Fusion of Audio and Video Information for Detecting Speech Events

Futoshi Asano, Yoichi Motomura, Hideki Asoh, Takashi Yoshimura and Naoyuki Ichimura
AIST
Tsukuba, Japan
f.asano/y.motomura/h.asoh@aist.go.jp
yoshimur/nic@ni.aist.go.jp

Satoshi Nakamura
ATR
Kyoto, Japan
satoshi.nakamura@atr.co.jp

Abstract – In this paper, a method of detecting speech events in a multiple-sound-source condition using sound and vision information is proposed. Detection of speech event is an important issue for automatic speech recognition operated in a real environment. Furthermore, as stated in this paper, the performance of sound source separation using adaptive beamforming is greatly improved by knowing when and where the target speech event occurs. For this purpose, 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, the information on time and location of speech events can be known in a multiple-sound-source condition. Results of an off-line experiment in a real environment with TV and music interference are shown.

Keywords: Sound localization, sound separation, vision tracking, Bayesian network, Speech event detection.

1 Introduction

When using an automatic speech recognizer (ASR) in a real environment with noise, speech-enhancement/noise-reduction system is indispensable. For speech enhancement, various kinds of methods, including single channel methods such as classical spectral subtraction and Wiener filtering [1] and multi-channel methods such as adaptive beamforming [2, 3] and blind source separation [4, 5], has been proposed.

The authors have proposed a method of sound source separation based on sound localization using subspace method and maximum likelihood adaptive beamforming [6] and have applied this system to ASR. This method is considered to have a higher noise reduction capability than the other single-channel and multi-channel methods in the environments such as we have tested, namely, offices and homes. However, this high performance is achieved only when the information of the speech event (when and where the target speaker is speaking) is provided, as stated in Section 2. The detection of speech events is also an important issue for ASR when it is used in a real environment.

When the noise are non-speech signals, a voice activity detector (VAD) can be used as a target detector. In environments such as offices and homes, however, not only the target but also the interference source such as a TV or a radio can be speech signals. In these cases, the detection of the target speech cannot be accomplished only by using the sound information, and fusion with the information from other modalities such as vision is necessary.

Chaudhury et al. [7] proposed a speech event detector using audio and video information. In this paper, an environment in which a dialog between multiple speakers and a Smart Kiosk terminal is considered. The speech events which are addressed only to the terminal by the speakers are detected by this system and are recognized. In this system, it is assumed that only a single sound event occurs at a single moment. Also, information on the location in the audio and video information was not utilized. In this paper, a method of detecting a speech event under the circumstance in which multiple sound events occur at the same time (speech event by a human is assumed to be single at a single moment) is proposed. The location and time information obtained from sound and vision are fused by a Bayesian network so that the time and location information of the speech event can be known. This information is used for the speech enhancement system. The information on the time of the speech event detected by this method is also useful for ASR.
2 Sound Source Separation

2.1 Signal Model

The Fourier transform of the input signal to the microphone array can be modeled as

\[ x(\omega, t) = [X_1(\omega, t), \cdots, X_M(\omega, t)]^T \]

\[ = \sum_{n=1}^{N} a_n(\omega)S_n(\omega, t) + n(\omega, t), \quad (1) \]

where \( M \) and \( N \) denote the number of microphones and the number of sound sources, respectively. The symbol \( \cdot^T \) denotes the transpose. The symbol \( X_m(\omega, t) \) denotes the short-time Fourier transform of the input signal to the \( m \)th microphone. The vector \( a_n(\omega) \) is termed the location vector and is defined as

\[ a_n(\omega) = [A_{1,n}(\omega), \cdots, A_{M,n}(\omega)]^T, \quad (2) \]

where \( A_{m,n}(\omega) \) denotes the transfer function of the direct path from the \( n \)th source to the \( m \)th microphone. The symbol \( S_n(\omega, t) \) denotes the spectrum of the \( n \)th sound source. The vector \( n(\omega, t) = [N_1(\omega, t), \cdots, N_M(\omega, t)]^T \) denotes the noise vector, in which \( N_m(\omega, t) \) denotes the spectrum of the ambient noise observed at the \( m \)th microphone.

2.2 Sound Separation by Adaptive Beamforming

In this paper, a simple adaptive beamformer using maximum likelihood (ML) estimation is employed (e.g., [3]). Assuming that the \( n \)th sound source is the target speaker, the source spectrum is estimated by the ML adaptive beamformer as

\[ y(\omega, t) = w^H(\omega)x(\omega, t), \quad (3) \]

where the beamformer coefficient vector is defined as

\[ w(\omega) = \frac{K^{-1}(\omega)a_n(\omega)}{a_n^H(\omega)K^{-1}(\omega)a_n(\omega)}. \quad (4) \]

The symbol \( \cdot^H \) denotes the Hermitian transpose. The matrix \( K(\omega) \) is the noise correlation defined as

\[ K(\omega) = E[\tilde{x}(\omega, t)\tilde{x}^H(\omega, t)], \quad (5) \]

where \( \tilde{x}(\omega, t) \) is the input signal when the target source is absent (mixture of interference signals):

\[ \tilde{x}(\omega, t) = \sum_{n \neq n} a_n(\omega)S_n(\omega, t) + n(\omega, t). \quad (6) \]

Figure 1 shows a block diagram of the ML adaptive beamformer. In this system, the input signal is segmented into a block with a duration of \( T \). In this block, the absence/presence of the target signal is judged by the “Target Detector.” When the target is present, the location vector corresponding to the target source \( a_{\hat{n}}(\omega) \) is estimated by the sound localization. When the target is absent, the noise correlation \( K(\omega) \) is estimated. Therefore, the performance of the target detector is a key factor of the system. Using the latest estimation of \( a_{\hat{n}}(\omega) \) and \( K(\omega) \), the beamformer coefficient vector \( w(\omega) \) is calculated at every time block using (4). The coefficient vector calculated at every frequency, is then transformed into the time domain by inverse FFT, and, finally, the microphone array input in this block is filtered by these beamformer coefficients.

Figure 2: An example of the audio information (spatial spectrum) and the video information (human tracking).
3 Basic Techniques for Speaker Tracking

In this section, basic techniques for speaker tracking used in this paper, i.e., the sound localization using a microphone array and the human tracking by vision, are described.

3.1 Sound Localization

For the sound localization, a method termed MUSIC [8] that is extended to the broadband signal with eigenvalue weighting (see Appendix for the detail) is used. In this method, the spatial correlation is estimated as

\[ R(\omega) = E[x(\omega, t)x^H(\omega, t)]. \]

Then, the eigenvalue decomposition is performed as

\[ R(\omega) = E(\omega)\Lambda(\omega)E^{-1}(\omega). \]

Here, \( E(\omega) = [e_1(\omega), \ldots, e_M(\omega)] \) denotes the eigenvector matrix while \( \Lambda(\omega) = \text{diag}(\lambda_1(\omega), \ldots, \lambda_M(\omega)) \) denotes the eigenvalue matrix. The eigenvalues \( \{\lambda_m(\omega)\} \) and eigenvectors \( \{e_m(\omega)\} \) are assumed to be sorted in the descending order of the eigenvalues. Using the eigenvectors corresponding to the smallest \( M-N \) eigenvalues, the MUSIC spatial spectrum estimator is defined as

\[ P(\theta, \omega) = \frac{|a(\theta, \omega)|^2}{\sum_{m=N+1}^{M} |e_m^H a(\theta, \omega)|^2}. \]

The symbol \( a(\theta, \omega) \) denotes the location vector of the virtual source in the arbitrary direction \( \theta \). This spatial spectrum is estimated at every frequency bin \( \omega \). To estimate the final spatial spectrum for the broadband input, (9) is averaged over the frequency of interest as

\[ \bar{P}(\theta) = \frac{1}{\bar{\omega}_h - \bar{\omega}_l} \int_{\bar{\omega}_l}^{\bar{\omega}_h} \lambda P(\theta, \omega) d\omega, \]

where \( \bar{\lambda} \) is the weight defined in Appendix. The range \( [\omega_l, \omega_h] \) denotes the frequency range of interest.

Figure 4: Audio information as a function of time. In (b), the marks ‘•’, ‘×’ and ‘·’ correspond to the 1st, 2nd and 3rd largest peaks, respectively.

3.2 Human Tracking by Vision

There are many methods to track humans using video information. As a human tracker used in the framework proposed in this paper, 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 (e.g., [9]). This range image is obtained by a trinocular stereo camera (Pointgray Research, Diclops). Figure 3 shows the process of the background subtraction. The main panel shows an example of the subtracted image (difference between the current input image and the pre-recorded background image). The upper and the left panel show histograms of the difference in a vertical and a horizontal slice, respectively. Regions with more than a certain area in which the value of the histogram is above a certain threshold are recognized as foreground objects (humans).

Figure 2(a) shows an example of the spatial spectrum estimated by (10). From the peaks of this spectrum, the location (direction) of the sound sources can be estimated.
Figure 12: The detected and the true speech events.

Figure 12(a) shows the results of inference, which show the detected speech events in the time-direction plane. The input measurement vector corresponding to this inference result is shown in Fig. 6. Comparing this with Fig. 12(b), which shows the true speech events, it can be seen that the speech events were detected fairly correctly.

Using the information of the detected speech events, the ML beamformer was updated at every time-block (every 1 s) and the microphone-array input was processed. The input/output waveform is shown in Fig. 13. In Fig. 13(b), the bars indicating the detected and the true speech events are also shown. From these data, it can be seen that the speech signal which was almost buried by the interference was recovered by the ML beamforming.

7 Conclusion

In this paper, a method of fusing the information of sound localization using a microphone array (audio information) and that of human tracking by stereo vision (video information) has been proposed. The purpose of fusing these two types of information is to detect speech events in the time and spatial domain (when and where the speech event occurs). The detection results were then fed to the sound separation system, in which speech signal from human speakers were separated from multiple interferences.

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 pixes) with the ambiguity in the estimation being taken into account.

In a future study, we plan to include more information sources. One of the candidates is a mouth motion detector developed by one of the authors [12]. By using this, the timing of speech events can also be known. By employing multiple information sources, more robust detection of speech events is expected to be realized. In the Bayesian networks, adding another in-
Figure 5: Video information as a function of time. The lines correspond to the center position of humans. The squares correspond to the bins in which humans are detected.

4 Audio-Video Information Fusion

4.1 Extracting Sound and Vision Feature

The information from the sound localization is a spatial spectrum as depicted in Fig. 2(a). In this spectrum, the region for the observation is divided into $N_a$ bins. Then, for each bin, whether a peak exists in this bin or not is detected. Figure 4(a) shows a running spatial spectrum, each time slice of which corresponds to the spatial spectrum such as depicted in Fig. 2(a). Figure 4(b) shows the peaks detected in Fig. 4(a). This peak information is then converted to a state of 0/1 ("1" corresponds to peak being detected.) Using this, the measurement vector $a(t) = \{A_1(t), \ldots, A_{N_a}(t)\}$ is formed, in which $A_n(t)$ denotes the state at the $n$th bin at the time $t$.

The information from the vision tracker is the center position of human(s) as depicted in Fig. 5. In a manner similar to that of sound, the region of observation is divided into $N_v$ bins, and then whether a human exists in each bin or not is confirmed. The squares in Fig. 5 are the detected bins. From these detected bins, the measurement vector $v(t) = \{V_1(t), \ldots, V_{N_v}(t)\}$ is formed.

Figure 6 shows the data vectors $a(t)$ and $v(t)$. The vertical slice in this figure corresponds to the data vector at time $t$.

4.2 Information Fusion using a Bayesian Network

By fusing the audio and video information, speech events can be detected.

Let $S$ denote the state of the speech event and be in the following values: $S = \{S_1, \ldots, S_{N_s}, NoEvent\}$.

The state $\{S_1, \ldots, S_{N_s}\}$ corresponds to the speaker's position (angle) such as $\{S_1, \ldots, S_{N_s}\} = \{-30^\circ, \ldots, +30^\circ\}$. When $S = -30^\circ$, the speaker is located in the direction of $30^\circ$ and is speaking. When $S = NoEvent$, there are no speech events.

For estimating $S$ from the audio and video information, a Bayesian network is used. The Bayesian network is a way of modeling a joint probability distribution of multiple random variables [10, 11], and is considered as a powerful tools for information fusion. Figure 7 shows the Bayesian network used for fusing audio and video information. The nodes for “Audio” and “Video” correspond to the elements in the audio and video measurement vectors $a(t)$ and $v(t)$, respectively.

In this paper, it is assumed that the value of all $A_i$ and $V_j$ are conditionally independent when the value of $S$ is given. Then the conditional probability distribution $P(S|A_1, \ldots, A_{N_a}, V_1, \ldots, V_{N_v})$ can be factored into the product of local conditional probabilities $P(A_i|S)$ and $P(V_j|S)$:

$$P(S|A_1, \ldots, A_{N_a}, V_1, \ldots, V_{N_v}) = P(S) \prod_{n=1}^{N_a} P(A_n|S) \prod_{n=1}^{N_v} P(V_n|S)/Z \quad (11)$$
Figure 7: Bayesian network for fusing audio and video information.

Table 1: An example of audio and video CPTs.

(a) Audio CPT

<table>
<thead>
<tr>
<th>S</th>
<th>-30</th>
<th>-20</th>
<th>-10</th>
<th>0</th>
<th>+10</th>
<th>+20</th>
<th>+30</th>
<th>N/E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.0</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>

(b) Video CPT

<table>
<thead>
<tr>
<th>S</th>
<th>-30</th>
<th>-20</th>
<th>-10</th>
<th>0</th>
<th>+10</th>
<th>+20</th>
<th>+30</th>
<th>N/E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>0.9</td>
<td></td>
</tr>
</tbody>
</table>

where

$$Z = \int_S P(S) \prod_{n=1}^{N_a} P(A_n|S) \prod_{n=1}^{N_v} P(V_n|S) dS$$  \hspace{1cm} (12)$$

The conditional probabilities, \(P(A_i|S)\) and \(P(V_j|S)\), can be estimated from training samples. For the training samples, the value of \(S\) is given as a “label” for each measurement vector. An example of the conditional probability tables (CPT) representing \(P(A_i|S)\) and \(P(V_j|S)\) are shown in Table 1.

In a operation phase, the measurement vectors for audio and video are obtained at every time block as evidence. Using the evidence and the CPTs obtained above, the conditional probability (11) is calculated and the most probable state of \(S\), given the values of audio and video nodes, can be obtained.

Figure 8 shows an example of the state of the Bayesian network used in this paper. For audio nodes, the 17 nodes correspond to the 17 regions (-90° - +90° , every 10 °). For video nodes, the 10 nodes correspond to the 10 regions (1 - 480 pixels, every 48 pixels). The entire video region approximately corresponds to the audio angle of ±35°. Based on the CPT and evidences, the probability of each state for the “Speaker” node is determined.

5 Entire System

Figure 9 shows a block diagram of the entire system. In the Sound Localization module, the spatial spectrum as depicted in Fig. 2(a) is estimated at every time block with a duration of \(T\). Also, in the Vision Tracking module, the tracking results as depicted in Fig. 2 are obtained using the frame rate of the vision system. To synchronize the audio and video data, a histogram is taken of the video tracking results during the time block and the human position in this time block is estimated. For example, when \(T = 1\) s for the audio system and the frame rate in the video system is 10 frames/s, a histogram for 10 frames is taken to determine the human position in this time block. Using these audio and video data, measurement vectors \(a(t)\) and \(v(t)\) are formed.

These measurement vectors are then fed to the Bayesian Network module. Using these measurement vectors and CPTs, the speech event is detected. From this detection, the location and the time of the speech event can be determined.

The information of the speech event functions as a target detector in Fig. 1. When the speech event occurs, the location vector for the target \(a_n(\omega)\) is updated in the Sound Separation module. On the other hand, when there is no speech event, the noise correlation \(K(\omega)\) is updated. Using the latest updates \(a_n(\omega)\) and \(K(\omega)\), the filter coefficient vector \(w(\omega)\) is updated at every time block and the input signal to the microphone array in this time block is processed. The information of the speech event can also be used in the Speech Recognizer module.
6 Experiment

The experiment was conducted in a medium-sized meeting room with a reverberation time of 0.5 s. A scene of the experiments is shown in Fig. 10. The parameters of the sound localization and the human tracking by vision are listed in Table 2.

For obtaining the training data, a single human speaker spoke in the directions of $-30^\circ$ to $+30^\circ$ (the range covered by the camera) at every $5^\circ$ . In every direction, speaker spoke intermittently for 30 s. As a label (the state of $S$), the physical direction of the speaker and the time of speaking was supplied to the training samples. Figure 11(a) and (b) shows the audio and video CPTs obtained by learning using the training samples. The examples of the CPTs in Table 1 correspond to the first (top) row of this figure (only the case for the state “1” is shown in this figure.)

In the detection experiment, two human speakers were located in the direction of $0^\circ$ and $-20^\circ$ , respectively. The first speaker spoke two short sentences, and then, after a short pause, the second speaker spoke two sentences. Figure 12(b) shows the time and the direction of these speech events. As interference sources, a TV (human voice $+$ music) and a loudspeaker (music) were located at $+30^\circ$ and $-90^\circ$ , respectively, as shown in Fig. 10.
formation sources is realized simply by adding another input node. This is considered to be an advantage of using Bayesian networks.

**Appendix: Eigenvalue Weighting of the Spatial Spectrum**

As stated in Section 3, sound localization using the subspace method is performed in each frequency bin. To obtain the final localization results for the broadband signal, a method to unite this information over the frequency of interest is required. In the coherent subspace method [13], the spatial correlation $R(\omega)$ is transformed and averaged using the coherence between the frequencies for this purpose. This method employs two-stage estimation, i.e., an initial rough estimation using the delay-and-sum beamforming and a second more precise estimation using the subspace method. The performance of the second estimation is highly dependent on the first stage estimation. However, the performance of the first stage is not always sufficient, especially for an array which does not perfectly satisfy Nyquist’s sampling theorem.

In this appendix, averaging of the MUSIC spectrum over the frequency of interest with eigenvalue weighting is proposed. The eigenvalue weight in (10) is defined as

$$\bar{\lambda} = \sum_{m=1}^{N} \lambda_m(\omega)$$  \hspace{1cm} (13)

where $\{\lambda_1(\omega), \ldots, \lambda_N(\omega)\}$ are the largest $N$ eigenvalues. These $N$ largest eigenvalues correspond to the energy of the signal in the signal subspace (the energy of the direct sound from the $N$ sources.) By employing the sum of the energy in the signal subspace as a weight, the estimates in the frequencies which contain the large energy of the direct sound (the frequencies in which the SNR is high) are enhanced.

**References**


