EURASIP Journal on Applied Signal Processing 2004:11, 1727–1738

Detection and Separation of Speech Event Using Audio and Video Information Fusion and Its Application to Robust Speech Interface

Futoshi Asano,[1] Kiyoshi Yamamoto,[2] Isao Hara,[1] Jun Ogata,[1] Takashi Yoshimura,[1]
Yoichi Motomura,[1] Naoyuki Ichimura,[1] Hideki Asoh[1]

[1] *Information Technology Research Institute, National Institute of Advanced Industrial Science and Technology,*
Tsukuba 305-8568, Japan
Emails: f.asano@aist.go.jp, isao-hara@aist.go.jp, jun.ogata@aist.go.jp, yoshimur@ni.aist.go.jp,
y.motomura@aist.go.jp, nic@ni.aist.go.jp, h.asoh@aist.go.jp

[2] *Department of Computer Science,Tsukuba University, Tsukuba 305-8573, Japan*
Email: kyama@mmlab.is.tsukuba.ac.jp

Received 11 November 2003; Revised 3 February 2004; Recommended for Publication by Chin-Hui Lee

A method of detecting speech events in a multiple-sound-source condition using audio and video information is proposed. For detecting speech events, sound localization using a microphone array and human tracking by stereo vision is combined by a Bayesian network. From the inference results of the Bayesian network, information on the time and location of speech events can be known. The information on the detected speech events is then utilized in the robust speech interface. A maximum likelihood adaptive beamformer is employed as a preprocessor of the speech recognizer to separate the speech signal from environmental noise. The coefficients of the beamformer are kept updated based on the information of the speech events. The information on the speech events is also used by the speech recognizer for extracting the speech segment.

Keywords and phrases: information fusion, sound localization, human tracking, adaptive beamformer, speech recognition.

1. INTRODUCTION

Detection of speech events is an important issue in automatic speech recognition (ASR) in a real environment with background noise and interferences. Also, the detection of the presence or absence of the target speech signal is often important for noise reduction such as adaptive beamformer (see, e.g., [1]) or spectral subtraction (see, e.g., [2]), which can be used as a preprocessor of ASR. In the maximum likelihood (ML) adaptive beamformer employed in this paper, the spatial correlation of the noise must be estimated during the absence of the target signal as described later in this paper. In the spectral subtraction, the spectrum of the noise must be estimated in a way similar to that of the ML beamformer.

When environmental noise is nonspeech signals, a voice activity detector (VAD) can be used as a target speech detector (see, e.g., [3]). In environments such as offices and homes, however, not only the target but also interference from sources such as a TV or a radio can be speech signals. In such cases, the detection of the target speech cannot be accomplished only by using sound information, and fusion with the information from other modalities such as vision is necessary.

Chaodhury et al. [4] proposed a speech event detector using audio and video information. In their paper, an environment in which a dialog existed between multiple speakers and a Smart Kiosk terminal [5] was considered. The speech events which were addressed only to the terminal were detected by using a dynamic Bayesian network based on the video information of face and the power level of sound. Thus, the main focus of that paper was detection of the attention of speakers to the terminal. In this system, therefore, it was assumed that only a single sound event occurs at a single moment. Also, information on the location of the target and other sound sources in the audio and video information was not utilized.

The main focus of this paper is the detection of speech events under the circumstance in which multiple sound events occur at the same time (e.g., speaking in the presence of TV sound). The location and time information obtained from audio and video observation is fused by a Bayesian network so that the time and location of the speech event can be estimated [6, 7].

The estimated information on the speech events is then utilized in the speech interface so that the speech signals

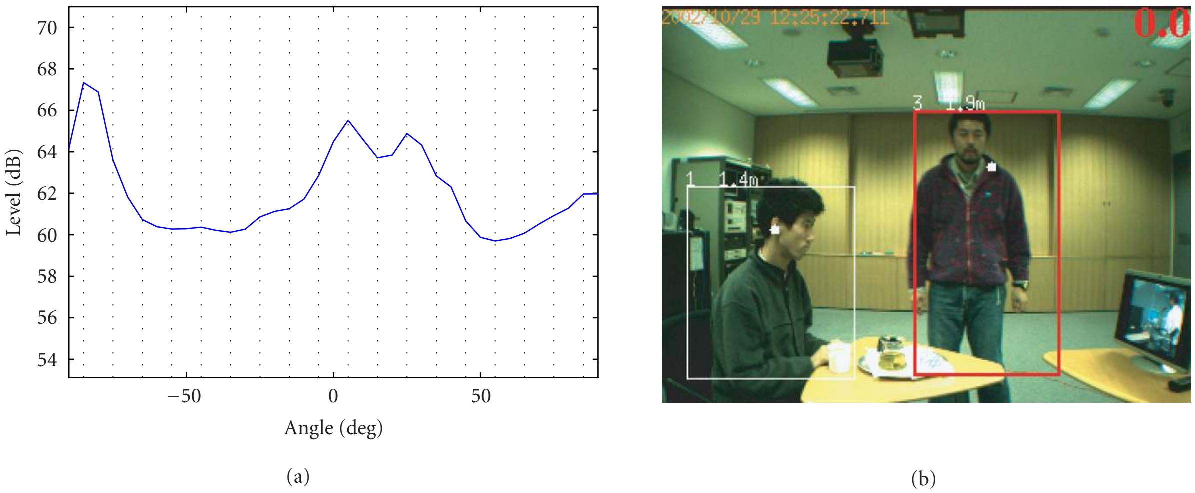


Figure 1: Example of (a) the audio information (spatial spectrum) and (b) the video information (human tracking).

spoken in environmental noise and interferences are robustly recognized. In this system, an ML adaptive beamformer is employed as a preprocessor of ASR to separate speech signals from environmental noise. The beamformer coefficients are updated based on the information on the speech events. The information on the speech events is also utilized in ASR for the segmentation of speech [8].

This paper is organized as follows. In Section 2, methods of extracting time and location information from the audio and video observation are briefly reviewed. In Section 3, a speech event detector based on information fusion is developed. In Section 4, a speech interface is constructed. In Section 5, the performance of the entire system is evaluated in a real environment.

2. FEATURE EXTRACTION FROM AUDIO AND VIDEO INFORMATION

In this section, the methods of sound localization and human tracking employed in this paper are briefly reviewed to facilitate the understanding of the following sections.

2.1. Sound localization

For sound localization, the MUSIC method [9] extended to a broadband signal with eigenvalue weighting [6] is used.

We denote the input vector as $\mathbf{x}(\omega, t) = [X_1(\omega, t), \ldots, X_M(\omega, t)]^T$, where $X_m(\omega, t)$ denotes the short-time Fourier transform of the input signal to the mth microphone. From this input vector, the spatial correlation is estimated as

$$\mathbf{R}(\omega) = E\left[\mathbf{x}(\omega, t)\mathbf{x}^H(\omega, t)\right]. \tag{1}$$

Using the eigenvectors of $\mathbf{R}(\omega)$ corresponding to the smallest $M - N$ eigenvalues, $\{\mathbf{e}_{N+1}, \ldots, \mathbf{e}_M\}$, the MUSIC spatial spectrum estimator is defined as

$$P(\theta, \omega) = \frac{\left|\mathbf{g}(\theta, \omega)\right|^2}{\sum_{m=N+1}^{M} \left|\mathbf{e}_m^H \mathbf{g}(\theta, \omega)\right|^2}, \tag{2}$$

where M and N denote the number of microphones and the number of sound sources, respectively. The symbol $\mathbf{g}(\theta, \omega)$ denotes the location vector of the virtual source in the arbitrary direction θ. The elements of the location vector are the transfer functions of the *direct* path from the virtual source to the microphones. To estimate the final spatial spectrum for the broadband input, (2) is averaged over the frequency of interest as

$$\bar{P}(\theta) = \sum_{\omega=\omega_l}^{\omega_h} \bar{\lambda}(\omega) P(\theta, \omega), \tag{3}$$

where $\bar{\lambda}$ is the eigenvalue weight [6] defined as

$$\bar{\lambda} = \sum_{n=1}^{N} \lambda_n. \tag{4}$$

The symbol λ_n is the nth eigenvalue of $\mathbf{R}(\omega)$. The eigenvalues are assumed to be sorted in descending order. By doing this, the frequency bins in which the power of the directional signal is dominant have larger weights. The range $[\omega_l, \omega_h]$ denotes the frequency range of interest.

Figure 1a shows an example of the spatial spectrum estimated by (3). From the peaks of this spectrum, the location (direction) of the sound sources can be estimated.

2.2. Human tracking by vision

There are many methods to track humans using video information. As a human tracker used in the proposed

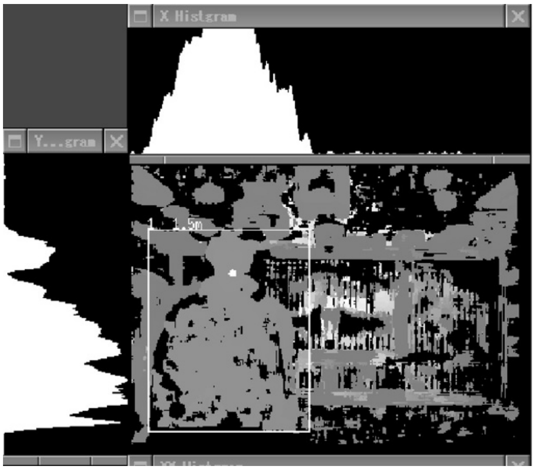


Figure 2: Example of the background subtraction for human tracking.

information fusion framework, any method which gives the position (pixels) of humans in the observed image can be used. In this paper, humans in a scene are detected by background subtraction based on the range image obtained by a stereo camera (see, e.g., [10]) for the sake of simplicity. Figure 2 shows the process of the background subtraction. The main panel shows an example of the subtracted range image (difference between the current input range image and the pre-recorded background range image). The upper and left panels show histograms of the difference in a vertical and a horizontal slice, respectively. Regions in which the value of the histogram of the difference is above a certain threshold are recognized as foreground objects (humans).

3. AUDIO-VIDEO INFORMATION FUSION

In this section, the audio and video information is fused to detect the speech events.

3.1. Basic concept

As described in the previous section, the location and time of the audio events (emission of sound from sound sources) can be estimated by examining the audio information (spatial spectrum). Using this audio information, "virtual audio sensors" which observe a certain region of the entire observation space and detect audio events are formed.

Also, the location and time of the video events (existence of humans) can be estimated from human tracking using vision. In the same way as in the case of audio sensors, virtual video sensors which detect the video events in a certain region of the video observation space are formed.

By combining the information of the audio and video events detected by these virtual audio and video sensors, the co-occurrence of audio and video events within a certain region can also be estimated. This co-occurrence of audio and video events is detected as "speech events."

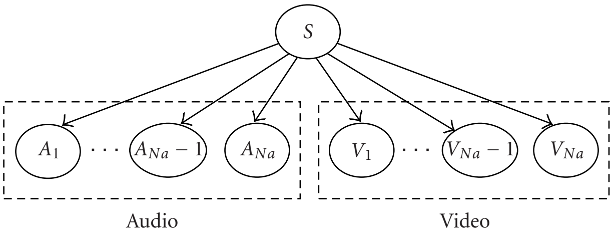


Figure 3: Bayesian network for fusing audio and video information.

3.2. Bayesian network used for information fusion

In this paper, a Bayesian network (see, e.g., [11]) is used to detect the co-occurrence of the audio and video events. The Bayesian network is a way of modeling a joint probability distribution of multiple random variables and is considered to be a powerful tool for information fusion [12].

Figure 3 shows the topology of the Bayesian network used in this paper. The network has N_a audio nodes, $\{A_1, \ldots, A_{N_a}\}$, and N_v video nodes, $\{V_1, \ldots, V_{N_v}\}$, as input nodes. These nodes correspond to the virtual audio and video sensors described in the previous subsection. The input nodes have the states of $\{0, 1\}$ according to the occurrence of the corresponding audio/video events.

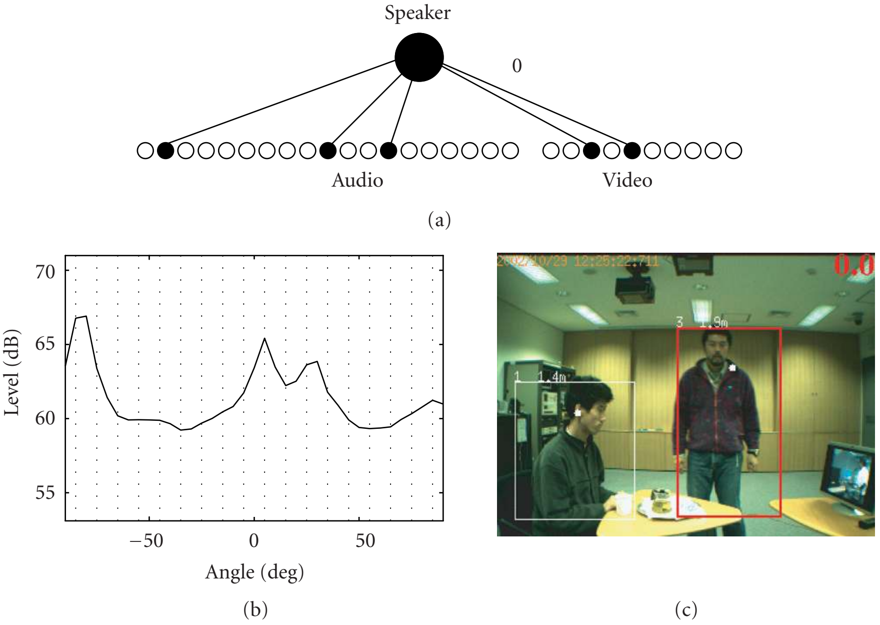
On the other hand, the output node S has the following $N_s + 1$ states: $S = \{S_1, \ldots, S_{N_s}, \text{NoEvent}\}$. The state $\{S_1, \ldots, S_{N_s}\}$ corresponds to the speaker's position (angle): $\{S_1, \ldots, S_{N_s}\} = \{-30^\circ, \ldots, +30^\circ\}$. For example, when $S = -30^\circ$, the speaker is located in the direction of 30°and is speaking. When $S = \text{NoEvent}$, there are no speech events.

3.3. Feature vector

Figure 4 shows an example of the state of the Bayesian network and the corresponding audio and video information.

Regarding the audio information, the spatial spectrum (Figure 4b) is divided into the $N_a(= 19)$ regions (-90°–$+90^\circ$, every 10°). The vertical lines in this panel indicate the divided regions, which correspond to the virtual audio sensors and are assigned to the 19 audio nodes of the Bayesian network. In each region, the peak in the spatial spectrum is detected. According to the peak detection, the state of the corresponding node is set at $\{0, 1\}$ ("1" corresponds to the peak being detected). Hereafter, the vector containing the state of the audio input node, $\mathbf{a}(t) = \{A_1(t), \ldots, A_{N_a}(t)\}$, is referred to as the audio feature vector, in which $A_i(t)$ denotes the state of the ith node at time t. Figures 5a and 5b show an example of the audio information (running spatial spectrum) and the corresponding feature vector, respectively. The vertical slice in Figure 5b corresponds to a single feature vector, $\mathbf{a}(t)$.

In the same way as in the case of audio information, the video observation space is divided into $N_v(= 10)$ regions (1–320 pixels, every 32 pixels), and these regions are assigned to the video input nodes. In each region, the existence of a human is examined by using a human tracker, and the video feature vector $\mathbf{v}(t) = \{V_1(t), \ldots, V_{N_v}(t)\}$ is formed. Figures 5c and 5d show the video information (results of human tracking) and the corresponding video feature vector $\mathbf{v}(t)$, respectively.



(a)

(b)

(c)

Figure 4: Example of a state of the Bayesian network and the corresponding audio and video information. (a) Bayesian network state; (b) sound localization; and (c) human tracking.

3.4. Inference of the Bayesian network

Next, a method of estimating the state of the output node S, given the observation (audio and video feature vectors), is described. The state of S can be determined by estimating the conditional probability of S, $P(S|A_1,\dots,A_{N_a},V_1,\dots,V_{N_v})$.

In this paper, it is assumed that the values of all A_i and V_j are conditionally independent, given the value of S. In general, audio and visual observations are strongly correlated through a common cause which is described by the value of S. However, when S is fixed to a certain value (state), the distributions of audio and visual observations become nearly independent because the effect of the common cause is removed from the distribution and only independent noisy fluctuation of observations remains.

Based on this assumption, the conditional probability distribution $P(S|A_1,\dots,A_{N_a},V_1,\dots,V_{N_v})$ can be factored into the product of the local conditional probabilities $P(A_i|S)$ and $P(V_j|S)$:

$$P(S|A_1,\dots,A_{N_a},V_1,\dots,V_{N_v}) = \frac{1}{Z}P(S)\prod_{i=1}^{N_a}P(A_i|S)\prod_{j=1}^{N_v}P(V_j|S), \tag{5}$$

where

$$Z = \int_S P(S)\prod_{i=1}^{N_a}P(A_i|S)\prod_{j=1}^{N_v}P(V_j|S)\,dS. \tag{6}$$

The state of S is estimated by evaluating (5) with the observations, $\mathbf{a}(t)$ and $\mathbf{v}(t)$, and the previously estimated conditional probabilities, $P(A_i|S)$ and $P(V_j|S)$. In this paper, the prior distribution of S, $P(S)$, is assumed to be uniform.

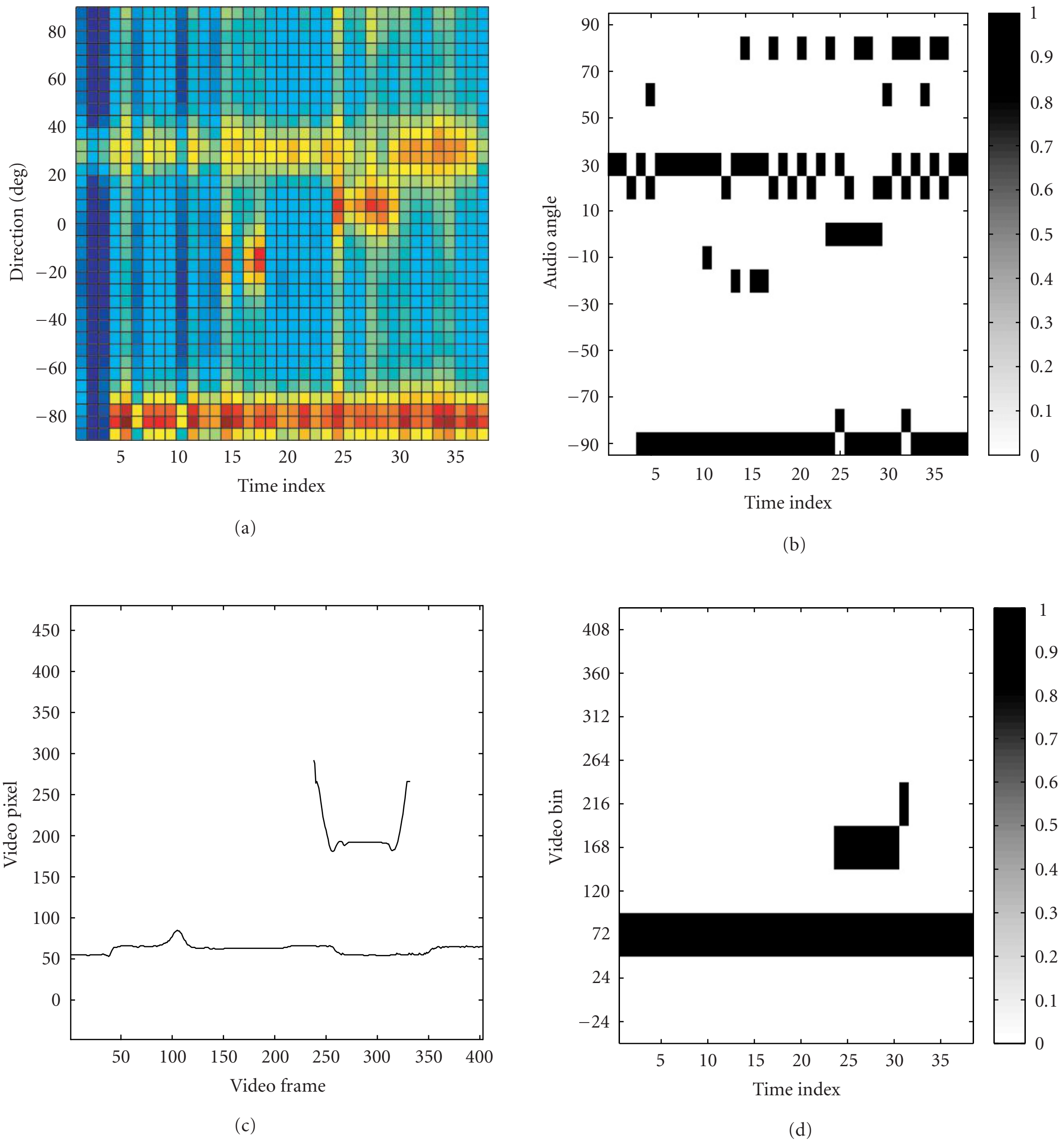
3.5. Learning of the Bayesian network

The conditional probabilities, $P(A_i|S)$ and $P(V_j|S)$, can be estimated from training samples prior to the actual operation. Since the process of estimating these conditional probabilities is supervised learning, the state of S is given as a supervisor for each set of observations, $\mathbf{a}(t)$ and $\mathbf{v}(t)$.

The table describing $P(A_i|S)$ and $P(V_j|S)$ is termed conditional probability table (CPT), and it functions as the provider of the correspondence between the audio and the video observation spaces. This correspondence can also be obtained by precise calibration. The advantage of obtaining this correspondence by learning with real data is to reflect fluctuations of the observation. By doing this, a more robust estimation can be expected.

In this paper, speech data spoken by a single speaker in an ordinary meeting room were used as training samples. The location of the speaker was varied between $-30°$and $+30°$every $5°$. In each sample, the speaker spoke intermittently for 30 seconds. There were no significant noise sources in the meeting room, but there was ordinary background noise such as that from an air conditioner and a PC fan. For these samples, the speech events were detected by a human operator, and the time and the location of the speech events were used as a supervisor for training CPT.

Figure 6 shows the audio and video CPTs obtained from the data described above. From this figure, the correspondence between the speaker's real location and the estimated location (angle) based on the audio information, and the estimated location (pixel) based on the video information can be determined. As shown in this figure, multiple regions in the audio and the video observation spaces are assigned to the real location of the speaker with weighting of the conditional probability. By using these CPTs, the fluctuation of observation is expected to be reduced to some extent.



(a)

(b)

(c)

(d)

Figure 5: Audio and video information and the corresponding feature vectors. (a) Running spatial spectrum (audio information); (b) audio feature vector corresponding to (a); (c) human tracking; (d) video feature vector corresponding to (c). A vertical slice of (a) and of (c) corresponding to Figures 1a and 1b, respectively.

4. SPEECH INTERFACE SYSTEM

Based on the speech event detector developed in the previous sections, the speech interface is constructed in this section.

4.1. Overview of the system

Figure 7 shows a block diagram of the entire system. The input signal is observed using a microphone array and the spatial spectrum as depicted in Figure 1a is estimated in the sound localization module. On the other hand, in the human tracking module, the tracking results as depicted in Figure 1b are obtained. Using these data, feature vectors, $\mathbf{a}(t)$ and $\mathbf{v}(t)$, depicted in Figure 5 are formed.

These feature vectors are then fed to the information fusion module. Using the feature vectors and the previously obtained CPTs, the location and the time of the speech event are estimated.

The information on the speech events is then sent to the sound separation module and the speech recognition

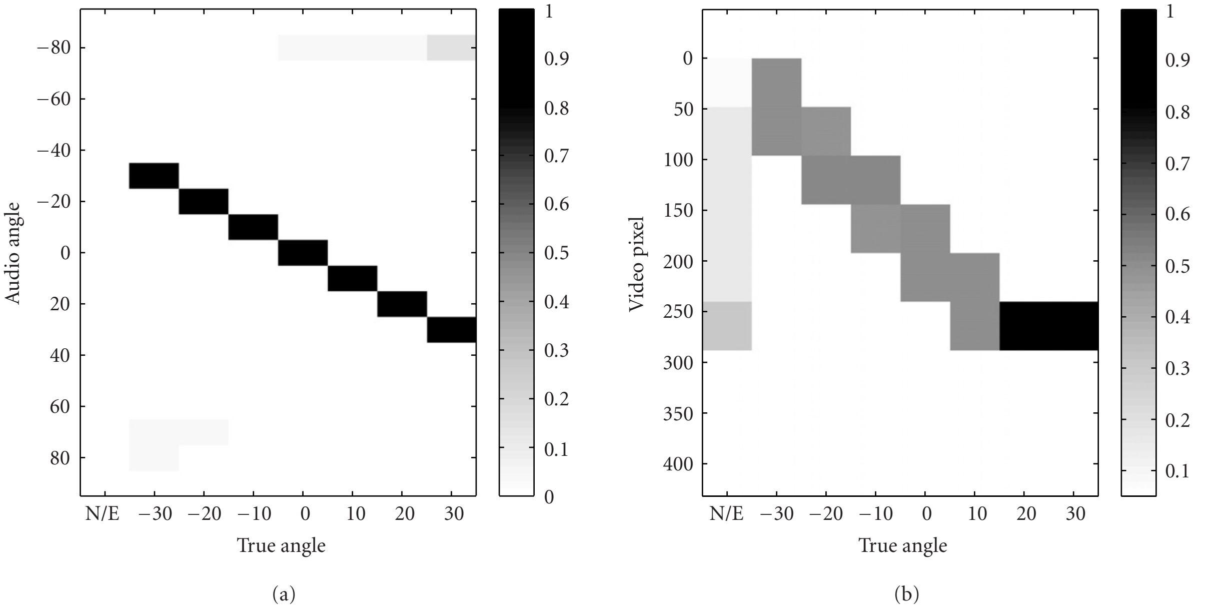


Figure 6: CPTs obtained from the training samples. "N/E" indicates the state "NoEvent." (a) CPT for audio; (b) CPT for video.

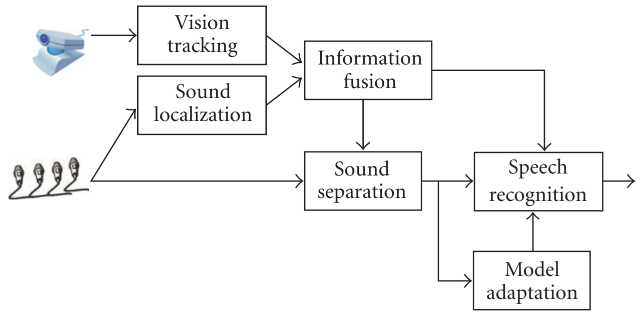


Figure 7: Block diagram of the proposed audio-video information fusion system.

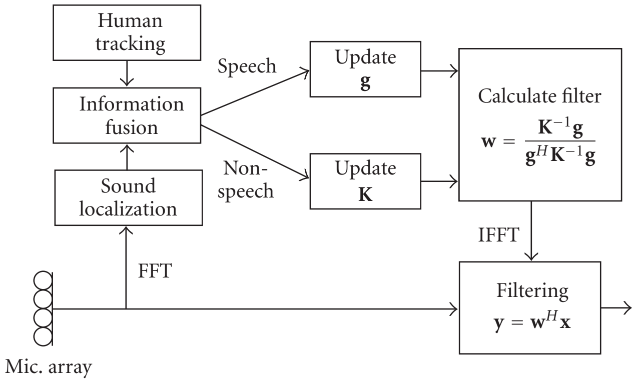


Figure 8: Block diagram of the ML beamformer.

module. In the sound separation module, the target speech is separated from the environmental noise and interferences by using the ML adaptive beamformer [1]. Based on the information on the speech events, the information of the ML beamformer is updated. As described in detail in Section 4.2, the location of the target speaker is updated when the speech event occurs. On the other hand, when there are no speech events, the spatial information of noise is updated. Using the updated beamformer coefficients, the input signal is processed.

In the speech recognition module, based on the speech event information, the segments corresponding to the speech events are extracted from the noise-reduced signal sent from the sound separation module and are recognized. The noise-reduced signal is also used for the adaptation of the acoustic model of the speech recognition in the model adaptation module. In this module, the acoustic model is updated so that it matches the residual noise of the sound separation.

4.2. ML beamformer

In the ML adaptive beamformer, the source spectrum is estimated as

$$y(\omega, t) = \mathbf{w}^H(\omega)\mathbf{x}(\omega, t), \tag{7}$$

where the beamformer coefficient vector is defined as

$$\mathbf{w}(\omega) = \frac{\mathbf{K}^{-1}(\omega)\hat{\mathbf{g}}(\omega)}{\hat{\mathbf{g}}^H(\omega)\mathbf{K}^{-1}(\omega)\hat{\mathbf{g}}(\omega)}. \tag{8}$$

The matrix $\mathbf{K}(\omega)$ is the noise spatial correlation observed in the *absence* of the target speech. The vector $\hat{\mathbf{g}}(\omega)$ is the location vector of the target speech source which can be estimated in the *presence* of the target speech.

Figure 8 is a block diagram of updating of the beamformer coefficients in the ML adaptive beamformer. Based on the information of the speech event detector, the noise

Table 1: Parameters of the speech recognizer.

Feature parameter	MFCC (26 dimensions)
Analysis frame length	25 ms
Analysis frame shift	10 ms
Number of phones	43
Number of mixtures	16/state
Vocabulary size	492

spatial correlation $\mathbf{K}(\omega)$ is updated in the absence of the target speech. In the presence of the target speech, on the other hand, the location vector for the target $\hat{\mathbf{g}}(\omega)$ is updated using the estimated location of the speech event.

The noise correlation $\mathbf{K}(\omega)$ used in the ML beamformer can be approximated by using the observed correlation matrix, $\mathbf{R}(\omega)$, in which both the target and the noise coexist in the speech segments. In this case, detection of the target speech is not necessary. The beamformer using the observed correlation matrix, $\mathbf{R}(\omega)$, is termed the minimum variance (MV) beamformer [1]. However, the performance of the noise reduction of the ML beamformer is much better than that of the MV beamformer if $\mathbf{K}(\omega)$ can be estimated separately [13]. In this paper, therefore, the ML beamformer was employed despite the cost of detecting the speech events.

4.3. Speech recognition and model adaptation

As a speech recognizer, HTK Ver.3.2 [14] was used. As the initial acoustic model, context-independent phone HMMs available in the Japanese dictation software [15] was used. The parameters of the speech recognizer are summarized in Table 1.

At the output of the sound separation module, the signal-to-noise ratio (SNR) is improved by the adaptive beamformer. However, less-directional noise, such as room reverberation in particular, cannot be removed perfectly. Therefore, adaptation of the acoustic model of the speech recognizer to residual noise is required. In this paper, model adaptation is executed simultaneously as a background process of the speech recognition and the acoustic model is kept updated. As a method of adaptation, a combination of MLLR [16] and MAP [17], in which MLLR-transformed means and variances were used as the priors for the MAP estimation, was employed [18]. Since online adaptation is assumed in this system, correctly labeled data for the adaptation are not available. In this paper, an unsupervised adaptation, in which the phonetic labels for parameter estimation were automatically generated using an initial recognizer, was employed [19].

5. EXPERIMENT

5.1. Condition

The experiments were conducted in a medium-sized meeting room with a reverberation time of 0.5 second. The setting of the experiments is shown in Figure 9. The microphone array used in the experiments was of a circular shape with a diameter of 0.5 m and had 8 microphones. The sampling frequency was 16 kHz. As a camera, Digiclops (Pointgray Research) was employed.

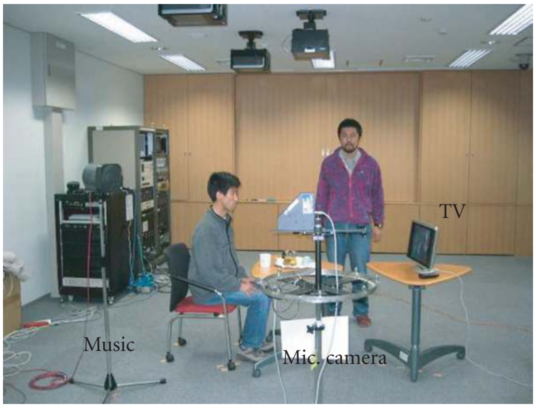


Figure 9: Setting of experiments.

5.2. Experiment 1

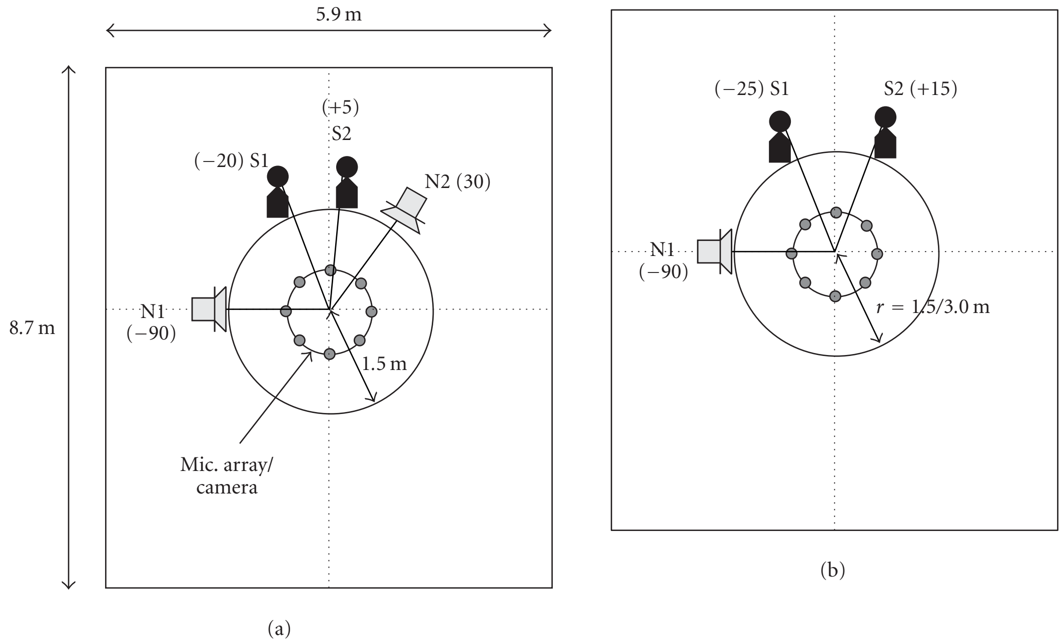
In experiment 1, the proposed speech event detection system was tested under a realistic scenario. The configuration of sound sources, humans, and microphone array/camera is shown in Figure 10a. The audio/video information and the feature vectors shown in Figure 5 correspond to this scenario. As noise sources, a TV and a loud speaker were employed. As shown in Figure 9, speaker #1 sat in front of the table during the entire observation and spoke a short sentence at around $t = 15$ seconds as depicted in Figure 5. Speaker #2 walked into the observation space at around $t = 20$ seconds, spoke a short sentence at around $t = 25$ seconds, then walked out.

Figure 11a presents the results of inference, which show the detected speech events in the time-direction plane. Comparing this with Figure 11b, which shows the true speech events, it can be seen that the speech events were detected fairly correctly.

Using the information on the detected speech events, the ML beamformer was updated at every time block (every 1 second) and the microphone-array input was processed. The input/output waveform is shown in Figure 12. In Figure 12b, bars indicating the detected and the true speech events are also shown. The segment in which the probability of S being in the state of $\{-30, \dots, +30\}$ is over 0.7 is shown as the detected speech segment. From these data, it can be seen that the speech signal which was almost buried by the interference was recovered by the ML beamforming.

5.3. Experiment 2

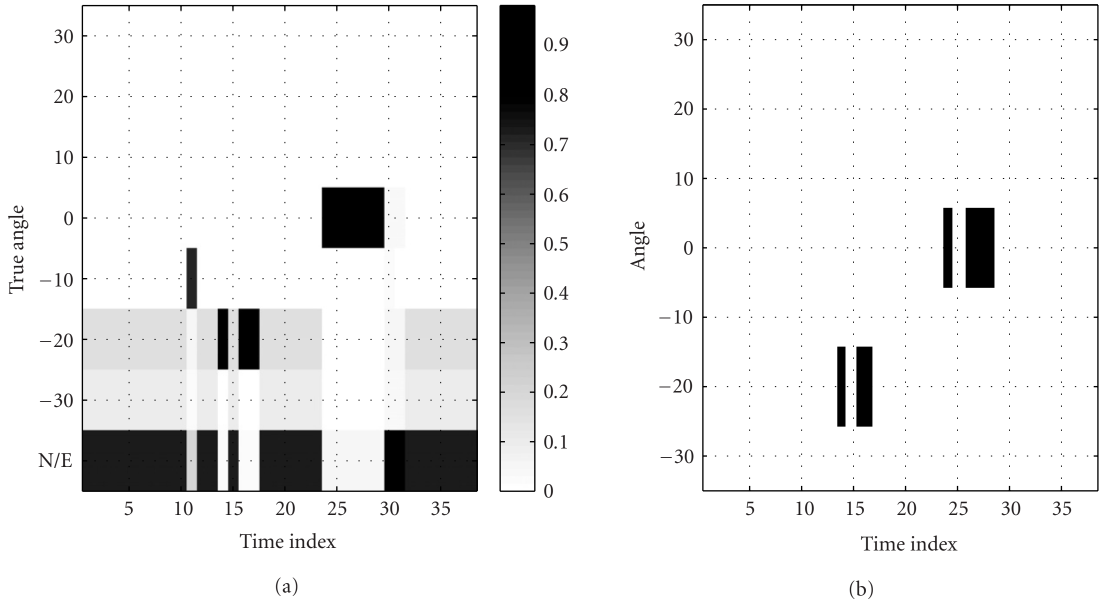
In experiment 2, a quantitative evaluation was conducted using detection rate and the score of ASR. The configuration of sound sources and humans is shown in Figure 10b. Speakers #1 and #2 spoke Japanese words alternatively without changing their position. The number of words in the test set was $N_w = 492$. The level of the noise was adjusted so that the SNR was around 0 dB in the previous test.



(a)

(b)

Figure 10: Sound source configuration in (a) Experiment 1 and (b) Experiment 2. The numbers in parentheses indicate angles. S1 and S2 denote humans (speech), N1 denotes loudspeaker (music), and N2 denotes TV (speech + music).



(a)

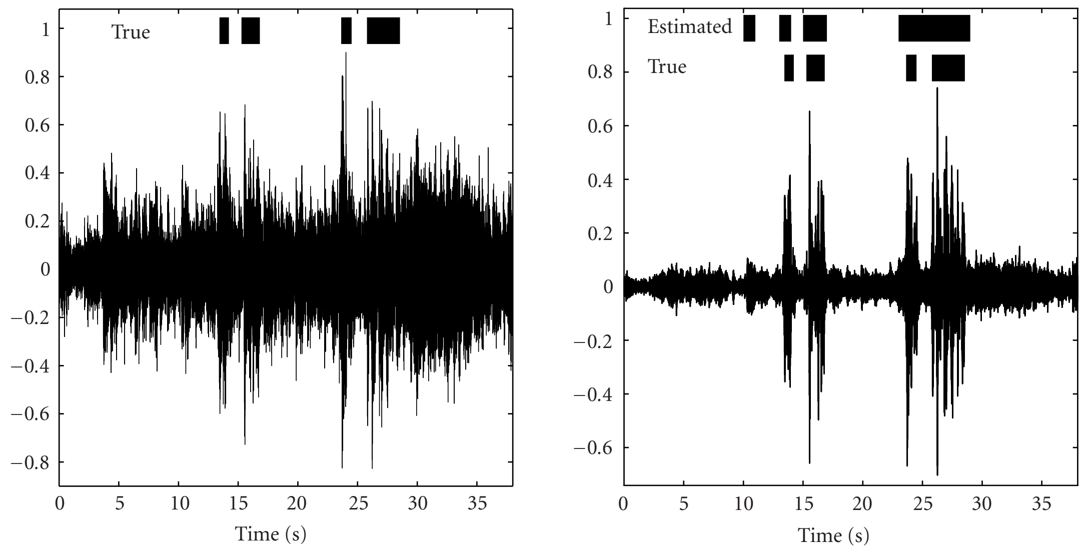
(b)

Figure 11: Detected and true speech events. (a) Inference results, (b) true speech segment.

Table 2 shows the detection rate R_d, the fitting rate R_f, and the coverage R_c defined as

$$R_d = \frac{\text{No. of correctly detected speech/nonspeech segments}}{\text{Total no. of segments}},$$
$$R_f = \frac{\text{No. of correctly detected speech segments in } N_d}{\text{No. of detected speech segments } (N_d)},$$
$$R_c = \frac{\text{No. of correctly detected speech segments in } N_s}{\text{No. of speech segments } (N_s)}. \tag{9}$$

In this table, "margin" indicates the additional speech segment attached before and after the detected speech segment so that a portion of speech with low power such as consonants at the beginning or the end of words could be included. The length of each margin is A: 0 second, B: 0.5 second and C: 1.0 second. From this table, it can be seen that there is a trade-off between R_d, R_f, and R_c. When the margin increases, the probability of a low-power speech segment at the beginning or the end of words being included increases. This is indicated by the increase of R_c with the increase of the margin.



(a)

(b)

FIGURE 12: Input/separated waveform. The bars indicate the true and the detected speech events. (a) Input, (b) separated.

TABLE 2: Detection rate.

Margin	R_d(%)	R_f(%)	R_c(%)
A (0 s)	85.7	81.0	82.2
B (0.5 s)	74.5	60.3	98.8
C (1.0 s)	53.4	44.4	99.8

When a margin of 0.5 second was employed, almost all the speech segments (98.8%) were covered by the detection. On the other hand, an increase of the margin also results in more nonspeech segments being included. This can be seen in the decrease of R_d and R_f with the increase of the margin. However, once a portion of speech is not detected and is discarded, there is no chance of recovering it in the later stage of processing. Therefore, the coverage R_c is considered to be more important. In the following speech recognition test, the margin of B (0.5 second), which exhibited an R_c of close to 100% with a smaller decrease of R_d and R_f, was employed.

In the speech recognition test, two cases, in which the ranges of sound sources were $r = 1.5$ m and $r = 3.0$ m, were tested. The word error rate (WER), defined as

$$\text{WER} = 100 \times \frac{\text{Subs.} + \text{Dels.} + \text{Ins.}}{N_w}, \tag{10}$$

where Subs. is the number of substitution errors, Dels. is the number of deletion errors, Ins. is the number of insertion errors, and $N_w(= 492)$ is the total number of words in the test set, is shown in Table 3. In this table, the abbreviations used are as follows: Det.: extraction of speech segments based on the speech event detection, Sep: sound separation, and Adp.: acoustic model adaptation. For example, when only Det. is "on," the extraction of speech based on speech event detection was employed and sound separation and acoustic model adaptation were not conducted.

TABLE 3: Word error rate.

Condition	Processing			WER	
	Det.	Sep.	Adp.	$r = 1.5$ m	$r = 3.0$ m
A	On			70.3%	90.4%
B	On	On		20.3%	45.7%
C		On		74.0%	136.0%
D	On	On	On	8.9%	19.9%

From Table 3, it can be seen that by employing front-end processing, that is, speech segment extraction and sound separation, a WER around 20% was achieved when $r = 1.5$ m as indicated in condition B. On the other hand, in condition C in which speech segment extraction was not conducted, continuous speech recognition was employed. In this case, it can be seen that WER is over 100% when $r = 3.0$ m, and that many insertion errors occurred. In case D in which all three types of processing were employed, a WER of less than 10% was achieved when $r = 1.5$ m, which is considered to be within a practical range.

6. DISCUSSION

6.1. Necessity of audio and video information fusion

In this section, a brief discussion is presented on the necessity of audio and video information fusion based on the experimental results.

In Experiment 1 shown in Section 5.2, the noise from N2 included a speech signal as well as a music signal. In the conventional approach using only audio information such as VAD, the noise from N2 would also have been detected as target speech and would have caused a serious degradation in ASR performance for the real target speech from S1 and S2. As shown in Figure 11, the speech signal from N2 was not detected by the proposed method. This is the effect of combining the audio and video information.

On the other hand, in signal separation, it is possible to use only video information (human tracking results) for steering the beamformer so that it tracks humans. When using only video information, information on the presence/absence of the target speech is not available. Thus, in this case, a beamformer which can be used without information on the presence/absence of the target (e.g., the MV beamformer [1]) must be used. However, the performance of the ML beamformer is higher than that of the MV beamformer, if the information on presence/absence of the target is given (see, e.g., [13], in which the ML beamformer is denoted as the MV beamformer designed with $\mathbf{K}(\omega)$). Therefore, the information on the presence/absence of the target obtained from the audio and video information plays an important role in the signal separation.

The information on the presence/absence of the target also plays an important role in speech recognition. As indicated in Table 3, even when signal separation is employed, WER was more than 70% (condition C). WER reached around 20% ($r = 1.5$ m) only when signal separation was combined with segmentation based on information on the presence/absence of the target (condition B).

6.2. Accuracy of estimation

In the proposed method, various types of estimation such as sound localization, human tracking, and the Bayesian network are included. In this section, we discuss the accuracy of each estimation and its effect over the entire performance of the system.

In sound localization, the number of sources, N, is required as indicated in (2). In the experiment shown in Section 5, the value of N was fixed to $N = 3$. Obviously, this was not always correct. For example, when the target speech was absent in Experiment 1, $N = 2$. However, the effect of an incorrect number of sources was small in the cases tested in this paper. In the MUSIC estimator described in Section 2.1, N peaks are always generated where N is the number of sources specified in (2). Since the MUSIC estimator employs principal component analysis (PCA), N principal components (the components with large power) are extracted from the sound field and the peaks corresponding to them are generated. When the actual number of sources is smaller than N in (2), the MUSIC estimator may yield ghost peaks such as those corresponding to wall reflections. For example, the active nodes in 80°in Figure 5b are considered to be wall reflections of noise source N1. By combining the audio and video information, these ghost peaks were not detected as target speech. However, when the location of the ghost coincides with the human position, this will be detected as a target speech event. Several methods of estimating the number of sources N have been proposed by the authors of [20] and other researchers (see, e.g., [21]). However, most of these methods are based on the eigenvalue analysis of the spatial correlation $\mathbf{R}(\omega)$, and the discrimination of the real sound sources and the virtual sound sources generated by reflections is difficult in these methods. Other information such as range information (e.g., comparison of range estimated by audio information and video information) may improve the performance in these critical situations.

In human tracking using video information, the position of a human subject on a video screen is not only a function of the direction of the human but also a function of the range. Therefore, the mapping between the audio coordinate and the video coordinate shown in Figure 6 will change as a function of range. For the range tested in this paper, that is, $r = 1.5-3.0$ m, the effect of range was small and the mapping shown in Figure 6 was valid in these ranges. This is considered to be due to the fact that a relatively wide video region (32 pixels) is assigned to each node of the Bayesian network. However, in the preliminary test, it was observed that the detection sometimes failed when $r < 1.0$ m and the image of the human subject covered a large part of the video screen. In this case, the CPTs shown in Figure 6 were no longer valid. In the close range, in particular, a different approach such as use of the mouth motion detector (MMD) or a combination of MMD and the audio information would be effective [4].

In connection with the above issue, the audio and the video region assigned to the nodes of the Bayesian network should be optimized in future studies. Improvement of the basic resolution of the sound localization and human tracking is beyond the scope of this study. However, the optimization of the region for extracting the feature vector from the audio and the video information may improve detection performance in a critical situation such as when the target source and the noise source are close. In the example shown in Experiment 1, the angle between the target source S2 and the noise source N2 was around 25°. This was the most critical condition of the cases tested in this paper, and S2 was well discriminated from N2. The limitation of the proposed approach should be investigated in a more critical situation in future studies.

Finally, room dependency of the detection is discussed. The CPTs shown in Figure 6 were obtained with data obtained in the room depicted in Figure 9. The same room was used for the experiments described in Section 5. Therefore, there is no mismatch between the environment for learning and the operation. Currently, the proposed method has been realized as a real-time system and has been tested in our demonstration room, which has a similar reverberation time (0.4–0.5 second) but has a completely different configuration. A similar performance has been obtained without changing the CPTs [22]. This is because the CPTs shown in Figure 6 mainly describe the mapping between the audio and video coordinates, and this mapping will not dramatically change in rooms with similar or less reverberation time. However, for rooms with longer reverberation time, the variance of the sound localization will increase, and this will affect the CPTs. Also, under different lighting conditions, the accuracy of the human tracker may change, and this may affect the CPTs. Currently, it takes 10–15 minutes to obtain data for training CPTs. Shortening of this procedure or online-training of CPT would increase the applicability of the proposed method to a wide variety of environments. These issues should also be addressed in future studies.

7. CONCLUSION

In this paper, a method of detecting speech events in a multiple-sound-source condition based on the fusion of audio and video information was proposed. As a result of the inference of the Bayesian network which was employed to fuse the information, the time and the location of the speech events can be estimated. The results of the experiment in a real environment showed that speech events were correctly detected by the proposed method in the presence of environmental noise and interference.

The information on the detected speech events was then utilized in the robust speech interface. By combining the proposed speech event detector with sound separation, speech recognition, and acoustic model adaptation, the system achieved a WER of approximately 10–20% in a noisy environment.

As a tool for fusing the audio and video information, a Bayesian network was used. In this paper, the main function of the Bayesian network is to establish the correspondence of the audio coordinate (in angles) and the video coordinate (in pixels) with the ambiguity in the estimation being taken into account.

In a future study, more information sources should be included to achieve greater robust speech event detection. In this paper, any sound coming from humans was detected as a speech event. This can be avoided by employing additional information such as a mouth motion detector [4, 23] or VAD. In the Bayesian network, adding other information sources is realized simply by adding other input nodes. This is considered to be a great advantage of using a Bayesian network.

ACKNOWLEDGMENT

A part of this work was supported by KAKENHI 14208033.

REFERENCES

[1] D. Johnson and D. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice Hall, Englewood Cliffs, NJ, USA, 1993.

[2] J. Lim, Ed., *Speech Enhancement*, Prentice Hall, Englewood Cliffs, NJ, USA, 1983.

[3] V. Gilg, C. Beaugeant, M. Schoenle, and B. Andrassy, “Methodology for the design of a robust voice activity detectory for speech enhancement,” in *Proc. International Workshop on Acoustic Echo and Noise Control*, pp. 131–134, Kyoto, Japan, September 2003.

[4] T. Chaodhury, J. Rehg, V. Pavlovic, and A. Pentland, “Boosted learning in dynamic Bayesian networks for multimodal detection,” in *Proc. International Conference on Information Fusion*, vol. 1, pp. 550–556, Annapolis, Md, USA, July 2002.

[5] A. Christian and B. Avery, “Digital smart kiosk project,” in *ACM’s Special Interest Group on Computer-Human Interaction*, pp. 155–162, Los Angeles, Calif, USA, April 1998.

[6] F. Asano, Y. Motomura, H. Asoh, T. Yoshimura, N. Ichimura, and S. Nakamura, “Fusion of audio and video information for detecting speech events,” in *Proc. International Conference on Information Fusion*, vol. 1, pp. 386–393, Cairns, Queensland, Australia, July 2003.

[7] F. Asano, Y. Motomura, H. Asoh, et al., “Detection and separation of speech segment using audio and video information fusion,” in *Proc. European Conference on Speech Communication and Technology*, pp. 2257–2260, Geneva, Switzerland, September 2003.

[8] T. Yoshimura, F. Asano, Y. Motomura, et al., “Detection of speech events in real environments through fusion of audio and video information using Bayesian networks,” in *Proc. International Workshop on Acoustic Echo and Noise Control*, pp. 319–322, Kyoto, Japan, September 2003.

[9] R. O. Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE Transactions on Antennas and Propagation*, vol. AP-34, no. 3, pp. 276–280, 1986.

[10] C. Eveland, K. Konolige, and R. C. Bolles, “Background modeling for segmentation of video-rate stereo sequences,” in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 266–271, Santa Barbara, Calif, USA, June 1998.

[11] F. Jensen, *Bayesian Networks and Decision Graphs*, Springer, New York, NY, USA, 2001.

[12] ISIF, Ed., *Proc. 5th International Conference on Information Fusion*, ISIF, Annapolis, Md, USA, 2002.

[13] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, “Speech enhancement based on the subspace method,” *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 497–507, 2000.

[14] S. Young, G. Evermann, D. Kershaw, et al., *The HTK Book, Version 3.2*, Cambridge University Engineering Department, Cambridge, UK, 2002.

[15] K. Itou, K. Shikano, T. Kawahara, et al., “Ipa japanese dictation free software project,” in *Proc. International Conference on Language Resources and Evaluation*, pp. 1343–1349, Athens, Greece, May 2000.

[16] C. Leggetter and P. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[17] J. Gauvain and C. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[18] E. Thelen, X. Aubert, and P. Beyerlein, “Speaker adaptation in the Philips system for large vocabulary continuous speech recognition,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1035–1038, Munich, Germany, April 1997.

[19] J. Ogata and Y. Ariki, “Unsupervised acoustic model adaptation based on phoneme error minimization,” in *Proc. International Conference on Spoken Language Processing*, vol. II, pp. 1429–1432, Denver, Colo, USA, September 2002.

[20] K. Yamamoto, F. Asano, W. Rooijen, T. Yamada, and N. Kitawaki, “Estimation of the number of sound sources using support vector machines and its application to sound source separation,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. V, pp. 485–488, Hong Kong, China, April 2003.

[21] M. Wax and T. Kailath, “Detection of signals by information theoretic criteria,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 2, pp. 387–392, 1985.

[22] K. Yamamoto, N. Kitawaki, F. Asano, I. Hara, J. Ogata, M. Goto, H. Furukawa, and T. Kamashima, “Real-time implementation and evaluation of speech event detection and separation based on the fusion of audio and video information,” in *Proc. of GSPx*, Santa Clara, Calif, USA, September 2004.

[23] K. Murai and S. Nakamura, “Real time face detection for multimodal speech recognition,” in *Proc. IEEE International Conference on Multimedia and Expo*, vol. 2, pp. 373–376, Lausanne, Switzerland, August 2002.

Futoshi Asano received the B.S. degree in electrical engineering, and the M.S. and Ph.D. degrees in electrical and communication engineering from Tohoku University, Sendai, Japan, in 1986, 1988, and 1991, respectively. From 1991 to 1995, he was a Research Associate at Research Institute of Electrical Communication (RIEC) in Tohoku University. From 1993 to 1994, he was a Visiting Researcher at The Applied Research Laboratory (ARL) in Pennsylvania State University. From 1995 to 2001, he was with the Electrotechnical Laboratory, Tsukuba, Japan. Currently, he is a Group Leader in the National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan. His research interests include array signal processing, adaptive signal processing, statistical signal processing, and speech recognition.

Kiyoshi Yamamoto received the B.E. and M.E. degrees from the University of Tsukuba, Tsukuba, Japan, in 2001 and 2003, respectively. Presently, he is a candidate for the Ph.D. degree in the Graduate School of Systems and Information Engineering, University of Tsukuba. His research interests include array signal processing and sound source separation. He is a Member of the Acoustical Society of Japan (ASJ).

Isao Hara is a Senior Researcher at the National Institute of Advanced Industrial Science and Technology (AIST), Japan. His interests include humanoid robot, multiagent system, and robot control. He received his B.S., M.S., and Ph.D. in electrical engineering from Kyushu University.

Jun Ogata is a Research Scientist at the National Institute of Advanced Industrial Science and Technology, Japan. His research interests include speech recognition and understanding, and spoken dialogue system. He received his B.S., M.E., and Ph.D. from the Ryukoku University, Japan.

Takashi Yoshimura received the B.E. degree in electrical and electronic engineering and the M.E. degree in information processing from Tokyo Institute of Technology, Tokyo, Japan, in 1989, and 1991, respectively. From 1991 to 2001, he was a Researcher at Machine Understanding Division, the Electrotechnical Laboratory, Tsukuba, Japan. Since 2001, he has been with National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan, and is currently a Researcher of Media Interaction Group, Information Technology Research Institute. His main areas of research interests are voice activity detection, speech processing, and multimodal signal processing. He is a Member of the Acoustic Society of Japan and the Institute of Electronics, Information and Communication Engineers.

Yoichi Motomura is working at the National Institute of Advanced Industrial Science and Technology Digital Human Research Center as a Senior Research Scientist. He joined the Electrotechnical Laboratory in 1993. He was working on statistical learning algorithms, probabilistic models, and their applications to intelligent robots. He developed Bayesian network software that constructs models from data and executes approximate probabilistic reasoning. In 1999, he was invited to the University of Amsterdam, where he worked for robot localization using probabilistic prediction methods. His current research interests include applying Bayesian networks for user modeling, modeling of human behaviors in living rooms for accident prediction, adaptive and interactive systems in cars, and probabilistic reasoning algorithms.

Naoyuki Ichimura received the B.E. degree in communication engineering in 1989 and the M.S. and Ph.D. degrees in electronic engineering in 1991 and 1996, all from the University of Electro-Communications, Tokyo, Japan. He is a Senior Research Scientist at the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan. He was a Visiting Scholar of the Computer Science Department at Columbia University from 2002 to 2004. His current research interests are computer vision, computer graphics, and time series analysis. He is a Member of IEEE, the Institute of Electronics, Information and Communication Engineers (IEICE), the Society of Instrument and Control Engineers (SICE), and the Information Processing Society of Japan (IPSJ).

Hideki Asoh received his B.Eng. degree in mathematical engineering and M.Eng. degree in information engineering from the University of Tokyo, in 1981 and 1983, respectively. In April, 1983, he joined the Electrotechnical Laboratory as a Researcher. From 1990 to 1991, he stayed at the German National Research Center for Information Technology as a Visiting Research Scientist. Since April 2001, he is a Senior Research Scientist of Information Technology Research Institute, National Institute of Advanced Industrial Science and Technology, Japan. His research interests are in constructing intelligent systems which learn through man-machine interaction.