Information Fusion of Microphone Array and Camera Array for Robust Speech Interface

Futoshi Asano\textsuperscript{1}, Kiyoshi Yamamoto\textsuperscript{2}, Hideki Asoh\textsuperscript{1}, Takashi Yoshimura\textsuperscript{1}
Isao Hara\textsuperscript{1}, Yoichi Motomura\textsuperscript{1}, Naoyuki Ichimura\textsuperscript{1}, Jun Ogata\textsuperscript{1}

\textsuperscript{1}AIST, Japan
\textsuperscript{2}Tsukuba University, Japan
f.asano@aist.go.jp

Abstract

In this paper, a method of detecting speech events based on the fusion of audio and video information is introduced. Moreover, this speech event detector has been integrated into a robust speech interface. The performance of this system is evaluated using a word recognition task in a real environment with interference. Finally, a real-time implementation of the system using hardware developed by the authors is also briefly introduced.

1. Introduction

Detection of speech events is an important issue for speech recognition and speech enhancement in a real environment. The authors have developed a method of detecting speech events based on the fusion of audio information (sound localization using a microphone array) and video information (human tracking using a stereo camera) [1, 2]. This speech event detector has been integrated into a robust speech interface using automatic speech recognition (ASR) [3]. In this interface, an adaptive beamformer and acoustic model adaptation of ASR are also employed to improve the robustness of the system against environmental noise. Performance evaluation using a word recognition task in a real environment with noise is herein reported. Finally, a real-time implementation of the system is also briefly introduced. In this system, hardware for array processing developed by the authors, RASP-1, is used.

2. Speech event detection

2.1. Sound localization

For sound localization, the MUSIC method [4] extended to a broadband signal with eigenvalue weighting [1] is used. By using this, the spatial spectrum depicted in the lower left panel of Fig. 1 is obtained. From the peaks of the spatial spectrum, the location (direction) and time of the audio events (emission of sound from sound sources) can be estimated.

2.2. Human tracking by vision

Humans in a scene are tracked by background subtraction (e.g., [5]) based on the range image obtained by a stereo camera as depicted in the lower right panel of Fig. 1. By using this, the position (pixels) and time of the video events (existence of humans in video information) can be estimated.

2.3. Information fusion

By combining the information of the audio and video events detected by sound localization and human tracking by vision, the co-occurrence of the audio and video events in a certain area of the observed space can also be estimated. This co-occurrence of the audio and video events is detected as a “speech event.”

As a tool for combining the multi-modal information, the Bayesian network (e.g., [6]) is used. 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 [7]. The upper panel of Fig. 1 shows the topology of the Bayesian network used in this study. The input nodes correspond to the presence/absence of the audio and video events in a certain region of the observed space. The output node indicates the state of the speech events. The output node $S$ has the following states: $S = \{-30^\circ, \cdots, +30^\circ, \text{NoEvent}\}$. For example, when $S = -30^\circ$, the speaker is located in the direction of $30^\circ$ and is speaking. When $S = \text{NoEvent}$, there are no speech events.

3. Robust speech interface

3.1. Overview

Figure 2 shows a block diagram of the speech interface. The input signal is observed using a microphone array and the spatial spectrum as depicted in the lower left panel of Fig. 1 is estimated in the Sound Localization module. On the other hand, in the Human Tracking module, the tracking results as depicted in the lower right
The information from the sound localization module and that from the human tracking module are then fed to the Information Fusion module. In this module, the location and the time of the speech events are estimated based on the co-occurrence of the audio and video events.

The information on the speech events are then sent to the Sound Separation module and the Speech recognition module. In the Sound Separation module, the target speech is separated from the environmental noise by application of the ML adaptive beamformer [8]. Based on the information on the speech events, the filter coefficients of the ML beamformer are kept updated.

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.

3.2. Sound separation

In the ML adaptive beamformer, the beamformer coefficient vector is given by

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

The matrix $K(\omega)$ is the noise spatial correlation observed in the absence of the target speech. The vector $\hat{g}(\omega)$ is the location vector of the target speech source which can be estimated in the presence of the target speech [1, 2].

Figure 3 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 spatial correlation $K(\omega)$ 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 $\hat{g}(\omega)$ is updated using the estimated location of the speech event.

3.3. Speech recognition

Table 1 shows the parameters used in ASR. The acoustic model of ASR is kept updated using the beamformer output so that the mismatch of the acoustic model in ASR and the input to ASR (output of the beamformer) is reduced. As a method of adaptation, a combination of MLLR [9] and MAP [10], in which MLLR-transformed means and variances are used as the priors for the MAP estimation, was employed [11]. Since online adaptation is assumed in this system, correctly labelled data for the adaptation are not available. In this study, an unsuper-

Table 1: Parameters of the speech recognizer.

<table>
<thead>
<tr>
<th>Feature parameter</th>
<th>MFCC (26 dimensions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis frame length</td>
<td>25ms</td>
</tr>
<tr>
<td>Analysis frame shift</td>
<td>10ms</td>
</tr>
<tr>
<td>Number of phones</td>
<td>43</td>
</tr>
<tr>
<td>Number of mixtures</td>
<td>16/state</td>
</tr>
<tr>
<td>Vocabulary size</td>
<td>492</td>
</tr>
</tbody>
</table>
4. Evaluation

4.1. Experimental condition

The microphone array used in the experiments was circular in 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. The microphone array and the stereo camera used in the experiment are shown in Fig. 4. The experiments were conducted in a medium-sized meeting room with a reverberation time of 0.5 s.

4.2. Exp.1 - with a realistic scenario

The setting of the experiment is shown in Fig. 5. Two human speakers spoke short phrases in the presence of TV sound and music presented by a loudspeaker. An example of the input/output waveform of the beamformer is shown in Fig. 6. In Fig. 6(b), bars indicating the detected and the true speech events are also shown. From these, it can be seen that the speech events were largely correctly detected. Also, the speech signal which was almost buried by the interference was recovered by the ML beamforming.

4.3. Exp.2 - word recognition task

The configuration of sound sources and humans is shown in Table 2. Speakers #1 and #2 spoke Japanese words alternatively at range of $r = 1.5$ m or $r = 3.0$ m. The number of words were 492. The level of the noise (music) was adjusted so that the SNR was around 0 dB in the previous test.

The word accuracy score, $R_w = (H - I)/N_w$, where $H$ is the number of correctly recognized words, $I$ is the number of insertions, and $N_w$ is the number of words, is shown in Table 3. In this table, the following abbreviations are used: Det.: extraction of speech segments based on the speech event detection; Sep: sound separation; and Adp.: acoustic model adaptation. In case D in Table 3, in which all three types of processing were employed, an ASR rate of 90% was achieved when $r = 1.5$ m, which is considered to be within a practical range.
Table 3: Speech recognition rate, $R_a$.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Det.</th>
<th>Sep.</th>
<th>Adp.</th>
<th>$r=1.5$ m</th>
<th>$r=3.0$ m</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>On</td>
<td></td>
<td></td>
<td>29.7%</td>
<td>9.6%</td>
</tr>
<tr>
<td>B</td>
<td>On</td>
<td>On</td>
<td></td>
<td>79.7%</td>
<td>54.3%</td>
</tr>
<tr>
<td>C</td>
<td>On</td>
<td></td>
<td></td>
<td>26.0%</td>
<td>-36.0%</td>
</tr>
<tr>
<td>D</td>
<td>On</td>
<td>On</td>
<td>On</td>
<td>91.1%</td>
<td>80.1%</td>
</tr>
</tbody>
</table>

Figure 7: Architecture of the system.

5. Real-Time system implementation

5.1. System configuration

The system consists of 2 PCs and the array signal processing hardware, RASP-1, described in the next section. Human tracking by vision and information fusion are conducted in the first PC. Speech recognition and model adaptation are processed in the second PC. Array signal processing, such as sound localization and sound separation, is conducted in RASP-1.

5.2. Array signal processor RASP-1

The array signal processing hardware, RASP-1, was developed by authors. The architecture of RASP-1 is depicted in Fig. 7. It consists of three boards. The analog board consists of an 8-channel A/D, a 2-channel D/A, microphone amplifiers and anti-aliasing filters. The signal processing board serves as a mother board of the system. Also, this board includes an FPGA circuit (Xilinx,VirtexII). By using this, low level signal processing such as FIR Filter is realized by this hardware. By doing this, real-time processing is guaranteed. In the CPU board, high level signal processing such as sound localization and the estimation of the sound separation filters are conducted. Figure 8 shows the appearance of the system.

6. References