BIMODAL SPEECH RECOGNITION USING LIP MOVEMENT MEASURED BY OPTICAL-FLOW ANALYSIS

Koji Iwano, Satoshi Tamura and Sadaoki Furui

Department of Computer Science
Graduate School of Information Science and Engineering
Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552, Japan
{iwano, tamura, furui}@furui.cs.titech.ac.jp

ABSTRACT

This paper proposes bimodal speech recognition using lip movements measured by optical-flow analysis. The optical flow is defined as the distribution of apparent velocities of brightness pattern movements. Since the optical flow can be computed without extracting the speaker's lip contours and location, robust visual features can be obtained on lip movements. Our method calculates two visual features in each frame: variances of horizontal and vertical components of flow velocities. Since these features represent movement of the speaker's mouth, they are especially useful for estimating pause/silence periods in noise-corrupted speech. The visual features are combined with acoustic features in the framework of HMM-based recognition. Phoneme HMMs are trained using the combined features extracted from clean speech data. In recognizing noise-corrupted speech, the observation probability of visual features are weighted. Experiments have been carried out using audio-visual data by 11 male speakers uttering connected digits. The following improvements of word accuracy over the audio-only recognition scheme were achieved by combining visual information only for silence HMM: 5% at SNR=5dB and 12% at SNR=10dB.

1. INTRODUCTION

Recently, there have been increasing interests in bimodal speech recognition systems incorporating both visual and acoustic information. These systems have been reported to increase the robustness and improve the performance over audio-only ASR, especially in the presence of additive noise including cross-talk [1-4]. In most of these methods, speaker's mouth locations and lip contours are extracted from image sequences with parametric models and the model parameters are used as visual features. In order to ensure robust extraction of visual features, preprocessing such as lip marking is often necessary.

Optical flow is the distribution of apparent velocities of movement of brightness patterns in an image[5], and it can be extracted without prior knowledge about the shape of the object. Therefore, the optical-flow analysis can detect visual features without extracting lip locations and contours.

In this paper, we propose a bimodal speech recognition scheme using the optical-flow velocities of lip images as visual information. Our recognition scheme uses a “feature fusion” technique; the visual features obtained from optical-flow velocities are combined with acoustic features in each frame. We describe the recognition results of noise-corrupted speech and compare the results with the audio-only recognition method.

2. OPTICAL-FLOW ANALYSIS

We use the Horn-Schunck optical-flow analysis[5]. Image brightness at a point \((x,y)\) in an image plane at time \(t\) is denoted by \(E(x,y,t)\). We assume that brightness of each point is constant during a movement for a very short time, then the equation becomes as follows:

\[
\frac{dE}{dt} \approx \frac{\partial E}{\partial x} \frac{dx}{dt} + \frac{\partial E}{\partial y} \frac{dy}{dt} + \frac{\partial E}{\partial t} = 0.
\]

If we let

\[
\frac{dx}{dt} = u \quad \text{and} \quad \frac{dy}{dt} = v,
\]

then a single linear equation

\[
E_x \cdot u + E_y \cdot v + E_t = 0
\]

is obtained. The vectors \(u\) and \(v\) denote apparent velocities of brightness constrained by this equation. Since flow velocity \((u, v)\) cannot be determined by this equation only, we use the additional constraint which minimizes the square magnitude of the gradient of the optical flow velocity:

\[
\left( \frac{\partial u}{\partial x} \right)^2 + \left( \frac{\partial u}{\partial y} \right)^2 \quad \text{and} \quad \left( \frac{\partial v}{\partial x} \right)^2 + \left( \frac{\partial v}{\partial y} \right)^2.
\]
This is called “smoothness constraint”. As a result, an optical-flow pattern under the condition that the apparent velocity of brightness pattern varies smoothly almost everywhere in the image is obtained. A flow velocity of a point is practically computed by an iterative scheme using the average of flow velocities estimated over neighboring pixels.

3. BIMODAL SPEECH RECOGNITION USING OPTICAL-FLOW ANALYSIS

3.1. Feature Extraction

Figure 1 shows our bimodal speech recognition system using the optical-flow analysis. First, we record speech and lip images around lips using a digital video camera. Audio signals are sampled at 16kHz with 16bit resolution. Each speech frame is converted into 39 acoustic parameters: 12 MFCCs, normalized log-energy, and their first and second order derivatives. These acoustic features are computed with frame rate of 100 frames/s. Visual signals are represented by RGB video captured with frame rate of 15 frames/s, and each image has 720×480 pixel resolution. Before computing the optical-flow, we reduce the image size to 180×120, keeping the aspect ratio. The image is transformed to gray-scale. In addition, we adjusted the gradient of the brightness pattern by low-pass filtering and adding low level random noise. Optical-flow velocities are calculated from a pair of connected images, using 3 iterations. An example of the optical-flow analysis result is shown in Figure 2.

Finally, two visual features are computed: horizontal and vertical variances of flow vector components. Since these features indicate whether speaker’s mouth is moving or not, they are especially useful for estimating pause/silence periods.

3.2. Feature Fusion

The acoustic and visual parameters are combined into a single vector, and this vector is used in model training and recognition processes. In order to cope with the frame rate difference, visual parameters are interpolated from 15Hz to 100Hz by a 3-degree spline function. After this step, acoustic and interpolated visual features are simply concatenated to build 41-dimensional audio-visual features.

3.3. Training and Recognition

Triphone HMMs are trained using the audio-visual features. The HMM has 3 states and 2 mixtures in each state. After training, 41-dimensional vectors in all triphones are separated into 2 streams: one stream is for 2-dimensional visual features (visual stream) and the other is for 39-dimensional acoustic features (audio stream).

In recognition, the probability \( b_j(o_{av}) \) of generating audio-visual observation \( o_{av} \) for state \( j \) is calculated by:

\[
b_j(o_{av}) = b_a(j, o_a)^{\lambda_a} \times b_v(j, o_v)^{\lambda_v},
\]

where \( b_a(j, o_a) \) is the probability of generating acoustic observation \( o_a \), and \( b_v(j, o_v) \) is the probability of generating visual observation \( o_v \). \( \lambda_a \) is a weighting factor for the audio stream, and \( \lambda_v \) is for visual stream. Their constraint is

\[
\lambda_a + \lambda_v = 1.
\]

4. EXPERIMENTS

4.1. Database

The audio-visual speech database was collected from 11 male speakers in a clean/quiet condition. The microphone and DV camera were located approximately 1m away from
mean word accuracy was computed as the measure of the recognition performance.

In the experiments reported in this paper, audio stream weight was varied only for silence HMM, and fixed at 1.0 for other triphones, that is, visual features were considered only for the silence triphone in the recognition process.

4.3. Experimental Results

Figure 3 shows digit recognition results at three SNR levels: 5, 10, 20dB. The horizontal axis shows the audio stream weight ($\lambda_o$). Solid and dotted lines indicate bimodal and audio-only speech recognition results, respectively. The difference between the bimodal and audio-only results at the audio stream weight of 1.0 is due to the difference of features used in training: 39 acoustic (audio-only) features vs. 41 audio-visual features. The optimum audio stream weight corresponding to the maximum accuracy increases with SNR. Table 1 shows the best bimodal recognition results and corresponding audio stream weights $\lambda_o$ in comparison with the audio-only results. These results show that the visual features are most effective at SNR=10dB and about 12% improvement was achieved. At SNR=20dB and in clean condition, almost no improvement was observed, probably because the accuracy was already very high without adding visual features.

As described above, each speaker uttered 250 digit strings with short pauses between them. In the above evaluation, the pause periods were removed from the speech signals for recognition, and pauses inserted as recognition results within digit string durations were not counted as errors. However, in actual applications, it is important that digits are not inserted in pause periods and pauses should not be inserted within digit string periods. Therefore we have conducted an additional evaluation, in which pause insertion within speech periods and substitution of pauses by digits as well as digit insertion within pauses were counted as errors. Table 2 shows the comparison between the previous and new evaluation results for both audio-only and audio-visual methods at the best audio stream weighting condition ($\lambda_o = 0.90$) for SNR=10. The table shows the number of errors per speaker. The difference between the results including/not including pauses in the audio-visual condition is significantly smaller than that in the audio only condition. This means that pause periods are much more correctly detected by the audio-visual method than the audio-only method.

5. CONCLUSIONS

This paper has proposed a bimodal speech recognition scheme using the optical flow obtained from image sequences of lip movements. So far we have combined au-
Figure 3: Word accuracy of audio-visual recognition in comparison with audio-only recognition at three SNR levels (top: SNR=5dB, middle: SNR=10dB, bottom: SNR=20dB).

Table 1: The best audio-visual recognition results at four SNR levels.

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>Audio-only</th>
<th>Audio-visual (λ≠)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>39.50%</td>
<td>44.95% (0.86)</td>
</tr>
<tr>
<td>10</td>
<td>58.55%</td>
<td>70.78% (0.90)</td>
</tr>
<tr>
<td>20</td>
<td>94.66%</td>
<td>94.51% (0.94)</td>
</tr>
<tr>
<td>∞</td>
<td>97.59%</td>
<td>97.96% (0.96)</td>
</tr>
</tbody>
</table>

Table 2: The number of deletion (Del), substitution (Sub) and insertion (Ins) errors per speaker. “including pause” treats pause periods as words (digits).

<table>
<thead>
<tr>
<th></th>
<th>Del</th>
<th>Sub</th>
<th>Ins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio-only</td>
<td>96.4</td>
<td>189.0</td>
<td>129.1</td>
</tr>
<tr>
<td>(including pause)</td>
<td>177.6</td>
<td>309.5</td>
<td>51.0</td>
</tr>
<tr>
<td>Audio-visual</td>
<td>145.6</td>
<td>129.2</td>
<td>17.4</td>
</tr>
<tr>
<td>(including pause)</td>
<td>190.7</td>
<td>118.9</td>
<td>22.4</td>
</tr>
</tbody>
</table>

6. ACKNOWLEDGMENTS

This research has been conducted in cooperation with NTT DoCoMo. The authors wish to express thanks for their support.

7. REFERENCES


