UDC 621.397.3:801.41

Speaker Position Detection System Using Audio-visual Information

●Naoshi Matsuo ●Hiroki Kitagawa ●Shigemi Nagata

(Manuscript received June 18, 1999)

This paper describes a speaker position detection system that achieves a high degree of accuracy using a multimodal interface that integrates audio and visual information from a microphone array and a camera. First, the system detects the position and angle of the microphone array relative to the camera position. Next, audio processing detects sound source positions and visual processing detects the positions of human faces. Finally, by integrating the sound source positions and face positions, the system determines the speaker's position. The system can integrate audio and visual information, even if the spatial relationship between the microphone array and camera is initially unknown. The system achieves a high detection rate for the speaker's position in a noisy environment.

1. Introduction

Multimodal interfaces use a variety of technologies such as speech processing and image processing to create a natural and friendly human interface and have recently been the subject of much study.[1),2)] The various modalities present in a multimodal interface, which correspond roughly to the five senses, and the information of the modalities need to be input, processed, and output concurrently or sequentially as circumstances require. Multimodality is an essential requirement for a true human interface which allows a user to interact with a personal computer or other equipment using such means as speech, gestures, and eye movements.

A sensing system that integrates audio and visual information is a kind of multimodal interface.[3)] We have previously studied a speaker position detection system that uses audio and visual information integration.[4)] The system integrates audio information from a microphone array and visual information from a camera to detect a speaker's position with a high degree of accuracy in a noisy environment. The system can, for example, be combined with a speech enhancement system[5)] to make a hand-free telephone system that can be used in a noisy environment.

However, the system cannot integrate audio and visual information if the relationship between the microphone array position and the camera position is unknown. Therefore, the microphone array and the camera must be set at predetermined locations or their positions must be input to the system.

This paper presents a new method that can overcome this problem. First, this method calculates the position and angle of the microphone array relative to the camera position and then integrates the results of the audio processing and visual processing according to the calculation results. In this way, the system can detect a speaker's position, even if the spatial relationship between the microphone array and camera is initially unknown.

2. Speaker position detection system

2.1 Outline

Figure 1 shows a prototype system that uses the proposed method, and **Figure 2** shows its block diagram. The system is connected to a microphone array, camera, and stereo loudspeakers. The microphone array consists of three nondirectional microphones, and the loudspeakers are used to detect the spatial relationship between the microphone array and the camera. The processing is performed by two personal computers that are connected with a network. One of the computers is used for processing the audio information and the other is used for processing the visual information and integrating it with the audio information.

Figure 3 shows the common coordinate system for the visual processing and integration, and **Figure 4** shows the coordinate system for the audio processing. The system detects sound source positions using audio processing and detects the positions of human faces using visual processing. The speaker's position $(x, y, z)_{SP}$ is detected by integrating the detected sound source positions and face positions in the coordinate system for integration. In Figure 3, the camera and loudspeakers are fixed in the coordinate system for integration processing. However, the microphone array is

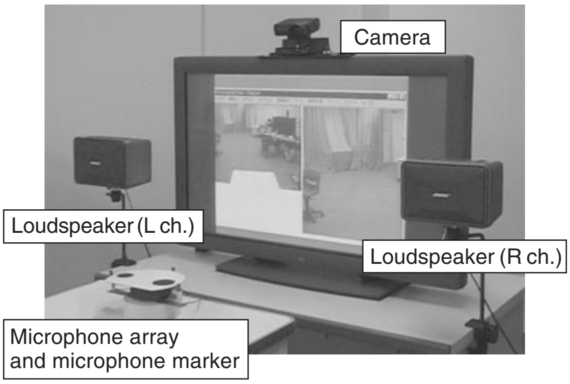

Figure 1
Prototype speaker position detection system.

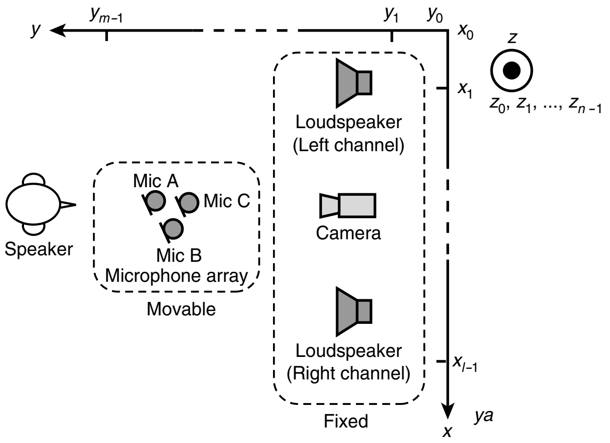

Figure 3
Coordinate system for visual processing and integration processing.

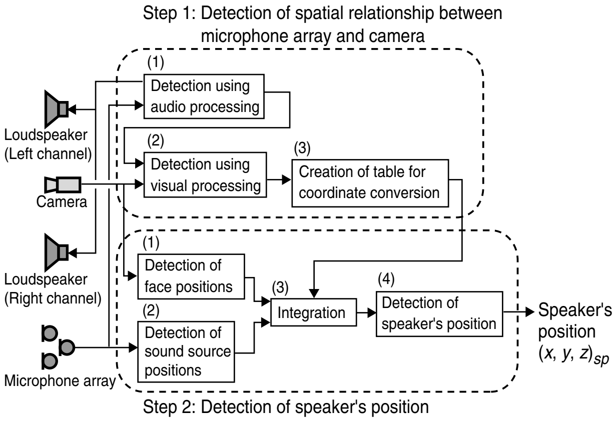

Figure 2
Block diagram of speaker position detection system.

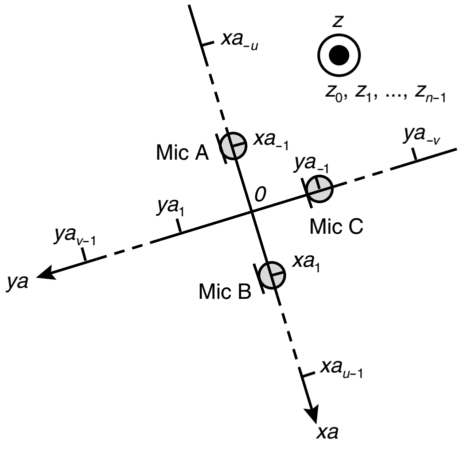

Figure 4
Coordinate system for audio processing.

movable, so its position and angle relative to the camera are initially unknown. Therefore, the system must detect the microphone array's position and angle before it can integrate the sound source positions and face positions.

Step 1 in Figure 2 shows the process for detecting the spatial relationship between the microphone array and the camera. First, based on the camera position, the microphone array position is roughly estimated using audio processing which can search 360 degrees (up/down, left/right) around the microphone array. Next, the system accurately detects the position and angle of the microphone array in the coordinate system for integration using visual processing. Finally, the system makes a coordinate conversion table from the relationship between the microphone array position and camera position. This table converts the coordinates for audio processing into the coordinates for integration. The coordinate conversion table is used in Step 2. Step 1 reduces the processing time and the detection error rate.

In Step 2, the system detects the sound source positions and face positions. The speaker's position is detected by integrating the results of the sound source position detection and the face position detection using the coordinate conversion table.

By using this method, therefore, the system can detect a speaker's position, even if the spatial relationship between the microphone array and the camera is initially unknown.

2.2 Step 1: Detection of spatial relationship between microphone array and camera

The spatial relationship between the microphone array and the camera is detected in Step 1 as follows:

1) Detection using audio processing

Figure 5 shows the audio processing for detecting the microphone array/camera relationship. First, the loudspeaker positions in the coordinate system for audio processing (Figure 4) are detected using the microphone array. Next, based on the camera position in the coordinate system for integration (Figure 3), which is fixed with respect to the loudspeaker positions, the microphone array position is roughly estimated using the detected loudspeaker positions.

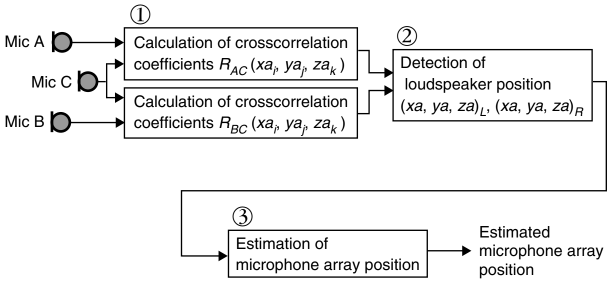
①
Mic A
Mic C
Mic B
Calculation of crosscorrelation coefficients $R_{AC}(xa_i, ya_j, za_k)$
Calculation of crosscorrelation coefficients $R_{BC}(xa_i, ya_j, za_k)$
②
Detection of loudspeaker position $(xa, ya, za)_L$, $(xa, ya, za)_R$
③
Estimation of microphone array position
Estimated microphone array position

Figure 5
Audio processing for detecting microphone array/camera relationship.

The various stages in the detection processing shown in Figure 5 are explained below.

①

- A signal is output to the left-channel loudspeaker.
- The crosscorrelation coefficients $R_{AC}(xa_i, ya_j, za_k)$ of the Mic A and Mic C input signals and $R_{BC}(xa_i, ya_j, za_k)$ of the Mic B and Mic C input signals in the coordinate system for audio processing are calculated.

②

- The position of the left-channel loudspeaker $(xa, ya, za)_L$ is calculated from the crosscorrelation coefficients $R_{AC}(xa_i, ya_j, za_k)$ and $R_{BC}(xa_i, ya_j, za_k)$ using triangulation.
  The position of the right-channel loudspeaker $(xa, ya, za)_R$ is calculated using a similar procedure.

③

- The position of the camera in the coordinate system for audio processing is calculated from loudspeaker positions $(xa, ya, za)_L$ and $(xa, ya, za)_R$.
- From the calculated camera position, the position of the microphone array in the coordinate system for integration is calculated.

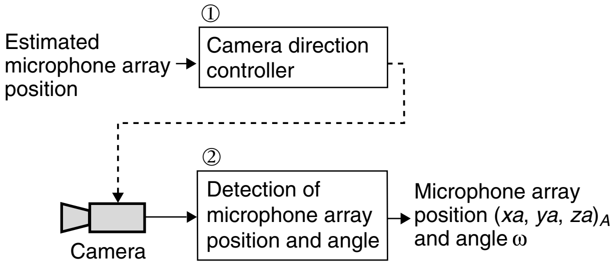
①
Estimated microphone array position
Camera direction controller
②
Camera
Detection of microphone array position and angle
Microphone array position (xa, ya, za)A and angle ω

Figure 6
Visual processing for detecting microphone array/camera relationship.

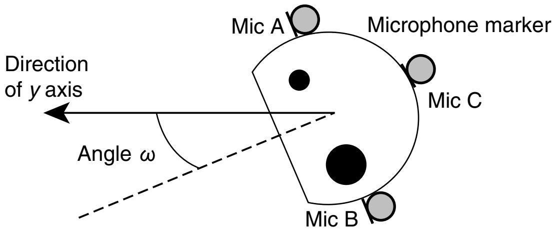

Figure 8
Angle of microphone array.

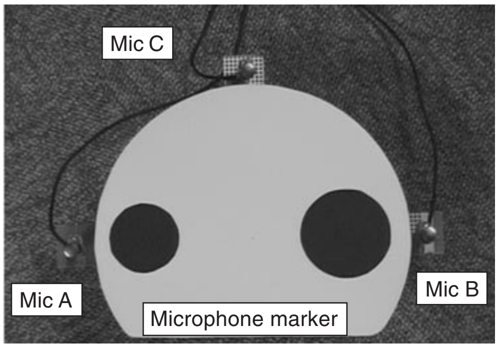

Figure 7
Microphone array and marker.

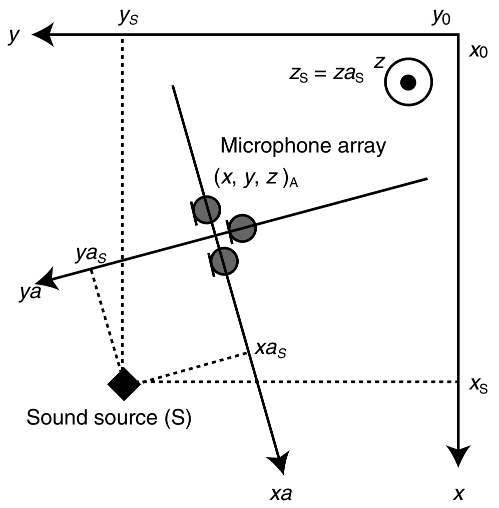

Figure 9
Coordinate conversion.

2) Detection using visual processing

Figure 6 shows the visual processing for detecting the microphone array/camera relationship. First, the camera is turned to the estimated position of the microphone array. Next, the position and angle of the microphone array are accurately calculated by the visual processing. The visual processing uses the microphone marker shown in **Figure 7**. This marker has two black circles of different diameters, and the position and angle of the marker are detected using the spatial relationship between these black circles (**Figure 8**).

The detection processing shown in Figure 6 is explained below.

①

– The camera is turned to the estimated position of the microphone array.

②

– The position $(x, y, z)_A$ and angle ωωof the microphone array in the coordinate system for integration are calculated by the visual processing using the microphone marker.

3) Creation of table for coordinate conversion

A sound source position $(xa, ya, za)_S$ in the coordinate system for audio processing corresponds to a position $(x, y, z)_S$ in the coordinate system for integration as shown in **Figure 9**. The coordinates are converted according to Equation (1) using a parallel displacement and a rotation around the z-axis based on the position and angle of the microphone array that were calculated in (2) above. The system creates a coordinate conversion table for the speaker position detection described in Section 2.3 below. **Table 1** shows an example for microphone array position $(x, y, z)_A$ = (20, 10, 5) and angle ωω= 0º.

Table 1
Example coordinate conversion table.

Coordinates for audio processing	(−10, −10, 0)	–	(0, 0, 0)	–	(9, 9, 4)
Coordinates for integration	(10, 0, 5)	–	(20, 10, 5)	–	(29, 19, 9)

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}_S = \begin{bmatrix} \cos(-\omega) & -\sin(-\omega) & 0 \\ \sin(-\omega) & \cos(-\omega) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} xa \\ ya \\ za \end{bmatrix}_S + \begin{bmatrix} x \\ y \\ z \end{bmatrix}_A \qquad (1)$$

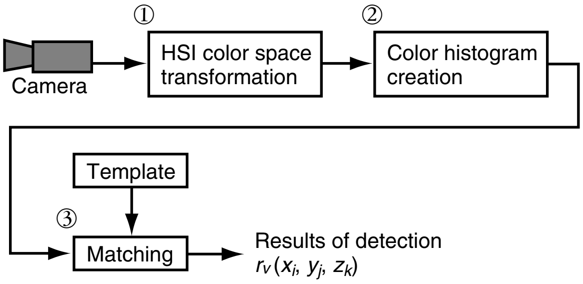

Figure 10
Visual processing for detecting face positions.

2.3 Step 2: Detection of speaker's position

The speaker's position is detected in Step 2 in Figure 2 as follows.

1) Detection of face positions

Figure 10 shows the visual processing for detecting the positions of human faces in the visual input. The face positions are detected by histogram matching[6)] between a template made in advance from an image of a human face and a color histogram obtained from an HSI color space transformation of the visual input. (In the detection experiments described in this paper, the template was made from an image of a person who was not a subject of the experiments.) The highest matching values $r_v(x_i, y_j, z_k)$ give the positions of faces in the visual input.

The detection processing shown in Figure 10 is explained below.

①

– The RGB data of the input image is transformed to HSI (Hue, Saturation, Intensity) data.

②

– A color histogram is created from the HSI data.

③

– The matching values $r_v(x_i, y_j, z_k)$ of the histogram and the template are calculated.

2) Detection of sound source positions

The processing for detecting the sound source positions in the coordinate system for audio processing is similar to the processing for detecting the loudspeaker positions shown in Figure 5. The products $r_a(xa_i, ya_j, za_k)$ of the crosscorrelation coefficients $R^S{}_{AC}(xa_i, ya_j, za_k)$ and $R^S{}_{BC}(xa_i, ya_j, za_k)$ are calculated. $R^S{}_{AC}(xa_i, ya_j, za_k)$ and $R^S{}_{BC}(xa_i, ya_j, za_k)$ are calculated in the same way as $R_{AC}(xa_i, ya_j, za_k)$ and $R_{BC}(xa_i, ya_j, za_k)$ in Figure 5. The positions of the highest $r_a(xa_i, ya_j, za_k)$ values are taken as the positions of sound sources.

3) Integration

In the integration stage, the $r_a(xa_i, ya_j, za_k)$ values which give the sound source positions in the coordinate system for audio processing are converted to $r'_a(x_i, y_j, z_k)$ values in the coordinate system for integration. Next, the products $r(x_i, y_j, z_k)$ of $r'_a(x_i, y_j, z_k)$ and $r_v(x_i, y_j, z_k)$ are calculated to obtain the speaker's position.

4) Detection of speaker's position

The position of the highest $r(x_i, y_j, z_k)$ is taken as the speaker's position, $(x, y, z)_{SP}$.

3. Experiments

We measured the accuracy of the detected microphone array position and angle and the detected speaker's position to evaluate the proposed method.

The coordinates for the integration were set as follows:

• x axis: $x_0, \dots, x_{39}$	(Interval: 10 cm)
• y axis: $y_0, \dots, y_{39}$	(Interval: 10 cm)
• z axis: $z_0, \dots, z_9$	(Interval: 20 cm)

- Camera position: (x_{20}, y_0, z_7)

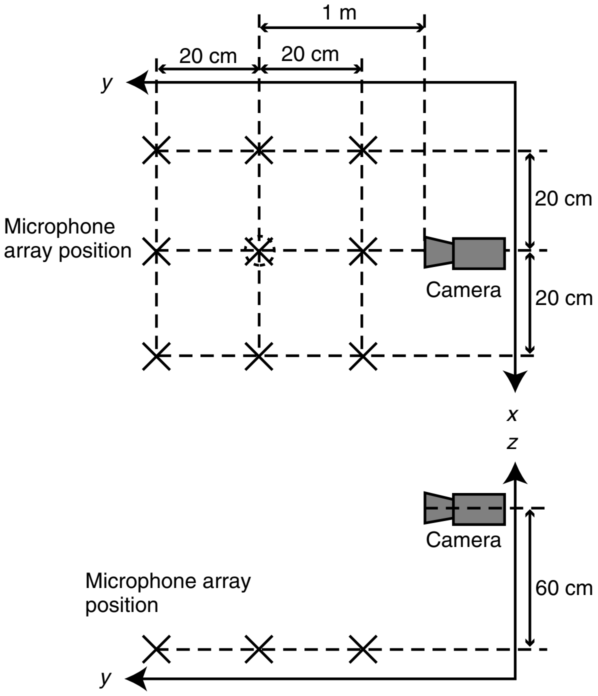

Figure 11
Microphone array positions in experiments.

- Left-channel loudspeaker position: (x_{13}, y_3, z_5)
- Right-channel loudspeaker position: (x_{27}, y_3, z_5)

The coordinates for the audio processing were set as follows:

- *xa* axis: $xa_{-10}, \dots, xa_9$ (Interval: 10 cm)
- *ya* axis: $\mathrm{ya}_{-10}, \dots, ya_9$ (Interval: 10 cm)
- *za* axis: $za_0, \dots, za_4$ (Interval: 20 cm)
- Microphone array position: (xa_0, ya_0, za_0)

The sampling frequency was set to 8 kHz.

3.1 Experimental detection of microphone array position and angle

We conducted the following experiments for measuring the accuracy of microphone array position and angle detection.

3.1.1 Method

We tested the system as described in (a) and (b) below at each of the nine positions shown in **Figure 11**, ten times each at microphone array angles 0°, ±45°, and ±90° (total of $9 \times 10 \times 5 = 450$ tests).

Table 2
Correct estimates for microphone array position.

Angle (°)	−90	−45	0	45	90
Correct estimates (%)	100	82	76	76	84

(a) Detection using audio processing

First, we measured the accuracy of estimations of microphone array position in the coordinate system for integration using audio processing. Each estimation was judged to be a success if the camera could see the microphone array after it was turned to the estimated position.

(b) Detection using visual processing

Then, we measured the accuracy of microphone array position and angle detection using visual processing with the microphone marker placed near the position estimated in (a). The errors in position detection were recorded as distance and direction errors relative to the camera position. Then, to help us evaluate the system's ability to detect the position and angle, we repeated this test using visual information only and compared the processing times.

3.1.2 Result

(a) Detection using audio processing

Table 2 shows the measurement results for microphone array position estimation using audio processing at the five different angles. The average rate of successful detection is about 84%. The variation in the rate over the five angles is due to the particular arrangement of microphones we used (see Figure 4). We will study this matter and look for an arrangement that reduces this variation.

(b) Detection using visual processing

The maximum detection errors for all the positions shown in Figure 11 are shown below. These results show that the errors are small and did not affect the experiment for speaker position detection described in the next section.

– Position errors
Distance : ±10 cm
Direction : ±1°
– Microphone array angle error : ±3°

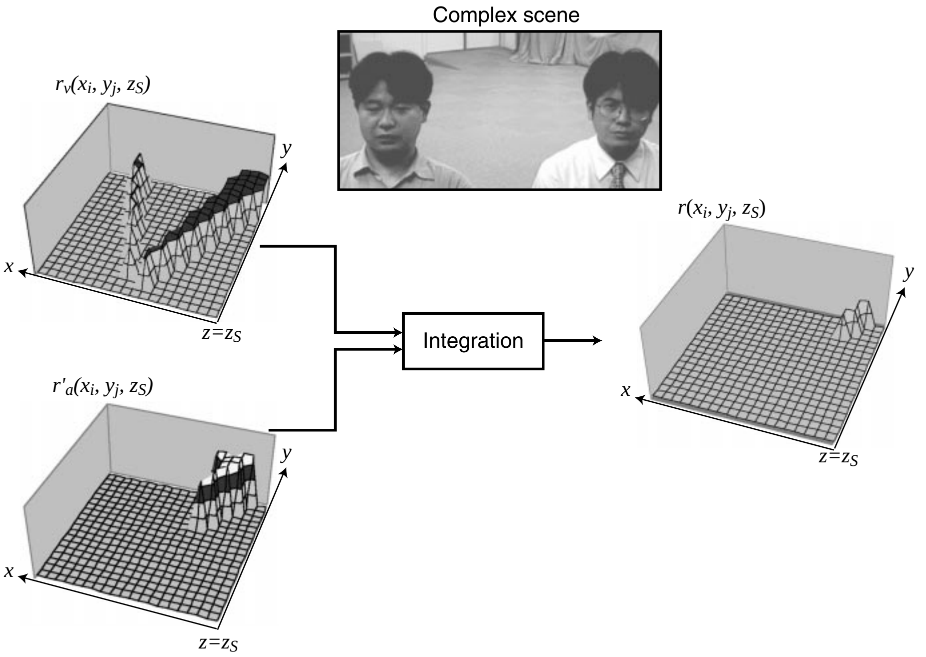

Figure 12
Example of integration.

When only visual information is used, the detection accuracies for distance, direction, and angle are similar to the results of the proposed method, but the processing time is about eight times longer than the time taken by the proposed method.

3.2 Experiment for speaker position detection

We conducted experiments for measuring the detection rates for a speaker's position in noisy environments.

3.2.1 Method

We measured the detection rates for a speaker's position when the microphone array was at the circled position shown in Figure 11 and its angle was 0º. First, the position and angle of the microphone array in the coordinate system for integration were detected. Next, a coordinate conversion table similar to Table 1 was made. Finally, the speaker's position was detected by integrating the audio and visual information. To further evaluate the proposed method, we also detected the speaker's position with ordinary methods that use audio or visual information only.

In the experiments, we used three kinds of noise:

– Audio noise

We placed a loudspeaker that was outputting human speech beside the speaker. The level of sound output was about 64 dB(A).

– Visual noise

We changed the illumination conditions by switching fluorescent and incandescent lamps on and off and visually complicated the scene by putting up posters. We also measured the detection rates without noise for reference.

In the experiments, we set the admissible range of errors to ±30 cm based on human body size and the distances shown in Figure 11.

3.2.2 Results

Figure 12 shows an example of integration on the *xy* plane ($z = z_S$) for a complex scene. The person on the right is speaking and the person on the left is silent. The example shows that the visual influence of the person on the left is eliminated by integrating the audio information $r'_a(x_i, y_j, z_k)$ and the visual information $r_v(x_i, y_j, z_k)$.

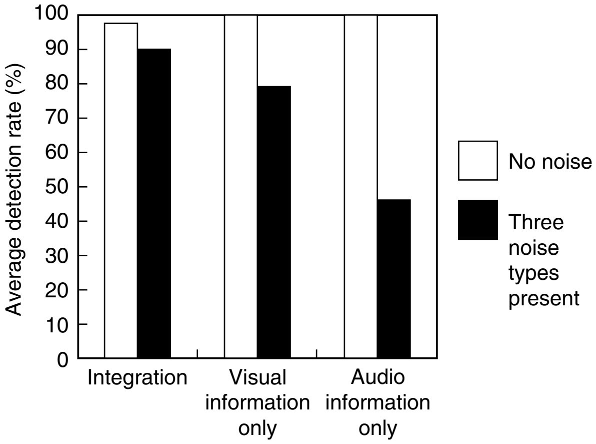

Figure 13
Results of speaker position detection experiments.

Figure 13 shows the average detection rates for the speaker's position when the three types of noise described in Subsection 3.2.1 were present. When there was no noise, the detection rates of the three methods were roughly equal. However, when the three types of noise were present, the detection rate of the proposed method was higher than the rates of the ordinary methods.

These results prove that the speaker's position can be detected with a high degree of accuracy in a noisy environment by integrating the audio information and visual information.

4. Conclusion

We proposed a method of speaker position detection that integrates audio and visual information from a microphone array and a camera. The speaker's position is detected by integrating the audio and visual information based on the detected spatial relationship between the camera and microphone array. Experiments have shown that the system can integrate audio and visual information even if the spatial relationship between the microphone array and the camera is initially unknown and that it can detect the speaker's position with a high degree of accuracy in a noisy environment.

We are planning to study a method for improving the detection accuracy for the microphone array/camera spatial relationship, and a method for updating the relationship without interrupting an application by using the speaker position detection system. We will use this system in various personal-computer applications.

References

1) K. Maruyama and T. Sasaki: Intersensory effects across vision and audition. (in Japanese), *J. ASJ*., **52,** 1, pp.34-39 (1996).

2) N. Hataoka and H. Kikuchi: Topics on Multimodal Interfaces Which Use Speech Technologies. (in Japanese), *J.IEICE*., **80**, 10, pp.1031-1035 (1997).

3) K. Takahashi: Sensing System Integrating Audio and Visual Information. (in Japanese), *J. IEICE*., **79**, 2, pp.155-161 (1996).

4) N. Matsuo, H. Kitagawa, and S. Nagata: A Speaker Position Detection System using Audio-Visual Information. (in Japanese), Proc. of the 13th Human Interface Symposium, pp.469-474 (1997).

5) N. Matsuo and S. Nagata: Study on Directional Microphone Technology using Estimated Signal. Proc. of ITC-CSCC'97, pp.425-428 (1997).

6) M. J. Swain and D. H. Ballard: Indexing via color histograms. Proc. of Image Understanding Workshop, pp.623-630 (1990).

Naoshi Matsuo received the B.E. and M.E. degrees in Electric Engineering from Kyushu Institute of Technology, Fukuoka, Japan in 1987 and 1989, respectively. He joined Fujitsu Laboratories Ltd., Kawasaki, Japan in 1989 and has been engaged in research and development of speech and audio signal processing and multimodal interfaces. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the Information Processing Society of Japan (IPSJ), and the Acoustic Society of Japan (ASJ).

Hiroki Kitagawa received the B.E. degree in Mechanical Engineering from Osaka University, Osaka, Japan in 1992. He joined Fujitsu Laboratories Ltd., Kawasaki, Japan in 1992 and has been engaged in research and development of image processing and multimodal interfaces. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and the Human Interface Society (HIS).

Shigemi Nagata received the B.E. degree in Mechanical Engineering Science and the M.E. degree in Systems Science from Tokyo Institute of Technology, Tokyo, Japan in 1975 and 1977, respectively. He joined Fujitsu Laboratories Ltd., Kawasaki, Japan in 1977 and has been engaged in research and development of pattern recognition, image processing, computer vision, neurocomputing, sensory data fusion, and multimodal interfaces. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, Information Processing Society of Japan (IPSJ), Japanese Society for Artificial Intelligence (JSAI), Japanese Neural Network Society (JNNS), Robotics Society of Japan (RSJ), and Japan Society of Mechanical Engineers (JSME). He received the best paper award from JSAI in 1998.