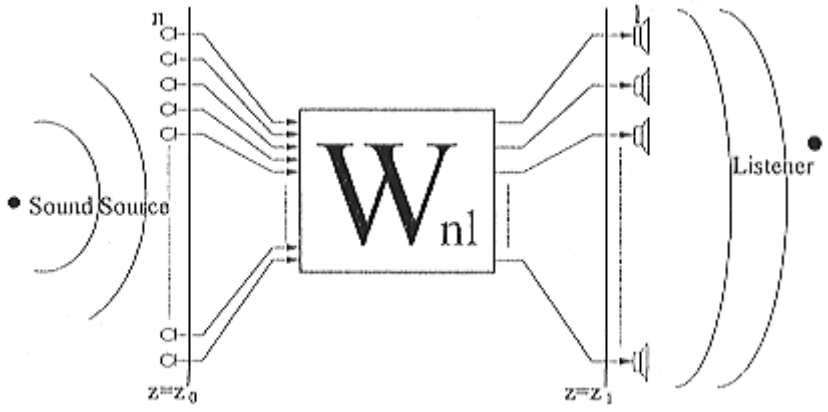

Fig.5 Wave-front synthesis

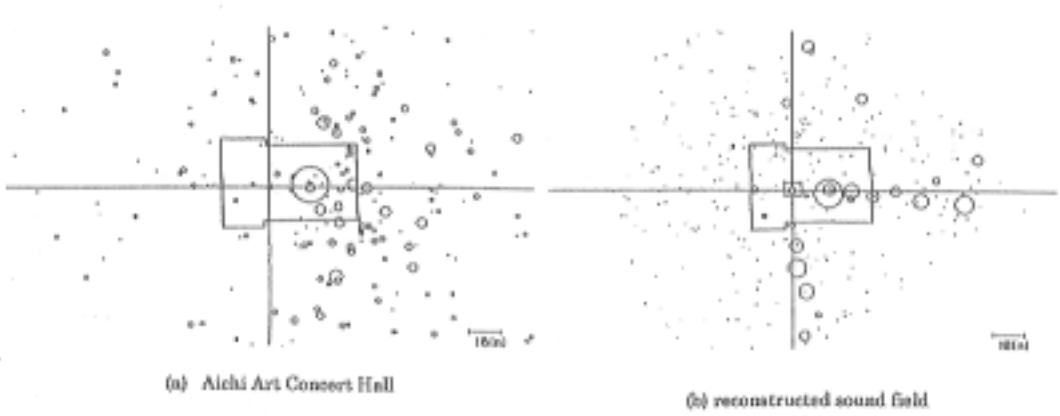

Fig.6 Wave-front synthesis

6. CONCLUSIONS

The visual information is recognized well to be very important in robot system and a lot of researches are done on robot vision. On the other hand, not so many studies on auditory information processing for robot are reported, although it is important as well. Sound plays an important role not only for the conversation but also for the environment understanding in humandaily life.

In this paper we introduced three topics on sound technologies studied in Humanoid projeet. The first and the second research havc been implemented in the experimental robot Hadaly which perform the task of the campus information guide with speech conversation and gesture. Our next target is to construct a totaled auditory system which works with vision system for Humanoid robot.

REFERENCE

[1] T.Kurata, D.Chang and S.Hashimoto: "Multimcdia Sensing System for Robot", Proc. Ro-man 95, 83-88, 1994.
[2] S.Hashimoto et al: "Humanoid-Development of an Information Assistant Robot Hadaly-", Proc.Robotics Symposia, 17-20, 1996. (in Japanese)
[3] E.Morikawa, M.Yokoyama, Y.Sugita and K.Shirai: "Spoken Dialogue Modeling Based on Statistical Approach", IPSJSIG-SLP 15-23, Feb. 1997. (in Japanese)
[4] K.Hoashi, M.Yokoyama, D.Arai, Y.Ando and K.Shirai: "Nonverbal Information of a Humanoid Robot", IPSJ SIG-SLP 15-11, Feb 1997. (in Japanese)
[5] Y.Iwano, Y.Sugita, M.Matsunaga and K.Shirai: "Difference in Face-to-face and Telcphone Dialogues: Analysis of the Role of Head Movements", IPSJ SIG-SLP 15-19, Feb 1997. (in Japanese)
[6] Y.Yamasaki and T.Itow, "Measurement of spatial information in sound field by closely located four point microphone method," J.Acoust.Soc, Jpn(E), 10, 101-110, 1989.
[7] A.J.Berkhout, D.de Vries and P.Vogel," Acoustic Control By Wave Field Synthesis," J.Acoust.Soc.Am.. 1993.
[8] A.J.Berkhout, Marinus M.Booncand Diemer de Vries" Generation of Sound Fields Using Wave Fields Synthesis, An Overview, ” Active95, 1995.