State Estimation of Meetings by Information Fusion using Bayesian Network

Michiaki KATOH[1], Kiyoshi YAMAMOTO[1], Jun OGATA[2], Takashi YOSHIMURA[2], Futoshi ASANO[2], Hideki ASOH[2], and Nobuhiko KITAWAKI[1]

[1] University of Tsukuba, Japan

[2] AIST, Japan

michy@mmlab.cs.tsukuba.ac.jp

f.asano@aist.go.jp

10.21437/Interspeech.2005-81

Abstract

In this paper, a method of structuring the multi-media recording of a small-sized meeting based on various information such as sound source localization, multiple-talk detection, and the detection of non-speech sound events, is proposed. The information from these detectors is fused by a Bayesian network to estimate the state of the meeting. Based on the estimated state, the recording of the meeting is structured using a XML-based description language and is visualized by a browser.

1. Introduction

Audio and video recordings of meetings include various information. However, extracting useful information from the recordings by surveying all of the recorded data requires considerable time and is inefficient. Recently, structuring of meetings has been focused on by several research groups. In [1,2], the speaker-turn features were extracted from microphone array inputs and were utilized in an HMM and a dynamic Bayesian network approach.

In this paper, various features are extracted from single/multiple microphone inputs to detect various events in the meeting. The most important events for structuring the meetings are the changes of speaker turns. Speaker turns are identified by estimating the direction of sound sources using a microphone array as in the previous approach. On the other hand, there are some small events such as multiple talk, speech pauses, noises and laughter, which sometimes carry important information. If listening or speech recognition is conducted for the recordings, however, these small events may also be interference. Therefore, the detection of these small events is also important.

The events detected by various detectors are then fed into the Bayesian network for fusing the information and estimating the state of the meeting. The estimated state is utilized for structuring the recordings of meetings. A language for describing the structure termed MADL (Meeting Archiver Description Language) and the browser which visualizes the structure is also briefly introduced.

2. Detection of Events

2.1. Events in Meetings

The events detected in this paper are listed in Table 2.1. The turn-takings of the speakers are termed Major events in this paper. On the other hand, small events such as MT, BN, IN and CO are termed Minor events.

Table 2.1 Events focused on in this paper.

Symbols	Events
$S_1,\cdots S_N$	Speech events by the speakers $S_1,\cdots S_N$
MT	Simultaneous speech by several speakers
BN	Background noise (pause between talks)
IN	Impulsive noise (short-duration noise)
CO	Sound of coughing

2.2. Speaker Identification Using Sound Source Localization

For identifying the speaker, the direction of a speech signal is estimated using the microphone array input. The MUSIC method extended for the broadband signal using eigenvalue weighting [3] is used for the estimation of the speaker direction. The advantage of this method over a simple delay-and-sum type estimation is a high resolution in direction finding, which makes it possible to distinguish adjacent talkers. Figure 2.1(a) shows an example of a spatial spectrogram estimated by the above method. From this, peaks are detected as shown in (b). Figure 2.2 shows a histogram of the peaks. From this, clusters of the peaks corresponding to the speakers participated in this meeting can be seen. Using this histogram, the region of direction corresponding to each speaker is determined.

2.3. Multiple talk (MT) Detection Using Support Vector Machine

The eigenvalue distribution of the spatial correlation matrix, $\mathbf{R} = E\left[\mathbf{x}(\omega,t)\mathbf{x}^H(\omega,t)\right]$, where $\mathbf{x}(\omega,t)$ consists of the short term Fourier transforms of the microphone array inputs, reflects the information on the number of the active sound sources in the field. Figure 2.3 shows an example of the eigenvalue distributions for cases of single and double speakers. The number of dominant eigenvalues roughly corresponds to the number of active sources. In this paper, the eigenvalue distributions are classified by support vector machines (SVM) into cases of single (ST) and multiple (MT) speakers [4].

2.4. Backgound Noise (BN) and Impulsive Noise (IN) Detection Using Kurtosis

Signals in a meeting situation can roughly be classified by the kurtosis measure, which reflects their time-domain

distribution [5]. The normalized kurtosis is given by

$$Kt = \frac{E[x^4(t)]}{E[x^2(t)]^2} - 3 \qquad (2.1)$$

where $x(t)$ denotes the time domain signal of a single microphone input. By using the normalized kurtosis, the input signal can be classified as follows:

- Background noise (BN) such as the sound of air-conditioner or fans of computer: Sub-Gaussian distribution (negative kurtosis)
- Speech (SP): Super-Gaussian distribution (relatively small positive kurtosis)
- Impulsive noise (IN) such as the sound of tapping on desks: Super-Gaussian distribution (relatively large positive kurtosis)

Figure 2.4 shows an example of the power and the kurtosis for impulsive noise. By examining the value of the normalized kurtosis, the signals are roughly classified into the three classes shown in Table 2.2. The threshold for the classification was determined so that the classified data matches the label given by the human labeler.

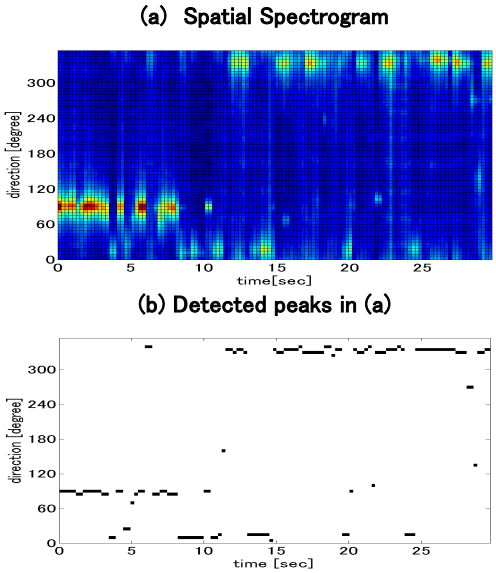

Figure 2.1 Sound localization for speaker identification.

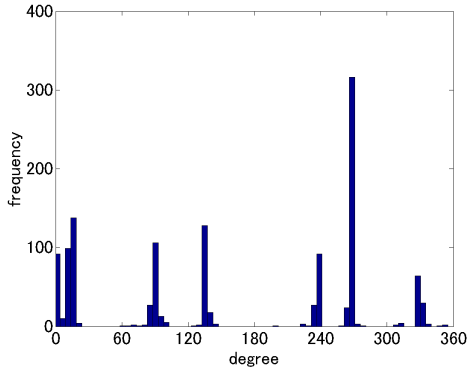

Figure 2.2 Histogram of peak location.

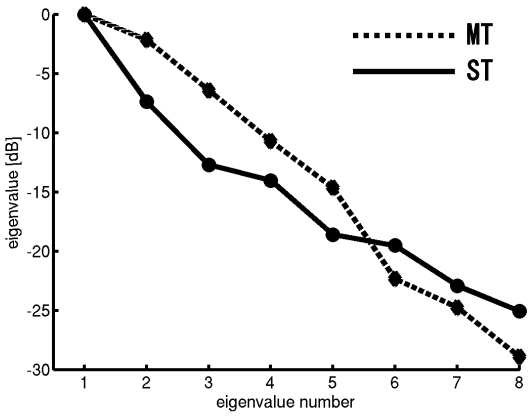

Figure 2.3 Sample of eigenvalue distribution.

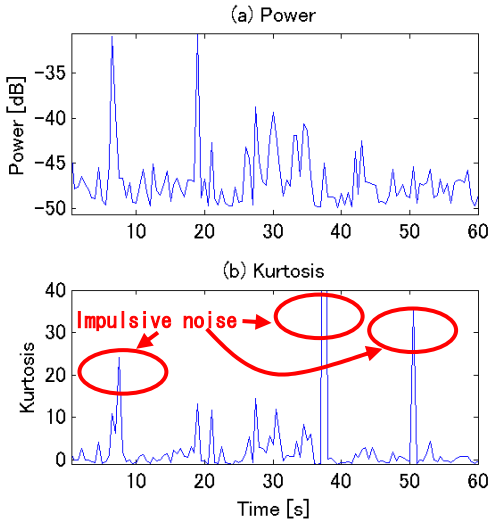

Figure 2.4 Power and kurtosis for impulsive noise

Table 2.2 Class of signals and the corresponding region of kurtosis

Class	Kurtosis value
Background noise (BN)	$Kt < -0.9$
Speech (SP)	$-0.9 \le Kt \le 15$
Impulsive noise (IN)	$Kt > 15$

2.5. Detection of Coughing Sound (CO) by HMM

The sound of coughing is somewhat impulsive. However, the corresponding kurtosis value is sometimes close to that of a speech signal, and thus the sound of coughing cannot always be detected by using kurtosis. In this paper, the sound of coughing is modeled using HMM and is detected by a phone recognizer which has 43 context-independent Japanese phone models with the cough model. The cough model was trained using 50 samples uttered by two subjects, and the other phone models were trained by using a database of Japanese read speech.

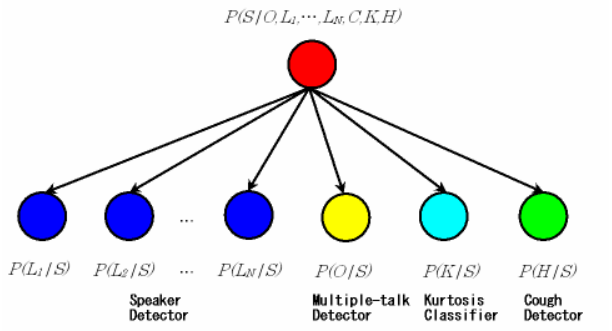

Figure 3.1 Bayesian network used for the information fusion.

3. Information Fusion using Bayesian Network

In the previous section, we introduced several detectors for detecting Major and Minor events. These events may occur simultaneously. However, Minor events treated in this paper can be a strong interference for Major events and may affect its detection. In this paper, therefore, the detection of the Minor events has a higher priority. Control of the priority is done by fusing the information from different detectors by a Bayesian network [2].

Figure 3.1 shows the topology of the Bayesian network used in this paper. The input and the output nodes of the network can take the values shown in Table 3.1. The output node S indicates the state of the meeting. The symbol S_i represents the speech event of the ith speaker (Major event). The symbols {BN, IN, MT, CO} represent Minor events. The input nodes L_i represents the state of the ith speaker detector. When the results of sound localization have a peak in the region corresponding to the ith speaker, the speaker detector L_i becomes "On." The input node O represents the multiple-talk detector and takes the value of Single (single-talk) or Multi (multiple-talk). The node K denotes the kurtosis classifier and takes the value of BN (background noise), SP (speech), and IN (Impulsive noise) according to the kurtosis value shown in Table 2.2. The node H denotes the cough detector using HMM and becomes "On" when a cough sound is detected.

Table 3.1 Node and the value of state

Node	Value
S (State of Meeting)	$S_1/\cdots/S_N$ / BN/IN/MT/CO
$L_1,\cdots,L_N$ (Speaker Detector)	On/Off
O (Multiple-talk Detector)	Single/Multi
K (Kurtosis Classifier)	BN/SP/IN
H (Cough Detector)	On/Off

Assuming that the events detected by each detector are independent, the conditional probability $P(S \mid L_1,\cdots,L_N,O,K,H)$ that represents the state of the meeting is calculated using the network shown in Figure 3.1 as follows:

$$P(S \mid L_1,\cdots,L_N,O,K,H) = \sum_{i=1}^{N} P(S \mid L_i)P(S \mid O)P(S \mid K)P(S \mid H) \quad (3.1)$$

The value of the conditional probabilities on the right-hand side, i.e., $P(S \mid L_i)$, $P(S \mid O)$, $P(S \mid K)$ and $P(S \mid H)$, are determined based on the state of the detectors, L_i, O, K and H, and the conditional probability table (CPT). In this paper, the CPT was given by the user so that the intended priority control could be obtained.

4. Experiments

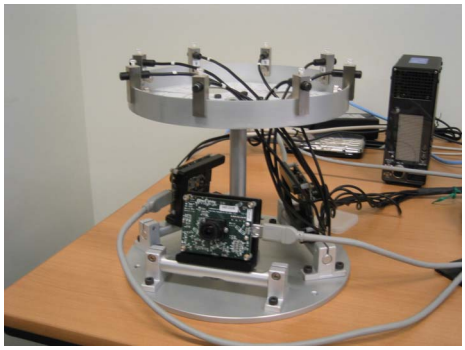

Figure 4.1 Microphone array and camera array used for the recording.

4.1. Condition

A meeting used in a Japanese market research termed "Group Interview" was recorded and used for testing the proposed method. In this meeting, one professional interviewer and five interviewees (university students in this paper) participated. The interviewer asked questions such as "What types of cellular phones are you using?," and the interviewee answered the questions in a discussion manner. Thus, speaker turns often changed, an ideal material for structuring.

The meeting was conducted in a middle-sized meeting room with a reverberation time of 0.5 s. The six participants were sat around a table. The microphone array and the camera array shown in Figure 4.1 were located in the middle of the table. The microphone array consisted of eight microphones in a circular shape with a diameter of 20 cm. The camera array consists of three cameras with a VGA resolution. The distance from the center of the array to the participants was approximately 1-1.5 m.

4.2. Results

Figure 4.2 shows an example of the state estimated by the proposed method. The estimation of the state was conducted every 0.5 s. By comparison with the state labeled by a human labeler, it can be seen that the outline of the meeting is well structured. Table 4.1 presents a confusion matrix. When the estimated state in a frame (with a duration of 0.5 s) was in accordance with that provided by the human labeler, this frame was judged to give a correct estimation. From this table, it can be seen that the estimation rate for Major events was high. On the other hand, some of Minor events such as MT and CO were not detected well. The performance of these detectors should be improved in the future.

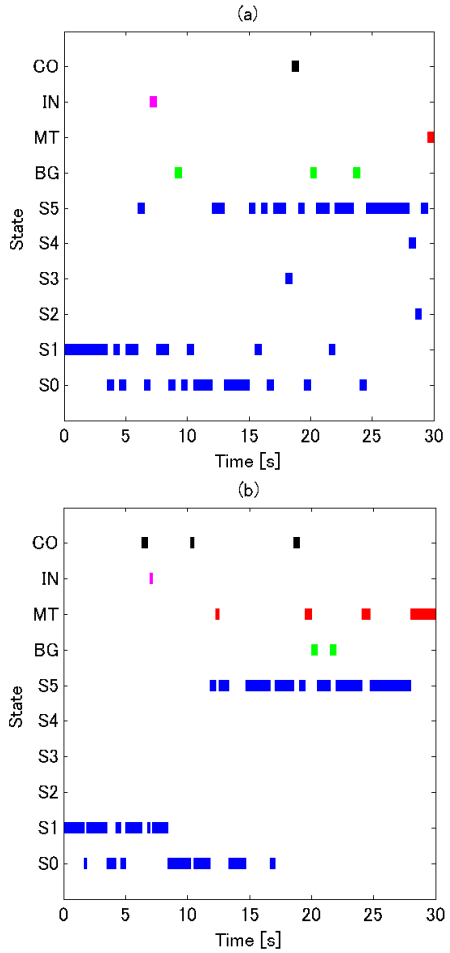

Figure 4.2 (a) State estimated by the proposed method; (b) State provided by a human labeler.

5. Description Language and Browser

Based on the estimated state above, the structure of the meeting is then described by the language termed MADL. MADL is an XML-based language designed for replaying audio-visual recordings based on the estimated structure. Figure 5.1 is a snapshot of the browser replaying the recordings. The top panels show videos recorded by the camera array. The middle panel shows the estimated structure (speaker turns.) The bottom panel shows the transcription obtained by automatic speech recognition.

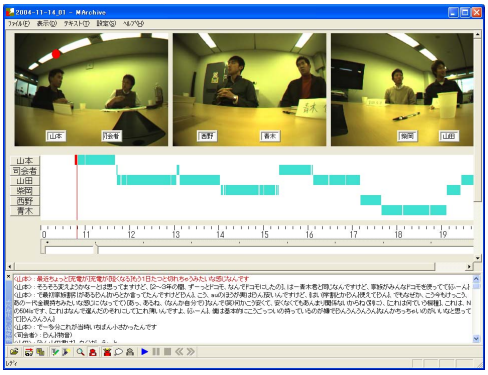

Figure 5.1 Snapshot of the browser.

6. Conclusions

In this paper, Major and Minor events in a small-sized meetings were detected by using various kinds of detectors. The state of the meeting was then estimated by fusing the information of the detected events with a Bayesian network. In the next stage, the information of Minor events should be actively utilized for sound separation and speech recognition to improve the performance of the transcription.

7. References

[1] Ajmera, J., Lathoud, G., McCowan, I., "Clustering and segmenting speakers and their locations in meetings", *Proc. ICASSP2004*, Vol.I, pp.605-608, Mar.2004.

[2] Dielmann, A. and Renals, S., "Dynamic Bayesian networks for Meeting structuring", *Proc. ICASSP2004*, Vol.V, pp.629-632, Mar.2004.

[3] Asano F. et al, "Detection and separation of speech event using audio and video information fusion and its application to robust speech interface", *EURASIP J. Applied Signal Processing 2004:11*, pp. 1727-1738, 2004.

[4] Yamamoto, K., et al., "Estimation of the number of sound sources using support vector Machines and its application to sound source separation", *Proc. ICASSP2003*, Vol.V, pp.485-488, Apr.2003.

[5] Yoshimura T., et al, "Investigation of Voice/Sound Activity Classifier using Distribution Models of Fourth-Order Statistics", IPSJ technical report, Vol.2004 No.74 (in Japanese).

Table 4.1 Confusion matrix of the estimated state (%).

	S1	S2	S3	S4	S5	S6	MT	BN	IN	CO
S1	**79.3**	2.5	2.6	1.4	5.1	2.2	2.3	3.7	0.8	0.0
S2	2.6	**87.0**	0.0	0.9	1.2	0.7	0.4	2.9	4.1	0.2
S3	0.0	0.9	**98.2**	0.0	0.0	0.0	0.0	0.9	0.0	0.0
S4	2.1	1.8	1.4	**87.6**	0.0	1.0	1.4	4.7	0.0	0.0
S5	0.4	0.2	0.2	0.0	**93.4**	0.0	3.0	0.6	2.2	0.0
S6	3.0	6.1	0.0	3.6	0.7	**80.4**	0.0	3.3	2.6	0.0
MT	8.6	12.7	9.1	7.0	27.3	5.6	**24.6**	0.1	3.3	1.6
BG	5.9	9.0	2.1	0.2	1.8	1.4	0.0	**79.2**	0.4	0.0
IN	4.6	2.4	0.8	1.3	2.0	15.6	4.0	0.0	**69.3**	0.0
CO	16.9	17.7	0.0	0.0	4.3	19.4	5.9	0.0	0.0	**35.7**