論文 / 著書情報
Article / Book Information

Title	Audio-visual speech recognition using new lip features extracted from side-face images
Authors	Tomoaki Yoshinaga, Satoshi Tamura, Koji iwano, Sadaoki Furui
Citation	Robust2004, Vol. , No. , pp. 33
Pub. date	2004, 8

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

Tomoaki Yoshinaga[*]*, Satoshi Tamura, Koji Iwano, and Sadaoki Furui*

Department of Computer Science, Tokyo Institute of Technology
t-yoshi@crl.hitachi.co.jp, {tamura, iwano, furui}@furui.cs.titech.ac.jp

Abstract

This paper proposes new visual features for audio-visual speech recognition using lip information extracted from side-face images. In order to increase the noise-robustness of speech recognition, we have proposed an audio-visual speech recognition method using speaker lip information extracted from side-face images taken by a small camera installed in a mobile device. Our previous method used only movement information of lips, measured by optical-flow analysis, as a visual feature. However, since shape information of lips is also obviously important, this paper attempts to combine lip-shape information with lip-movement information to improve the audio-visual speech recognition performance. A combination of an angle value between upper and lower lips (lip-angle) and its derivative is extracted as lip-shape features. Effectiveness of the lip-angle features has been evaluated under various SNR conditions. The proposed features improved recognition accuracies in all SNR conditions in comparison with audio-only recognition results. The best improvement of 8.0% in absolute value was obtained at 5dB SNR condition. Combining the lip-angle features with our previous features extracted by the optical-flow analysis yielded further improvement. These visual features were confirmed to be effective even when the audio HMM used in our method was adapted to noise by the MLLR method.

1. Introduction

In the recent mobile environment, necessity of noise-robust speech recognition is widely spreading. Audio-visual (bimodal) speech recognition techniques using face information in addition to acoustic information are promising directions for increasing the robustness of speech recognition, and many audio-visual methods have been proposed thus far[1-8]. However, most of them use lip information extracted from frontal images of the face, users need to hold a handset with a camera in front of their mouth. This is not only unnatural but also inconvenient for talking in a mobile environment. If the lip information can be captured in the usual way of holding the handset in telephone conversations, this would be more desirable for the users.

From this point of view, we previously proposed an audio-visual method using side-face images, assuming that a small camera can be installed near the microphone of the mobile device[9]. In our method, a bimodal speech recognition technique[10] was employed and lip-movement features extracted by an optical-flow analysis were used as visual features[9, 10]. Specifically, horizontal and vertical variances of optical-flow vector components were used, and the effectiveness of using lip-movement information in increasing the noise robustness was confirmed. However, since it is obvious that lip-shape information is also important for lip-reading by humans, in this paper we investigate using shape information extracted from side-face images for further increasing the performance.

In this paper, we first propose new visual features which can be measured using an angle between upper and lower lips. The angle is hereafter referred to as "lip-angle". Then, effectiveness of the proposed features is investigated using our audio-visual speech recognition scheme.

In Section 2, we explain the method for extracting the angle information. Section 3 describes our audio-visual recognition method. Experimental results are reported in Section 4, and Section 5 concludes this paper.

2. Lip-angle extraction

The lip-angle extraction process consists of three components: (1) detecting a lip area, (2) extracting a center point of lips, and (3) determining lip-lines and a lip-angle. Details are explained in the following subsections.

2.1. Detecting a lip area

Speaker's lips in the video data of the side view are tracked by using a rectangular window. An example of a detected rectangular image is shown in Figure 1.

For detecting a rectangular lip area from an image frame, two kinds of image processing methods are used: edge detection by Sobel filtering and binarization using hue values. Examples of the edge image and the binary image are shown in Figures 2 and 3, respectively. As shown in Figure 2, the edge image is effective in detecting horizontal positions of a nose, a mouth, lips and a jaw. Therefore, the edge image is used for horizontal search of the lip area; first counting the number of edge points on every vertical line in the image, and finding the image area which has a larger value of edge points than a preset threshold. Since lips, cheek, and chin areas have a hue value around $1.5\pi \sim 2.0\pi$, and they do not exist in the same horizontal level, the binary image is used for vertical search of the lip area, in the same way as the horizontal search in the edge image.

2.2. Extracting the center point of lips

The center point of lips is defined as an intersection of upper and lower lips, as shown in Figure 1. For finding the center point, a dark area considered to be inside of the mouth is first extracted from the rectangular lip area. The dark area is defined as a set of pixels having brightness values lower than 15. The left-most point of the dark area is extracted as the center point.

[*] Currently at Central Research Laboratory, Hitachi Ltd., Kokubunji, Tokyo 185-8601, Japan

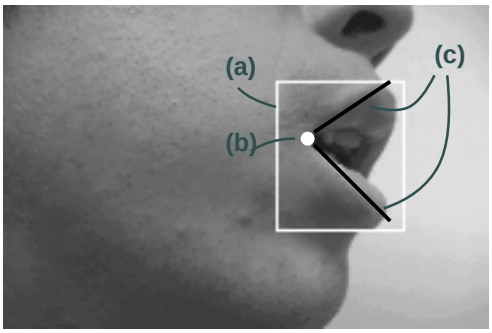

Figure 1: *An example of (a) a detected rectangular lip area, (b) an extracted center point of lips, and (c) detected lip-lines.*

Figure 2: *An example of the edge image.*

Figure 3: *An example of the binary image.*

2.3. Determining lip-lines and a lip-angle

Finally, two lines modeling upper and lower lips are determined in the lip area. These lines are referred to as "lip-lines". Examples of detected lip-lines are shown in Figure 1.

The detecting process is as follows:

1. An AND (overlapped) image is created for edge and binary images.
2. Line segments are radially drawn from the center point to the right in the image at every small step of the angle, and the number of AND points on each line segment is counted.
3. A line segment having the maximum number of points is detected as the "base line" which is used for detecting upper and lower lip-lines. A dotted line in Figure 4 shows an example of the base line.
4. The number of points on each line segment is counted in the binary image made by using hue values.
5. Line segments with the maximum value above and below the base line are respectively detected as upper and lower lip-lines. Solid lines in Figure 4 indicates examples of the lip-lines.

Finally, a lip-angle between the upper and lower lip-lines is measured.

3. Audio-visual speech recognition using lip-angle features

3.1. Overview

Figure 5 shows our bimodal speech recognition system using lip-angle features[10]. First, both speech signals and lip images of the side view are synchronously recorded. Audio signals

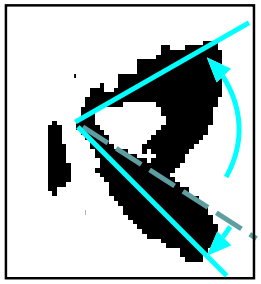

Figure 4: *An example of determined lip-lines.*

are sampled at 16kHz with 16bit resolution. The speech signal at each frame is converted into 38 acoustic parameters: 12 MFCCs, 12 ΔMFCCs, 12 $\Delta\Delta$MFCCs, Δ log energy, and $\Delta\Delta$ log energy. The window length and the frame rate are set at 25 ms and 100 frames/s, respectively. Cepstral mean subtraction (CMS) is applied to each utterance. Visual signals are captured as RGB video signals with a frame rate of 30 frames/s, where each image has a 720$\times$480 pixel resolution. Before computing the lip-angle, the image size is reduced to 180$\times$120.

Next, two dimensional visual features, consisting of a lip-angle and its derivative (delta), are calculated for each frame and normalized by the maximum values in each utterance. Figure 6 shows an example of a time function of the normalized lip-angle for a Japanese digit utterance, "7102, 9134". It is shown that the features are almost constant in pause/silence periods and have large values when the speaker's mouth is widely opened.

The lip-angle features and the visual features obtained by optical-flow analysis[10] are further combined to produce a four dimensional visual feature set which is then evaluated in comparison with the two dimensional feature set consisting of only the lip-angle features.

The acoustic and visual features are combined to construct a single vector. In order to compensate for the frame rate difference, the visual features are interpolated from 30Hz to 100Hz by a 3-degree spline function. After this step, the acoustic and interpolated visual features are simply concatenated to build a 40(42)-dimensional audio-visual feature vector.

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

$$b_j(o_{av}) = {b_{aj}(o_a)}^{\lambda_a} \times {b_{vj}(o_v)}^{\lambda_v}, \quad (1)$$

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, audio and visual HMMs are trained separately and combined using a mixture-tying technique as follows.

1. The audio HMMs are trained using 38-dimensional acoustic (audio) features. Each audio HMM has 3 states, except for the "sp (short pause)" model which has a single state.
2. Training utterances are segmented into phonemes by the forced (Viterbi)-alignment technique using the audio

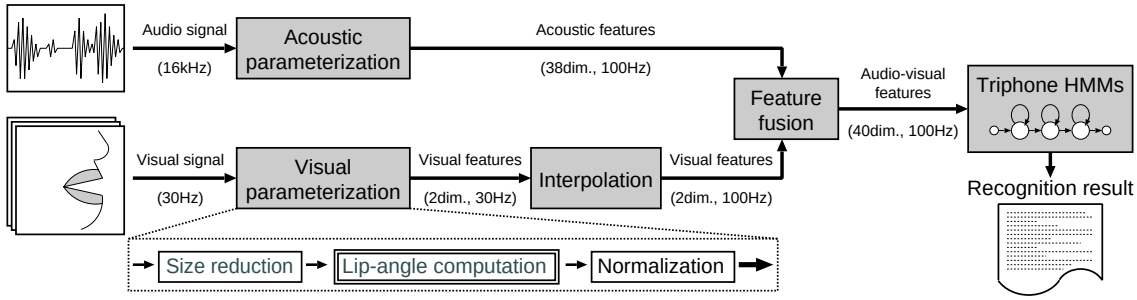

Figure 5: *Audio-visual speech recognition system using lip-angle features.*

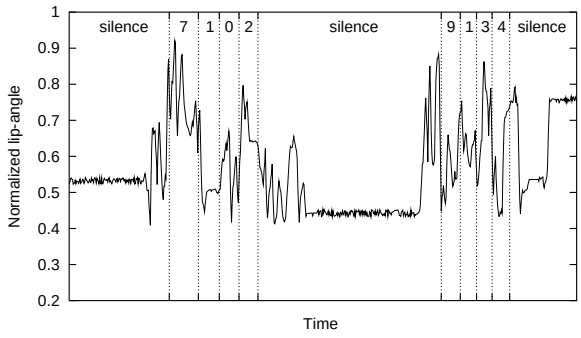

Figure 6: *An example of a time function of the normalized lip-angle value.*

HMMs, and time aligned labels are obtained.

3. The visual HMMs are trained by 2(4)-dimensional visual features using the phoneme labels obtained by step 2. Each visual HMM has 3 states, except for the "sp" and "sil (silence)" models which have 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 set at 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 sequences of four connected digits in Japanese. Short pauses were inserted between the sequences.

In order to simulate the situation in which speakers would be using 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 away from each speaker's right cheek.

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. Experiments were conducted using the leave-one-out method; data from one speaker were used for testing, while data from the remaining 37 speakers were used for training. Accordingly, 38 speaker-independent experiments were conducted, and a mean word accuracy was calculated as the measure of the recognition performance.

4.3. Experimental Results

Table 1 shows digit recognition accuracies obtained by the audio-only and the audio-visual methods at various SNR conditions. Accuracies using only optical-flow features[9] are also shown in the table for comparison. The audio and visual stream weights in the audio-visual methods were optimized at each condition. The optimized audio stream weights (λ_a) are shown next to the audio-visual recognition accuracies in the table. In all the SNR conditions, digit accuracies were improved by using lip-angle features compared to results obtained by the audio-only method. The best improvement from the baseline (audio-only) results, 8.0% in absolute value, was observed at the 5dB SNR condition.

Combining visual features improved digit accuracies more than either of the 2-dimensional audio-visual feature results, which used only optical-flow features or lip-angle features as visual information, at all the SNR conditions. The absolute improvement of the accuracy at the 5dB SNR condition was 10.9% from the audio-only (baseline) method.

Figure 7 shows the digit recognition accuracy as a function of the audio stream weight (λ_a) at the 5dB SNR condition. The horizontal and vertical axes indicate the audio stream weight (λ_a) and the digit recognition accuracy, respectively. The dotted straight line indicates the baseline (audio-only) results, and others indicate the results obtained by audio-visual methods. For all the visual features conditions, improvements from baseline are observed over a wide range of the stream weight. The range over which accuracy is improved is largest when the combined visual features are used. This means that the proposed audio-visual recognition methods are not sensitive to the change of stream weights, and the method using the combined visual features are the most robust for the change of the weight.

To confirm the effectiveness of the proposed visual features in a condition of combination with noise-adapted audio HMM, a supplementary experiment was conducted. The audio-visual HMM was constructed by integrating the audio HMM adapted by the MLLR (Maximum Likelihood Linear Regression) method and non-adapted visual HMM. Table 2 shows the results when using the adapted audio-visual HMM. Comparing these to the results of the baseline (audio-only) method in Tables 1, it can be observed that accuracies are largely improved by the MLLR adaptation. It can also be observed that the visual features further improve the performance. Consequently, the best

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

SNR (dB)	Audio-only (baseline)	Audio-visual (Optimized λ_a) Optical-flow	Lip-angle	Combined
20	91.5%	92.2% (0.60)	92.3% (0.55)	92.6% (0.70)
15	75.6%	78.7% (0.55)	79.1% (0.35)	79.9% (0.55)
10	51.9%	56.7% (0.60)	57.5% (0.30)	59.4% (0.45)
5	28.4%	34.7% (0.40)	36.4% (0.20)	39.3% (0.25)

Table 2: Comparison of digit recognition accuracies when MLLR-based audio-visual HMM adaptation is applied.

SNR (dB)	Audio-only (baseline)	Audio-visual (Optimized λ_a) Optical-flow	Lip-angle	Combined
20	97.0%	97.2% (0.90)	97.4% (0.60)	97.2% (0.90)
15	91.5%	93.3% (0.55)	93.3% (0.55)	93.4% (0.70)
10	69.4%	76.9% (0.45)	77.2% (0.30)	79.5% (0.35)
5	39.5%	52.6% (0.30)	53.1% (0.20)	58.4% (0.30)

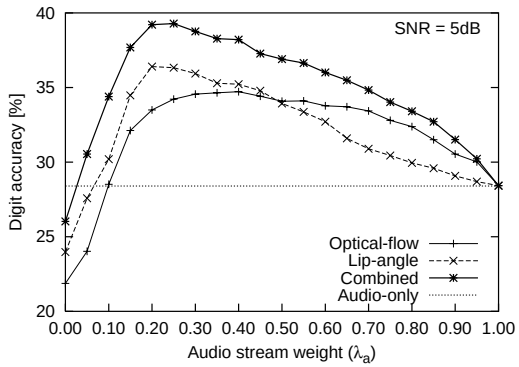

Figure 7: *Digit recognition accuracy as a function of the audio stream weight (λ_a) at 5dB SNR condition.*

improvement from the non-adapted audio-only result, 29.1% in absolute value at the 5dB SNR condition, was observed when using the adapted audio-visual HMM which included the combined features.

5. Conclusions

This paper has proposed new visual features for audio-visual speech recognition using lip information extracted from side-face images. The proposed features, consisting of the lip-angle between upper and lower lip-lines and its delta, achieved significant improvement at all SNR conditions. Combination with previously proposed optical-flow features further improved recognition accuracies. The improvement by using the visual features was confirmed even when MLLR-based noise adaptation was applied to the audio HMM.

Future works include (1) evaluation using more general recognition tasks, (2) developing an optimization method of stream weights, and (3) improving the combination method of lip-angle and optical-flow 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] Yoshinaga, T., Tamura, T., Iwano, K., Furui, S., "Audio-visual speech recognition using lip movement extracted from side-face images," *Proc. AVSP2003*, pp.117–120, St. Joriouz, France, 2003.

[10] 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.