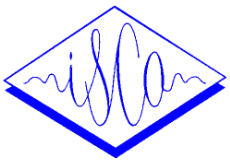

AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4–7, 2003

Audio-Visual Speech Recognition Using Lip Movement Extracted from Side-Face Images

Tomoaki Yoshinaga, Satoshi Tamura, Koji Iwano, and Sadaoki Furui

Department of Computer Science,
Tokyo Institute of Technology
{yossy, tamura, iwano, furui}@furui.cs.titech.ac.jp

Abstract

This paper proposes an audio-visual speech recognition method using lip movement extracted from side-face images to attempt to increase noise-robustness in mobile environments. Although most previous bimodal speech recognition methods use frontal face (lip) images, these methods are not easy for users since they need to hold a device with a camera in front of their face when talking. Our proposed method capturing lip movement using a small camera installed in a handset is more natural, easy and convenient. This method also effectively avoids a decrease of signal-to-noise ratio (SNR) of input speech. Visual features are extracted by optical-flow analysis and combined with audio features in the framework of HMM-based recognition. Phone HMMs are built by the multi-stream HMM technique. Experiments conducted using Japanese connected digit speech contaminated with white noise in various SNR conditions show effectiveness of the proposed method. Recognition accuracy is improved by using the visual information in all SNR conditions, and the best improvement is approximately 6% at 5dB SNR.

1. Introduction

Since speech input is a useful interface in mobile environments, there is increasing development of mobile devices incorporating automatic speech recognition (ASR) techniques. However, noise is a very serious problem for speech recognition, and increasing noise-robustness is one of the most important issues in speech recognition, especially in real mobile environments.

Audio-visual speech recognition, using face information in addition to acoustic features, has been investigated for increasing the robustness and improving the accuracy of ASR in noisy conditions[1-8]. Most use lip information extracted from frontal images of the face. When using these methods in mobile environments, users need to hold a handset with a camera in front of their mouth at some distance, which is not only unnatural but also inconvenient for talking. Furthermore, the recognition accuracy may worsen due to the decreasing SNR. If the lip information can be taken by using a handset held in the usual way of telephone conversation, this would greatly improve its desirability.

From this point of view, we propose an audio-visual speech recognition method using side-face images, assuming that a small camera is installed near the microphone of the mobile device. This method captures the images of lips located at a small distance from the microphone. The proposed method uses a bimodal speech recognition method [9] which extracts visual information of lip movements by optical-flow analysis.

In Section 2 of this paper, we will explain the optical-flow analysis method. Section 3 describes our audio-visual recognition method. Experimental results are reported in Section 4, and Section 5 concludes this paper.

2. Optical-flow analysis

The Horn-Schunck optical-flow analysis technique[10] was used. 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 period, then the equation becomes:

$$\frac{dE}{dt} \simeq \frac{\partial E}{\partial x}\frac{dx}{dt} + \frac{\partial E}{\partial y}\frac{dy}{dt} + \frac{\partial E}{\partial t} = 0. \quad (1)$$

If we let

$$\frac{dx}{dt} = u \quad \text{and} \quad \frac{dy}{dt} = v, \quad (2)$$

then a single linear equation

$$E_x \cdot u + E_y \cdot v + E_t = 0 \quad (3)$$

is obtained. The vectors u and v denote apparent velocities of brightness constrained by this equation. Since the flow velocity (u, v) cannot be determined only by this equation, we use an 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 is obtained, under the condition that the apparent velocity of brightness pattern varies smoothly in the image. The flow velocity of each point is practically computed by an iterative scheme using the average of flow velocities estimated from neighboring pixels.

3. Audio-visual speech recognition using optical-flow analysis

3.1. Overview

Figure 1 shows our bimodal speech recognition system using the optical-flow analysis[9].

First, both speech and lip images of the side view are synchronously recorded. Audio signals are sampled at 16kHz with 16bit resolution. Each speech frame is converted into 38 acoustic parameters: 12 MFCCs, 12 ΔMFCCs, 12 ΔΔMFCCs, Δ log energy, and ΔΔ log energy. The window length is 25ms. Cepstral mean subtraction (CMS) is applied to each utterance.

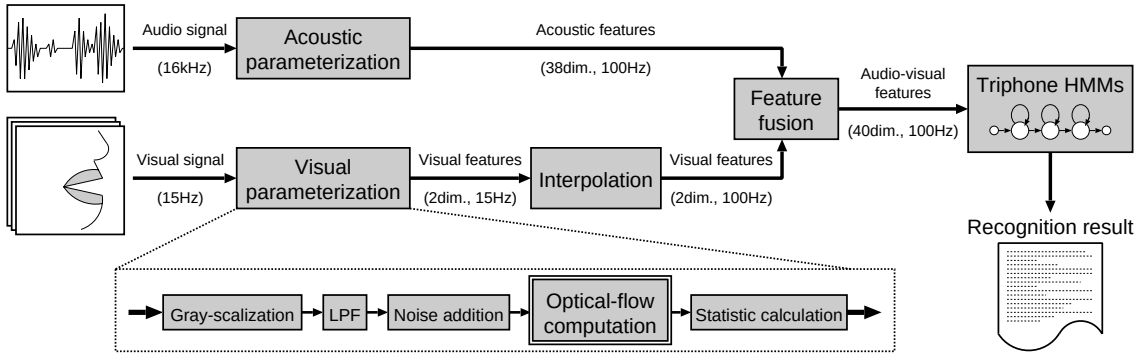

Figure 1: *Audio-visual speech recognition system using optical-flow analysis.*

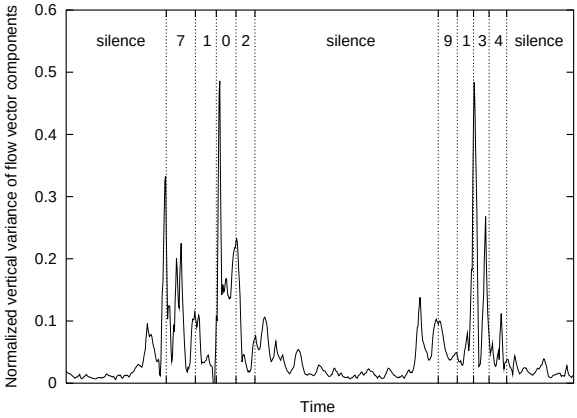

Figure 2: *An example of the vertical variance of flow vector components.*

The acoustic features are computed with a frame rate of 100 frames/s. Visual signals are represented by RGB video captured with a frame rate of 15 frames/s, and each image has a 720×480 pixel resolution. Before computing the optical-flow, we reduce the image size to 180×120, and then transform the images to gray-scale. Since the equation (1) assumes that the image plane has spatial gradient and correct optical-flow vectors cannot be computed at a point without spatial gradient, the visual signal is passed through a low-pass filter and low-level random noise is added to the filtered signal. Optical-flow velocities are calculated from a pair of connected images, using five iterations.

Next, two visual features, horizontal and vertical variances of flow vector components, are calculated for each frame and normalized by the maximum values in each utterance. Since these features indicate whether the speaker's mouth is moving or not, they are especially useful for detecting the onset of speaking periods. Figure 2 shows an example of a time function of the vertical variance for an utterance, "7102, 9134", as well as the period of each digit. It is shown that the features are almost 0 in pause/silence periods and have large values in speaking periods. It was found that time functions of the horizontal variance were similar to those of the vertical variance.

The acoustic and visual features are combined to make a single vector. In order to cope with the frame rate difference, the visual features are interpolated from 15Hz to 100Hz by a 3-degree spline function. After this step, the acoustic and interpolated visual features are simply concatenated to build a 40-dimensional audio-visual feature vector.

Triphone HMMs are modeled as multi-stream HMMs. In recognition, the probability $b_j(o_{av})$ of generating audio-visual observation o_{av} for state j is calculated by:

$$b_j(o_{av}) = b_{aj}(o_a)^{\lambda_a} \times b_{vj}(o_v)^{\lambda_v}, \quad (4)$$

where $b_{aj}(o_a)$ is the probability of generating acoustic observation o_a, and $b_{vj}(o_v)$ is the probability of generating visual observation o_v. λ_a and λ_v are weighting factors for the audio and the visual stream, respectively. They are constrained by $\lambda_a + \lambda_v = 1$.

3.2. Building multi-stream HMMs

In order to make the HMMs for recognition, we first trained audio and visual HMMs separately and combined them using a mixture-tying technique:

1. The audio HMMs are trained by 38-dimensional acoustic (audio) features. Each audio HMM has 3 states, except for the "sp (short pause)" model having a single state.
2. Training utterances are segmented into phonemes by the forced (Viterbi)-alignment technique using the audio HMMs, and time aligned labels are obtained.
3. The visual HMMs are trained by 2-dimensional visual features using the phoneme labels obtained by the step 2. Each visual HMM has 3 states, except for the "sp" and "sil (silence)" models having a single state.
4. The audio and visual HMMs are combined to build audio-visual HMMs. Gaussian mixtures in the audio stream of the audio-visual HMMs are tied with corresponding audio-HMM mixtures, while the mixtures in the visual stream are tied with corresponding visual HMM mixtures.

In all the HMMs, the number of mixtures is two.

4. Experiments

4.1. Database

An audio-visual speech database was collected from 38 male speakers in a clean/quiet condition. Each speaker uttered 50

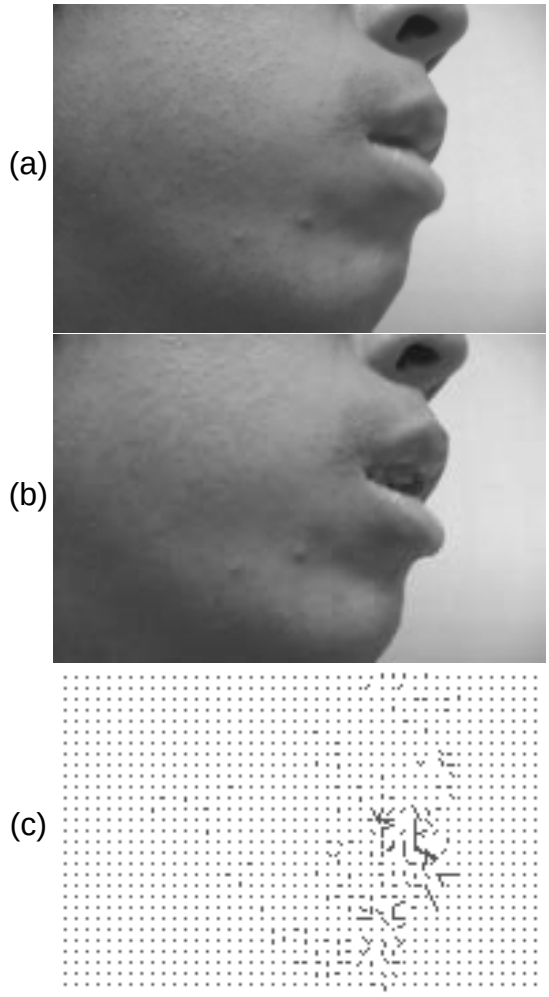

Figure 3: *An example of optical-flow analysis using a pair of lip images (a) and (b). Optical-flow velocities for lip image changes from (a) to (b) are shown in (c).*

sequences of 4 connected Japanese digits. Short pauses were inserted between the sequences.

With an assumption that speakers would use a mobile device with a small camera installed near a microphone, speech and lip images were recorded by a microphone and a DV camera located approximately 10cm from each speaker's right cheek. An example of consecutive two lip images is shown in Figure 3 (a) and (b). Figure 3 (c) shows the corresponding optical-flow analysis result indicating the lip image changes from (a) to (b).

4.2. Training and Recognition

The HMMs were trained using clean audio-visual data, and audio data for testing were contaminated with white noise at four SNR levels: 5, 10, 15, and 20dB. Nineteen speakers were selected for testing from the 38 speakers. Experiments were conducted using the leave-one-out method; data from one speaker were used for testing while data from other 37 speakers were used for training. Accordingly, 19 speaker-independent experiments were conducted, and the mean word accuracy was calculated as the measure of the recognition performance.

Table 1: Comparison of digit recognition accuracies with the audio-only and audio-visual methods in various SNR conditions.

SNR (dB)	Audio-only	Audio-visual (λ_a)
∞ (clean)	99.3%	99.4% (0.80)
20	89.6%	90.5% (0.60)
15	71.8%	75.4% (0.55)
10	48.4%	53.1% (0.60)
5	26.4%	32.3% (0.45)

Table 2: Comparison of the onset detection errors (ms) of speaking periods in various SNR conditions.

SNR (dB)	Audio-only	Audio-visual
20	41.2	35.1
15	54.9	45.9
10	76.2	63.1
5	104.1	86.6

4.3. Experimental Results

Table 1 shows digit recognition accuracies in each SNR condition with the audio-only and audio-visual methods. The audio and visual stream weights of the bimodal method were optimized at each condition. The optimized audio stream weights (λ_a) are also shown in combination with the audio-visual recognition accuracies in the table. In all the SNR conditions, digit accuracies are improved by using the visual features. The best improvement, 5.9% for the absolute value, is observed in the 5dB SNR condition.

Figure 4 shows the digit recognition accuracy as a function of the audio stream weight (λ_a) in 5, 10, 15, and 20dB SNR conditions. The horizontal axis indicates the audio stream weight (λ_a), and the solid and dotted lines indicate audio-visual and audio-only speech recognition results, respectively. Effectiveness of the visual features is observed over a wide range of stream weight in all the SNR conditions.

As a supplementary experiment, we compared audio-visual HMMs and audio HMMs in terms of the onset detection capability for speaking periods under noisy environments. Noise-added utterances and clean utterances were segmented by either of these models using the forced-alignment technique, and the detected boundaries between silence and beginning of each digit sequence were used to evaluate the performance of onset detection. The amount of errors (ms) was measured by averaging the differences of detected onset locations for noise-added utterances and clean utterances. Table 2 shows the onset detection errors in various SNR conditions. The audio and visual stream weights were optimized at each condition in the audio-visual HMMs. The mean onset detection error rate was reduced by approximately 17% for 10dB and 5dB SNR utterances using the audio-visual HMMs. We attribute the better recognition performance by the proposed method to the precise boundary detection.

5. Conclusions

This paper has proposed a bimodal speech recognition scheme using lip movements extracted from side-face images to achieve

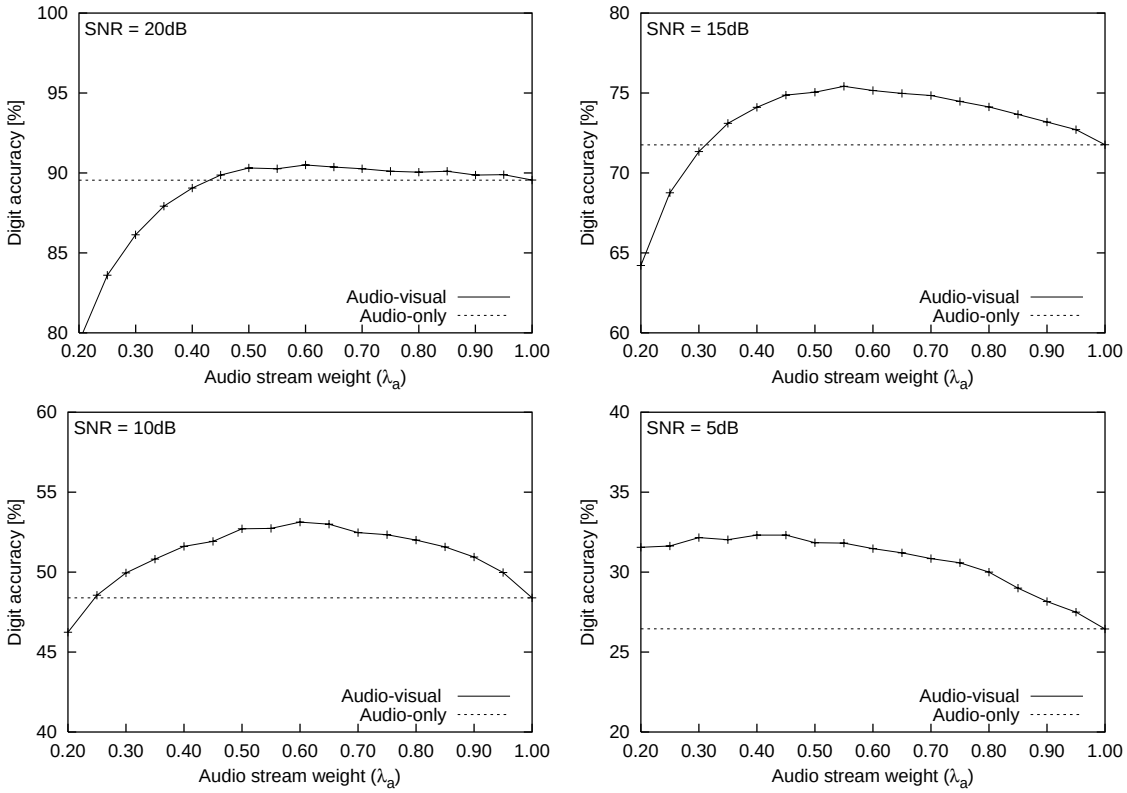

Figure 4: *Digit recognition accuracy as a function of the audio stream weight (λ_a) .*

robust, natural and easy voice communication in mobile environments. Visual features were extracted by optical-flow analysis of lip image sequences. The proposed method achieved approximately 6% and 5% word accuracy improvement in 5dB and 10dB SNR conditions, respectively. The visual features are significantly useful for detecting the onset of speaking periods and consequently improve recognition performance in noisy conditions.

Future works include the combination of the proposed method with model adaptation techniques such as MLLR and evaluation using more general recognition tasks. We are also planning to combine several techniques to increase robustness in extracting visual features.

6. Acknowledgements

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

7. References

[1] Bregler, C. and Konig, Y., " "Eigenlips" for robust speech recognition," *Proc. ICASSP94*, vol.2, pp.669–672, Adelaide, Australia, 1994.

[2] Tomlinson, M.J., Russell, M.J., and Brooke, N.M., "Integrating audio and visual information to provide highly robust speech recognition," *Proc. ICASSP96*, vol.2, pp.821–824, Atlanta, USA, 1996.

[3] Potamianos, G., Cosatto, E., Graf, H.P., and Roe, D.B., "Speaker independent audio-visual database for bimodal ASR," *Proc. AVSP'97*, pp.65–68, Rhodes, Greece, 1997.

[4] Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., and Zhou, J., "Audio-visual speech recognition," *Final Workshop 2000 Report*, Center for Language and Speech Processing, Baltimore, 2000.

[5] Dupont, S. and Luettin, J., "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol.2, no.3, pp.141–151, 2000.

[6] Zhang, Y., Levinson, S., and Huang, T.S., "Speaker independent audio-visual speech recognition," *Proc. ICME2000*, TP8-1, New York, USA, 2000.

[7] Chu, S.M. and Huang, T.S., "Bimodal speech recognition using coupled hidden markov models," *Proc. ICSLP2000*, vol.2, pp.747–750, Beijing, China, 2000.

[8] Miyajima, C., Tokuda, K., and Kitamura, T., "Audio-visual speech recognition using MCS-based HMMs and model-dependent stream weights," *Proc. ICSLP2000*, vol.2, pp.1023–1026, Beijing, China, 2000.

[9] Iwano, K., Tamura, S., and Furui, S., "Bimodal speech recognition using lip movement measured by optical-flow analysis" *Proc. HSC2001*, pp.187–190, Kyoto, Japan, 2001.

[10] Horn, B.K.P and Schunck, B.G., "Determining Optical Flow," *Artificial Intelligence*, vol.17, nos.1–3, pp.185–203, 1981.