Proceedings of 2004 IEEE/RSJ International Conference on
Intelligent Robots and Systems
September 28 - October 2, 2004, Sendai, Japan

Robust Speech Interface Based on Audio and Video Information Fusion for Humanoid HRP-2

Isao Hara, Futoshi Asano,
Hideki Asoh, Jun Ogata, Naoyuki Ichimura
Information Technology Res. Inst.
AIST
Tsukuba, Japan
Email: Isao-Hara@aist.go.jp

Yoshihiro Kawai, Fumio Kanehiro,
Hirohisa Hirukawa
Intelligent Systems Inst.
AIST
Tsukuba, Japan

Kiyoshi Yamamoto
Tsukuba Univ.
Tsukuba, Japan

Abstract— **For cooperative work of robots and humans in the real world, a communicative function based on speech is indispensable for robots. To realize such a function in a noisy real environment, it is essential that robots be able to extract target speech spoken by humans from a mixture of sounds by their own resources. We have developed a method of detecting and extracting speech events based on the fusion of audio and video information. In this method, audio information (sound localization using a microphone array) and video information (human tracking using a camera) are fused by a Bayesian network to enable detection of speech events. The information of detected speech events is then utilized in sound separation using adaptive beamforming. In this paper, some basic investigations for applying the above system to the humanoid robot HRP-2 are reported. Input devices, namely a microphone array and a camera, were mounted on the head of HRP-2, and acoustic characteristics for sound localization/separation performance were investigated. Also, the human tracking system was improved so that it can be used in a dynamic situation. Finally, overall performance of the system was tested via off-line experiments.**

I. Introduction

For communication between robots and humans using speech in a everyday situation with environmental noise, extraction of speech spoken by humans (target speech) from a mixture of sounds is essential. The authors have previously proposed a method of detecting and separating target speech events using audio and video information fusion [2], [3], [4]. From audio information obtained by a microphone array, the time and location of an audio event (emission of sound from sound sources) can be known. From video information, the time and location of a video event (existence of a human) can be known. By combining such audio and video information, co-occurrence of the audio and video event in a certain region can be detected. This co-occurrence is defined as a speech event in this study. By using both audio and video information, only speech signals emitted by humans are detected and noise, including speech signals such as sound from a TV, can be avoided.

A method of detecting and separating speech events based on audio and video information has also been proposed by Nakadai *et al.* [5]. The scope of thier research is considered to realize a more human-like robot audition system using only two microphones at the ear position of robots and the knowledge of human speech such as harmonic structures. In the method proposed by the authors [2], [3], [4], a microphone array with more microphones and a more general sound localization/separation approach were employed. By using these at the cost of larger computational load and system size, the limitations on source signal, system configuration and environments are considered to be relaxed to some extent.

In this paper, to facilitate introduction of the above framework to the humanoid robot HRP-2 as shown in Fig. 1 [1], some basic investigations are reported. We first evaluated the performance of sound localization used in the above approach with a microphone array mounted on the head of a humanoid. Especially, the effects of the complicated shape of the head and the size limitation on the sound localization performance were examined. Secondly, the background subtraction method of human tracking was replaced by a model-based approach. The background subtraction used in [2], [3] and [4] is a simple and effective method for human detection. However, this method cannot be used for robot application since the background often changes as the robot moves. Thus, a model-based approach is newly introduced into the information fusion framework. In this method, a human in a scene is first detected by using skin-color detection and template matching of face. Once a human is found, a model of that human is made and is tracked by the kernel-based tracking method [6]. In the present study, the performance of the newly introduced human tracking system was tested in an information fusion framework. Finally, an experiment on the detection and separation of speech events was conducted in an ordinary environment with interference by music.

II. Sound Localization

A. Method

The purpose of sound localization is to estimate the location of sound sources in the environment. For sound localization, the MUSIC method [7] extended to the broad-band signal with eigenvalue weighting [2] was employed.

0-7803-8463-6/04/$20.00 ©2004 IEEE

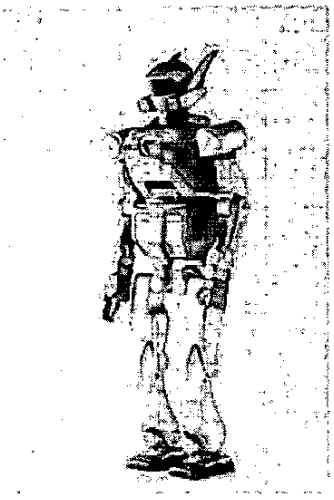

Fig. 1. Humanoid HRP-2.

This method can be summarized as follows:

$$\mathbf{R} = E[\mathbf{x}(\omega,t)\mathbf{x}(\omega,t)^H] \quad (1)$$

$$\mathbf{R} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^{-1} \quad (2)$$

$$\mathbf{P}(\theta,\omega) = \frac{\mathbf{v}^H(\omega,\theta)\mathbf{v}(\omega,\theta)}{\sum_{m=N+1}^{M}|\mathbf{v}^H(\omega,\theta)\mathbf{e}_m|^2} \quad (3)$$

$$\bar{P}(\theta) = \sum_{\omega=\omega_L}^{\omega_H}\bar{\lambda}(\omega)P(\omega,\theta) \quad (4)$$

$$\bar{\lambda}(\omega) = \sum_{m=1}^{N}\lambda_m(\omega) \quad (5)$$

First, the spatial correlation is calculated by (1). Here, $\mathbf{x}(\omega,t)$ is termed the input vector, which consists of the short-time Fourier transform of the input signal at microphones. Then, the eigenvalue decomposition is obtained using (2), where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1,\cdots,\lambda_M)$ and $\mathbf{E} = [\mathrm{e}_1,\cdots,\mathrm{e}_M]$ are the eigenvalue matrix and eigenvector matrix, respectively. The eigenvalues are assumed to be sorted in descending order. The standard MUSIC method is denoted as (3). The symbols M and N denote the number of microphones and sound sources, respectively. The vector $\mathbf{v}(\omega,\theta)$ is termed location vector, which consists of the transfer functions of the direct paths from the virtual sound source located in direction θ to the microphones. The broadband extension with the eigenvalue weighting is denoted as (4), where the weight $\bar{\lambda}$ is defined as (5). The eqs. (1)-(5) are evaluated and the location of the sound sources is estimated at every 0.5 s in this paper.

B. Measurement

For the operation of the sound localization shown in the previous section, the location vector, $\mathbf{v}(\omega,\theta)$, must be prepared. When the microphone array configuration is simple such as a linear or circular one, the location vector can be calculated by geometric information of the microphone array configuration. However, when a microphone array is mounted on the complicated surface as in the case of this paper, the location vector must be measured discretely in advance of the operation.

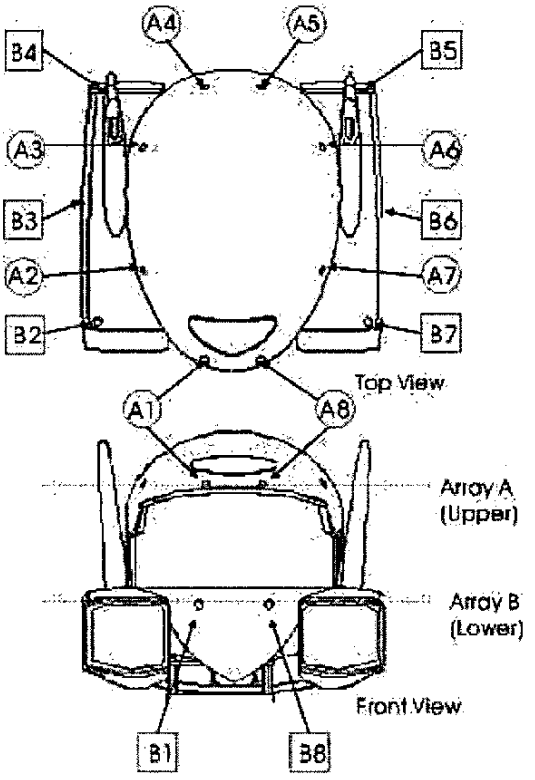

Fig. 2. Configuration of the microphone arrays.

In the measurement, a mock-up of the head of HRP-2, which was mounted on a camera stand, was used. Figure 2 shows the microphone array configuration and Fig. 3 shows its appearance. Two microphone-arrays, denoted as Array A (upper level) and Array B (lower level), were employed. Each microphone array consists of 8 microphones. Two microphone array sets were employed in order to evaluate the effect of the complicated surface of the robot head on the performance of the array processing.

Figure 4 shows a scene of the measurement of the location vector. The small square markers surrounding the mock-up show the points of virtual sound sources. The loudspeaker was placed at these points and the impulse responses from the loudspeaker to the microphones were measured. For measuring the impulse response, the time-stretched pulse (TSP) method was used [9]. The sampling frequency was 16 kHz. Then, the portion of the impulse response corresponding to the direct sound depicted in Fig. 5 was extracted. In this figure, the symbol T_w denotes the length from the largest peak of the impulse response which determines the length of the direct portion. In this paper, $T_w = 32$ was employed. The Fourier transform of the im-

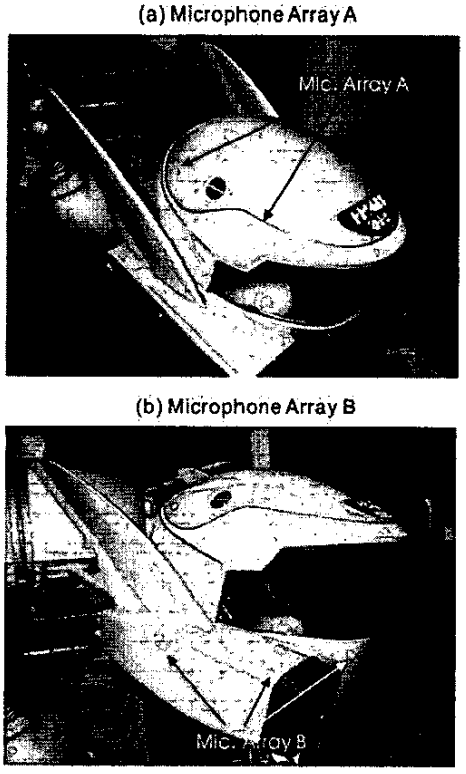

Fig. 3. Microphone arrays mounted on the head of HRP-2.

pulse response, $V_m(\omega, \theta)$, then becomes an element of the location vector as $\mathbf{v}(\omega, \theta) = [V_1(\omega, \theta), \cdots, V_M(\omega, \theta)]^T$. The length of the Fourier transform was 512.

C. Results

Figure 6 shows the spatial spectrum $\bar{P}(\theta)$ obtained by (4). The sound sources were located at 0° and 60° . The frequency range was [500, 3000] Hz. The location of the sound sources was well estimated by both array sets, and the effect of the complicated shape of the robot head is considered to be small. Comparing the performance of Array A and Array B, the spatial spectrum with Array A (upper level) has duller peaks than that with Array B. The main reason for this is the effective size of the array. Array A was mounted on the top of the head with a smaller interval of microphones. This effect is more clearly seen in the spatial spectrum for each frequency $P(\omega, \theta)$ depicted in Fig. 7. At the lower frequencies, no clear peaks were observed in the direction of 0° and 60° . This is due to the phase difference between the microphones being smaller. Therefore, Array B is employed hereafter in this paper.

III. Human Tracking by Vision

Our human tracking method consists of two processes. One is a finding a face by using a skin-colored model and template matching of that face in chromatic color space, and the other is tracking by a kernel-based face model.

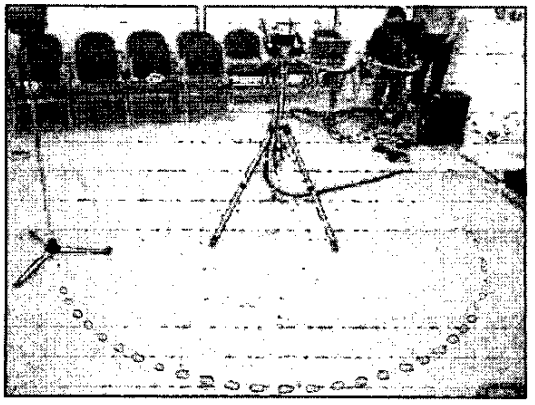

Fig. 4. Scene of measurement.

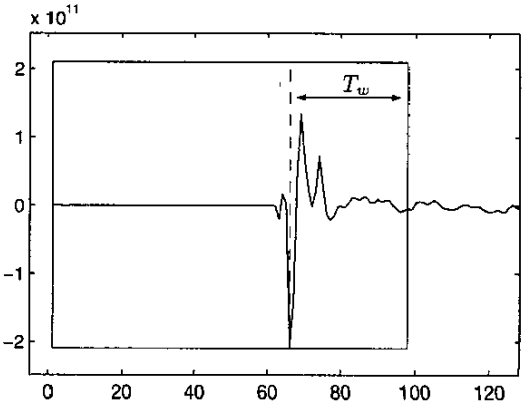

Fig. 5. Measured impulse response. The portion surrounded by a square is assumed to be direct sound.

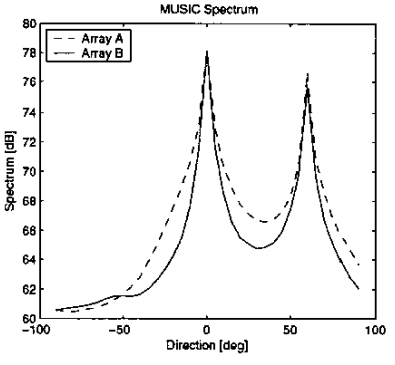

Fig. 6. Eigenvalue-weighted broadband MUSIC spectrum.

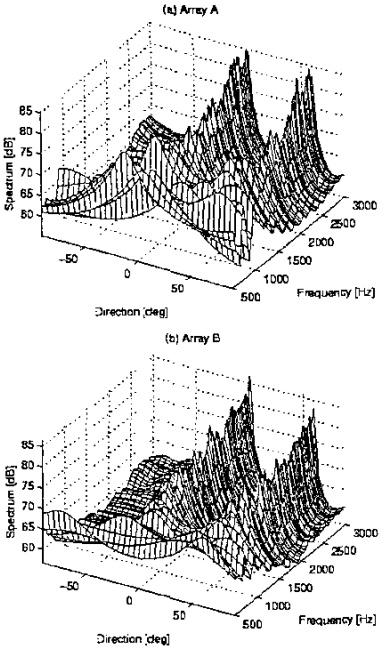

Fig. 7. MUSIC spectrum at each frequency.

A. Finding A Human Face

To reduce the computational costs of face finding, we used a skin-colored model in chromatic color space designed to characterize a human face. Since the color of human skin is made up of sufficiently distinctive chrominances, we believe it is an effective way to direct a robot's attention towards humans without requiring knowledge of he human shape or other high-level models. This model is adaptable to different people and different lighting conditions in dynamic environments.

Most devices for capturing images use RGB representation. However, RGB is not necessarily the best color representation for characterizing skincolor, because the RGB representation (r, g, b) includes not only color but also brightness. If the corresponding elements at two points (r_1, g_1, b_1) and (r_2, g_2, b_2) are proportional, they have the same color but different brightness.

Therefore, if the color space can be normalized as in the equations $R = r/(r+g+b), G = g/(r+g+b)$, general skin color becomes an identifiable constant. After this transform, the skin-color model can be represented by a two-dimensional Gaussian model $N(m, \sum^2)$ where m is the mean vector of (R, G) and $\sum$ is the covariance matrix.

Although human skin colors fall into a cluster in the chromatic color space, background colors, such as colored cloth or wooden bookshelves, may have an influence on the skin-color model. Therefore, to find the target face region in a scene, we apply a template matching method against each skin-color region.

Fig. 8. Result of skin-color detection

The procedure for detecting the face region is as follows:

1) Taking a face image or a set of face images, select the skin-colored region, and estimate the skin-color model from the mean and the covariance of their color distribution in chromatic color space in advance.
2) Calculate the likelihood of source pixels in a current frame matched against the model skin-color distribution by an appropriate error threshold.
3) Compute normalized squared differences between template images of the face and each skin-color region, and detect a candidate face region in a scene.

An experimental result of detecting a skin-color region and a face region is shown in Fig. 8. Black pixels in the right image indicate the skin-color region. Red rectangles and a green rectangle represent candidate face regions and a target face region, respectively.

B. Kernel-based Face Tracking

For real-time human tracking in a dynamic environment, it is desirable to keep the computational complexity of a human tracker as low as possible. The above-mentioned template matching process in face finding requires too much time to detect several faces at the same time. Therefore, a faster procedure is required to track face regions in realtime. The kernel-based approach [6] is used for the tracking procedure and can be summarized as follows:

First we define the target model and target candidate at location $\mathbf{y}$ in the subsequent frame as

$$\text{target model:} \quad \hat{\mathbf{q}} = \{\hat{q}_u\}_{u=1\ldots m} \qquad \sum_{u=1}^{m} \hat{q}_u = 1$$
$$\text{target candidate:} \quad \hat{\mathbf{p}}(\mathbf{y}) = \{\hat{p}_u(\mathbf{y})\}_{u=1\ldots m} \qquad \sum_{u=1}^{m} \hat{p}_u = 1$$

To find the location corresponding to the target in the current frame, the distance which is defined as $d(\mathbf{y}) = \sqrt{1 - \rho[\hat{\mathbf{p}}(\mathbf{y}), \hat{\mathbf{q}}]}$ should be minimized as a function of $\mathbf{y}$. Chooseing

$$\hat{\rho}(\mathbf{y}) \equiv \rho[\hat{\mathbf{p}}(\mathbf{y}), \hat{\mathbf{q}}] = \sum_{u=1}^{m} \sqrt{\hat{p}_u(\mathbf{y})\hat{q}_u} \quad ,$$

the next location $\mathbf{y}_{next}$ can be derived to maximize the $\hat{\rho}(\mathbf{y})$.

In our face tacker, we chose the target model $\hat{q}$ as normalized m-bin histograms of an ellipsoidal region in the chromatic color space. Thus, we have the model

$$\hat{q}_u = C \sum_{i=1}^{n} k(\|\mathbf{x}_i\|^2)\delta[b(\mathbf{x}_i) - u]$$

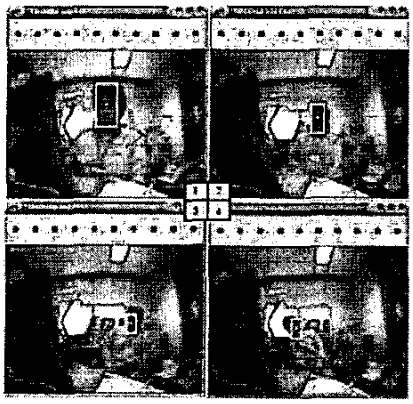

Fig. 9. Result of kernel-based tracking

where $\{\mathbf{x}_u\}$ is the normalized pixel location in the region defined as the target model, $k(x)$ is an isotropic kernel which is assigned a smaller weight to a pixel farther from the center, $b(\mathbf{x}_i)$ is the index of m-bin hostogram, and δ is the Kronecker delta function.

An experimental result of tracking face region is shown in Fig. 9.

IV. Information Fusion

A. System

In this section, an information fusion system is briefly introduced [2], [3], [4]. Figure 10 shows a block diagram of the information fusion system. The results of the sound localization and the human tracking by vision are fused by a Bayesian network. Figure 11 shows the Bayesian network used in this system and the corresponding audio and video information. The observation space of the sound localization is divided into small regions and each region is assigned to an audio input node of the network. In each node, occurrence of a sound event is detected by examining peaks in the spatial spectrum, and the node is activated when a sound event occurs in the corresponding region. In a similar manner, the video observation space is divided and assigned to video input nodes. In the Bayesian network, the co-occurrence of the audio event and the video event in a corresponding region is detected as a speech event. The information of the detected speech event is utilized in updating the separation filter in the sound separation and the segmentation of speech in the automatic speech recognition (ASR). In the separation filter, the target speech event is separated from noise and interferences by using a maximum likelihood beamformer.

B. Results

The experiment was conducted in the same room where the measurement was performed. Using the microphone array and the camera mounted on HRP-2, approximately 20 s of data was recorded and processed. Figure 12 shows a scene of the experiment taken by the camera mounted on HRP-2. In the scenario used in the recorded data, a

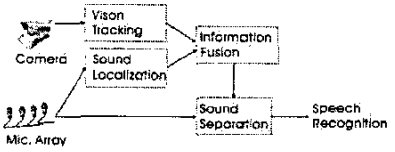

Fig. 10. Block diagram of the information fusion system.

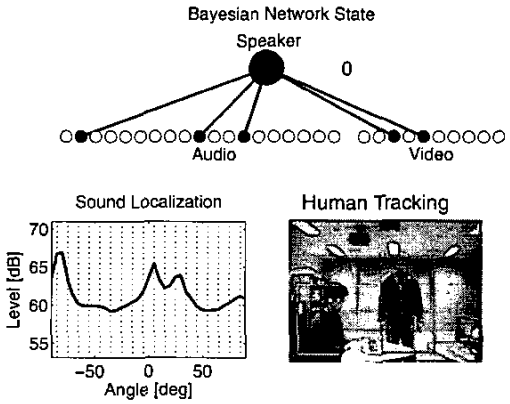

Fig. 11. Bayesian network and the audio and video information.

noise source (loudspeaker) was located in the direction of $-50°$. Also, as noise sources, computers were located at 60° and generated fan noise. Speaker 1 kept standing in the direction of $-20°$ and spoke between 0 s and 5 s. Then, Speaker 2 walked in, spoke at around 10 s and walked out.

These events can be observed in the audio and video information depicted in Fig. 13. In the video information in Fig. 13(b), the trajectory of Speaker 2 walking out can be seen but that of walking in cannot. This is because it takes a few seconds to find humans by the human finder. These data were then digitized and the feature vectors which indicate the state of the input nodes of the Bayesian network as depicted in Fig. 14 were obtained.

Figure 15 shows the results of the speech event detection obtained from the inference by the Bayesian network. In this figure, it can be seen that the speech events by Speaker 1 and 2 were extracted from the various audio events shown in Fig. 14(a). Figure 16 shows the waveform observed at one of the microphones (before separation) and that observed at the beamformer output. The segments of estimated and true speech events are also depicted by bars in the upper part of Fig. 16(b). From these bars, it can be seen that the estimated speech events are in good agreement with the actual ones. By comparing Fig. 16(a) and (b), it can be seen that the speech signal buried in noise was recovered by the beamforming.

V. Conclusion

In this paper, a method of detecting and separating speech events was applied to humanoid HRP-2 and some basic experiments were conducted. From the results, it was

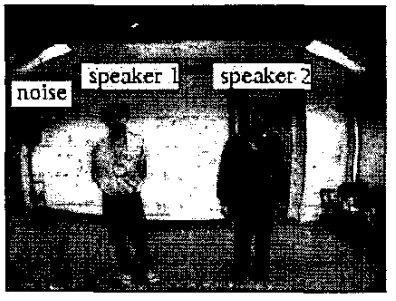

Fig. 12. An example of the images used for human tracking.

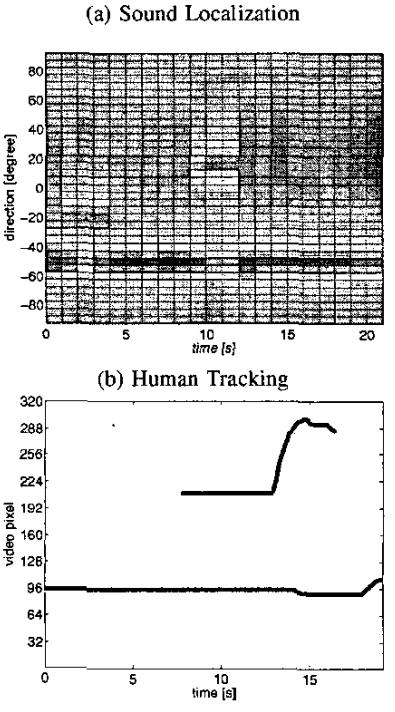

Fig. 13. Results of sound localization and human tracking.

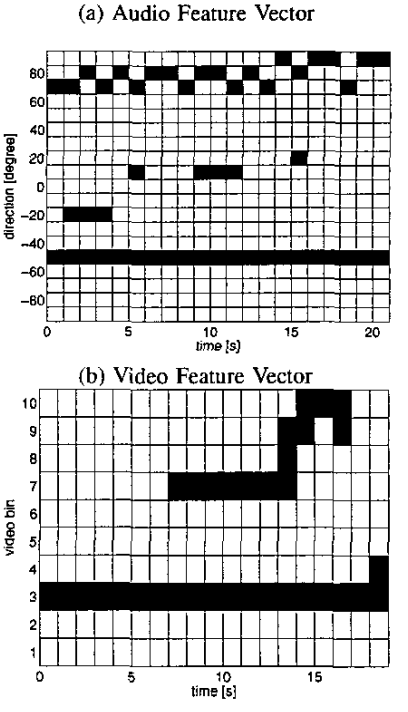

Fig. 14. Audio and video feature vectors. Black squares indicate that the corresponding nodes are active.

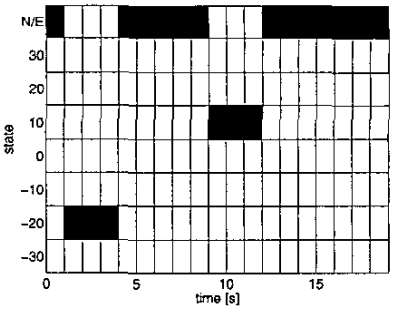

Fig. 15. Detected speech events. "N/E" indicates "no speech event."

shown that a performance similar to that obtained by a regular microphone array (circular in shape with a diameter of 0.5 s, 8 elements) shown in [4] was obtained by the microphone array mounted on the humanoid HRP-2. Also, a newly introduced human finding and tracking system performed well in the tested conditions. As a next step of this study, a real-time system, which can be mounted on the body of HRP-2, is currently being developed. By using this system, a more realistic evaluation, including speech recognition, should be realized.

REFERENCES

[1] Takakatsu Isozumi, Kazuhiko Akachi, Shigehiro Ota, Fumio Kanehiro, Kenji Kaneko, and Hirohisa Hirukawa, "Humanoid Robot HRP-2," in *The 21st Annual Conf. of RSJ*, 2003, 3A32.

[2] Futoshi Asano, Yoichi Motomura, Hideki Asoh, Takashi Yoshimura, Naoyuki Ichimura, and Satoshi Nakamura, "Fusion of audio and video information for detecting speech events," in *Proc. Fusion 2003*, 2003, pp. 386–393.

[3] Futoshi Asano, Yoichi Motomura, Hideki Asoh, Takashi Yoshimura, Naoyuki Ichimura, Kiyoshi Yamamoto, Nobuhiko Kitawaki, and Satoshi Nakamura, "Detection and separation of speech segment using audio and video information fusion," in *Proc. Eurospeech2003*, September 2003, pp. 2257–2260.

[4] F. Asano, K. Yamamoto, J. Hara, J. Ogata, T. Yoshimura, Y. Motomura, N. Ichimura, and H. Asoh, "Detection and separation of speech event using audio and video information fusion and its application to robust speech interface," *accepted by J. Applied Signal Processing*, 2004.

[5] K. Nakadai, K. Hidai, H. Mizoguchi, H.G. Okuno, and H. Kitano, "Real-Time Auditory and Visual Multiple-Object Tracking for Humanoid," *Proc. of IJCAI2001*, pp.1424–1432, 2001.

[6] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object

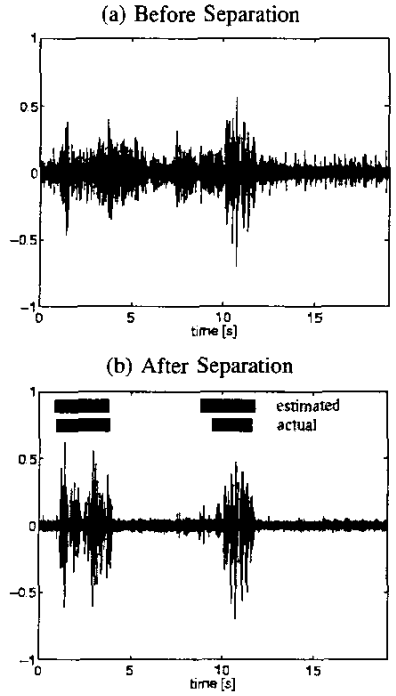

Fig. 16. Input and output waveform in the sound separation. In (b), the bars on the waveform indicate the "estimated" and "actual" speech segments, respectively. The "actual" speech segments were detected by a human listener.

tracking," *IEEE Trans. Pattern Analysis Machine Intell.*, vol. 25, no. 5, pp. 564–575, 2003.

[7] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag*, vol. AP-34, no. 3, pp. 276–280, March 1986.

[8] Yoiti Suzuki, Futoshi Asano, Hack Yoon Kim, and Toshio Sone, "An optimum computer-generated pulse suitable for the measurement of very long impulse response," *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1119–1123, 1993.

[9] Yoiti Suzuki, Futoshi Asano, Hack Yoon Kim, and Toshio Sone, "An optimum computer-generated pulse suitable for the measurement of very long impulse response," *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1119–1123, 1993.