Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2007, Article ID 64506, 9 pages
doi:10.1155/2007/64506

Research Article

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

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

Department of Computer Science, Tokyo Institute of Technology, 2-12-1-W8-77 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

Received 12 July 2006; Revised 24 January 2007; Accepted 25 January 2007

Recommended by Deliang Wang

This paper proposes an audio-visual speech recognition method using lip information extracted from side-face images as an attempt to increase noise robustness in mobile environments. Our proposed method assumes that lip images can be captured using a small camera installed in a handset. Two different kinds of lip features, lip-contour geometric features and lip-motion velocity features, are used individually or jointly, in combination with audio features. Phoneme HMMs modeling the audio and visual features are built based on the multistream 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. These visual features were confirmed to be effective even when the audio HMM was adapted to noise by the MLLR method.

1. INTRODUCTION

In the current environment of mobile technology, the demand for noise-robust speech recognition is growing rapidly. 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 [1–11]. Most use lip information extracted from frontal images of the face. However, 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 conversation. Since the distance between the mouth and the handset decreases SNR, recognition accuracy may worsen. If the lip information can be taken by using a handset held in the usual way for telephone conversations, this would greatly improve the usefulness of the system.

From this point of view, we propose an audio-visual speech recognition method using side-face images, assuming that a small camera can be installed near the microphone of the mobile device in the future. This method captures the images of lips located at a small distance from the microphone. Many geometric features, mouth width and height [3, 11], teeth information [11], and information about points located on a lip-contour [6, 7], have already been used for bimodal speech recognition based on frontal-face images. However, since these features were extracted based on "oval" mouth shape models, they are not suitable for side-face images. To effectively extract geometric information from side-face images, this paper proposes using lip-contour geometric features (LCGFs) based on a time series of estimated angles between upper and lower lips [12]. In our previous work on audio-visual speech recognition using frontal-face images [9, 10], we used lip-motion velocity features (LMVFs) derived by optical-flow analysis. In this paper, LCGFs and LMVFs are used individually and jointly [12, 13]. (Preliminary versions of this paper have been presented at workshops [12, 13].) Since LCGFs use lip-shape information, they are expected to be effective in discriminating phonemes. On the other hand, since LMVFs are based on lip-movement information, they are expected to be effective in detecting voice activity. In order to integrate the audio and visual features, a multistream HMM technique is used.

In Section 2, we explain the method for extracting the LCGFs. Section 3 describes the extraction method of the LMVFs based on optical-flow analysis. Section 4 explains our audio-visual recognition method. Experimental results are reported in Section 5, and Section 6 concludes this paper.

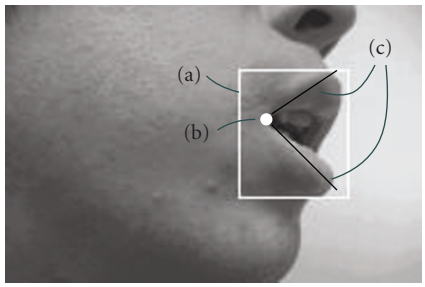


Figure 1: An example of the lip image extraction process: (a) an edge image detected using Sobel filtering, (b) a binary image obtained by thresholding hue values, and (c) a detected lip-area image.

2. EXTRACTION OF LIP-CONTOUR GEOMETRIC FEATURES

Upper and lower lips in side-face images are modeled by two-line components. An angle between the two lines is used as the lip-contour geometric features (LCGFs). The angle is hereafter referred to as "lip-angle." 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

In the side-view video data, speaker's lips are detected by using a rectangular window. An example of a detected rectangular area 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(a) and 2(b), respectively. As shown in Figure 2(a), 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 then finding the image area which has a larger value of edge points than a preset threshold. The area (1) in Figure 2(a) indicates the area detected by the horizontal search.

Since lips, cheek, and chin areas have hue values within $1.5\pi \sim 2.0\pi$, these areas are detected by thresholding the hue values in the above detected area. The region labeling technique [14] is applied to the binary image generated by the thresholding process to detect connected regions. The largest connected region in the area (1), indicated by (2) in Figure 2(b), is extracted as a lip area.

To determine a final square area (3), horizontal search on an edge image and vertical search on a binary image are sequentially conducted to cover the largest connected region. Since these two searches are independently conducted, the aspect ratio of the square is variable. The original image of the square area shown in Figure 2(c) is extracted for use in the following process.

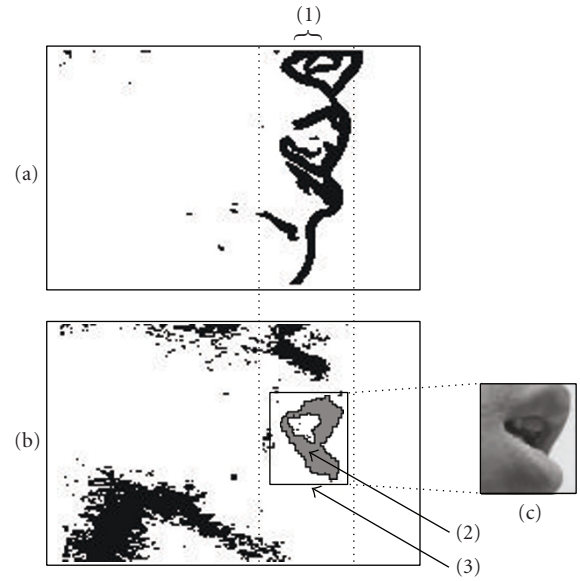


Figure 2: Examples of lip images used for lip-area detection: (a) an edge image detected by Sobel filtering, (b) a binary image obtained using hue values, and (c) a detected lip-area image.

2.2. Extracting the center point of lips

The center point of the lips is defined as an intersection of the upper and lower lips, as shown in Figure 1. For finding the center point, a dark area considered to be the 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 a preset threshold. In our experiments, the threshold was manually set to 15 after preliminary experiments using a small dataset.[1] The leftmost point of the dark area is extracted as the center point.

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." The detecting process is as follows.

(1) An AND (overlapped) image is created for edge and binary images. Figure 3(a) shows an example of an AND image. A gray circle indicates the extracted center point of the lips.
(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.

[1] The threshold value was manually optimized to achieve a good balance between dark and light areas.

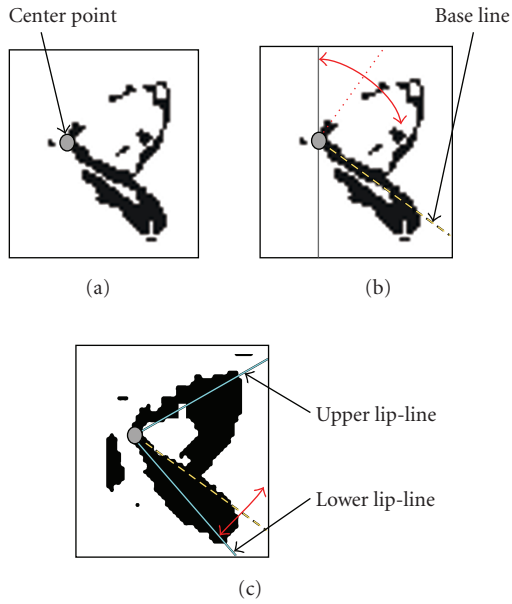


FIGURE 3: Selected stages in the lip-line determination process.

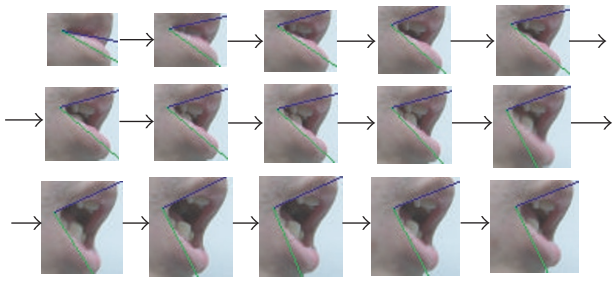


FIGURE 4: An example of the extracted lip-line feature sequence with a frame rate of 30 frames/s.

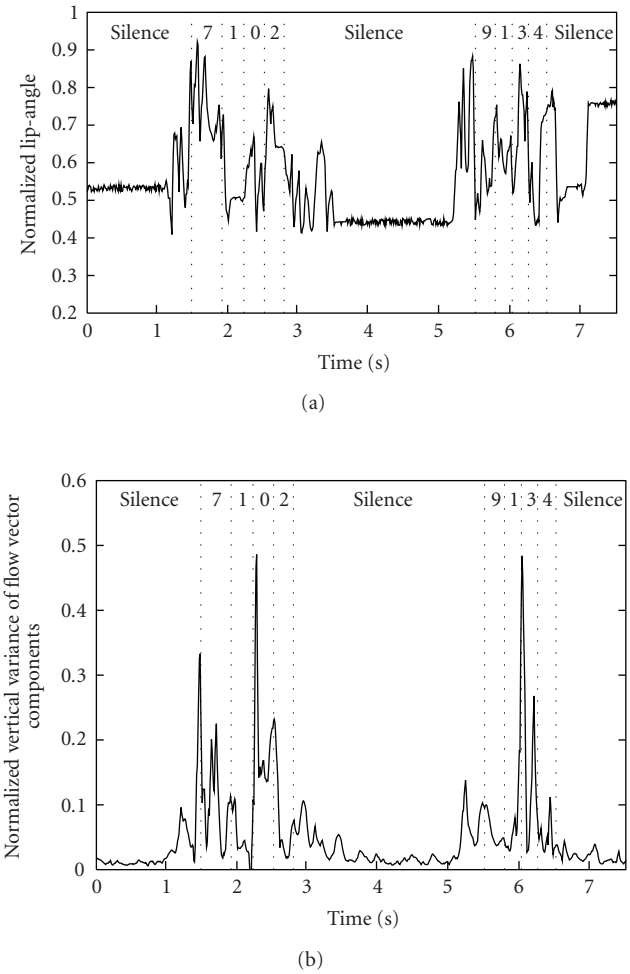


FIGURE 5: An example of a time function of (a) LCGF (normalized lip-angle value) and (b) LMVF (normalized vertical variance of optical-flow vector components).

(3) A line segment having the maximum number of points is detected as the "baseline" which is used for detecting upper and lower lip-lines. The dashed line in Figure 3(b) shows an example of the baseline.
(4) The number of points on each line segment drawn during stage 2 is counted in the binary image made by using hue values. Figure 3(c) shows an example of this binary image.
(5) Line segments with a maximum value above or below the baseline are, respectively, detected as upper or lower lip-lines. The two solid lines in Figure 3(c) indicate examples of the extracted lip-lines.

An example of the sequence of extracted lip-lines is shown in Figure 4. Finally, a lip-angle between the upper and lower lip-lines is measured.

2.4. Building LCGF vectors

The LCGF vectors, consisting of a lip-angle and its derivative (delta), are calculated for each frame and are normalized by the maximum values in each utterance. Figure 5(a) shows an example of a time function of the normalized lip-angle for a Japanese digit utterance, "7102, 9134," as well as the period of each digit. 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. As indicated by the figure, the speaker's mouth starts moving approximately 300 milliseconds before the sound is acoustically emitted. Normalized lip-angle values between 2.8 $\sim$ 3.5 seconds indicate that speaker's mouth is not immediately closed after uttering " 2 / n i /." A sequence of large lip-angle values, which appears after 7.0 seconds in Figure 5(a), is attributed to lip-lines determination errors.

3. EXTRACTION OF LIP-MOTION VELOCITY FEATURES

Our previous research [9, 10] shows that visual information of lip movements extracted by optical-flow analysis based on the Horn-Schunck optical-flow technique [15] is effective for bimodal speech recognition using frontal-face (lip) images.

Thus, the same feature extraction method [9] is applied to a bimodal speech recognition method using side-face images. The following subsections explain the Horn-Schunck optical-flow analysis technique [15] and our feature extraction method [9], respectively.

3.1. Optical-flow analysis

To apply the Horn-Schunck optical-flow analysis technique [15], image brightness at a point (x, y) in an image plane at time t is denoted by $E(x, y, t)$. Assuming that brightness of each point is constant during a movement for a very short period, the following equation is obtained:

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

If we let

$$\frac{dx}{dt} = u, \qquad \frac{dy}{dt} = v, \tag{2}$$

then a single linear equation

$$E_x \cdot u + E_y \cdot v + E_t = 0 \tag{3}$$

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:

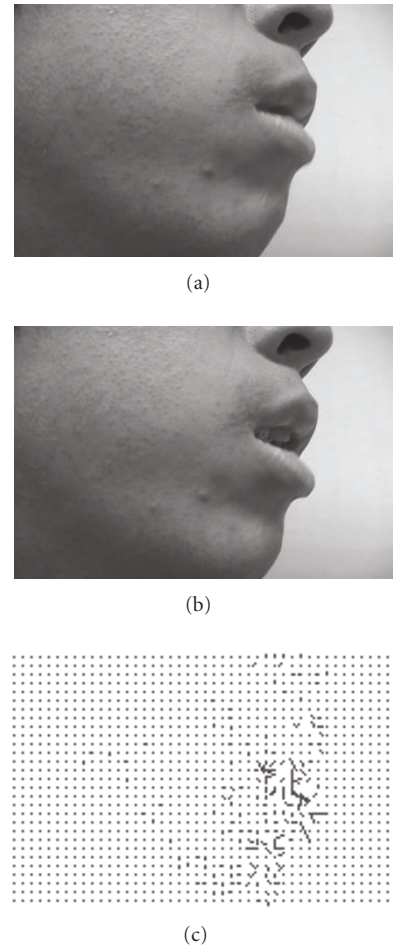
$$\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2, \qquad \left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2. \tag{4}$$

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.

3.2. Building LMVF vectors

Since (1) assumes that the image plane has a spatial gradient and that correct optical-flow vectors cannot be computed at a point without a spatial gradient, the visual signal is passed through a lowpass 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. An example of two consecutive lip images is shown in Figures 6(a) and 6(b). Figure 6(c) shows the corresponding optical-flow analysis result indicating the lip image changes from (a) to (b).

Next, two LMVFs, the horizontal and vertical variances of flow-vector components, are calculated for each frame and one 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 5(b) shows an example of a time function of the normalized vertical variance for the utterance appearing in Section 2.4. It is shown that the features are almost 0 in pause/silence periods and have large values in speaking periods. Similar to Figure 5(a), Figure 5(b) shows that the speaker's mouth starts moving approximately 300 milliseconds before the sound is acoustically emitted. It was found that time functions of the horizontal variance were similar to those of the vertical variance.



(a)

(b)

(c)

Figure 6: 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).

Finally, the two-dimensional LMVF vectors consisting of normalized horizontal and vertical variances of flow vector components are built.

4. AUDIO-VISUAL SPEECH RECOGNITION

4.1. Overview

Figure 7 shows our bimodal speech recognition system using side-face images.

Both speech and lip images of the side view are synchronously recorded. Audio signals are sampled at 16 kHz with 16-bit resolution. Each speech frame is converted into

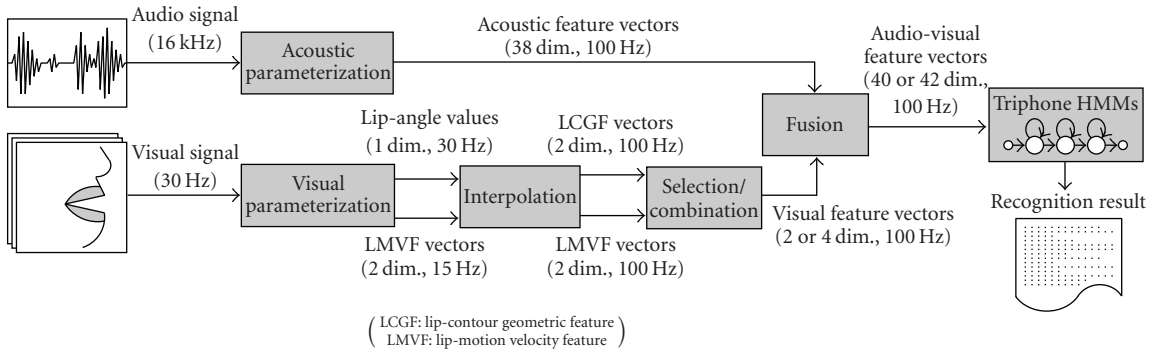


Figure 7: audio-visual speech recognition system using side-face images.

38 acoustic parameters: 12 MFCCs, 12 ΔMFCCs, 12 ΔΔMFCCs, Δ log energy, and ΔΔ log energy. The window length is 25 milliseconds. Cepstral mean subtraction (CMS) is applied to each utterance. 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 30 frames/s and 720 × 480 pixel resolution. Before computing the feature vectors, the image size is reduced to 180 × 120. For reducing computational costs of optical-flow analysis, we reduce a frame rate to 15 frames/s and transform the images to gray-scale before computing the LMVFs.

In order to cope with the frame rate differences, the normalized lip-angle values and LMVFs (the normalized horizontal and vertical variances of optical-flow vector components) are interpolated from 30/15 Hz to 100 Hz by a 3-degree spline function. The delta lip-angle values are computed as differences between the interpolated values of adjacent frames. Final visual feature vectors consist of both or either of the two features (LCGFs and LMVFs). In case that the two features are jointly used, a 42-dimensional audio-visual feature vector is built by combining the acoustic and the visual feature vectors for each frame. When using either LCGFs or LMVFs as visual feature vectors, a 40-dimensional audio-visual feature vector is built.

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

$$b_j(o_{av}) = b_{aj}(o_a)^{\lambda_a} \times b_{vj}(o_v)^{\lambda_v}, \tag{5}$$

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 streams, respectively. They are constrained by $\lambda_a + \lambda_v = 1$ $(\lambda_a, \lambda_v \geq 0)$.

4.2. Building multistream HMMs

Since audio HMMs are much more reliable than visual HMMs at segmenting the feature sequences into phonemes, audio and visual HMMs are trained separately and one combined using a mixture-tying technique as follows.

(1) The audio triphone HMMs are trained using 38-dimensional acoustic (audio) feature vectors. 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 forced alignment using the audio HMMs, and time-aligned triphone labels are obtained.
(3) The visual HMMs are trained for each triphone by four-dimensional visual feature vectors using the triphone labels obtained during 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. Figure 8 shows an example of the integration process. In this example, an audio-visual HMM for the triphone /n−a+n/ is built. The mixtures for the audio-visual HMM "`n-a+n,AV`" are tied with the audio HMM "`n-a+n,A`" and the visual HMM "`n-a+n,V`."

5. EXPERIMENTS

5.1. Database

An audio-visual speech database was collected from 38 male speakers in a clean/quiet condition. The signal-to-noise ratio (SNR) was, therefore, higher than 30 dB. Each speaker uttered 50 sequences of four connected digits in Japanese. Short pauses were inserted between the sequences. In order to avoid contaminating the visual data with noises, a gray monotone board was used as a background and speakers side-face images were captured under constant illumination conditions. The age range of speakers was 21 ∼ 30. Two speakers had facial hair.

In order to simulate the situation in which speakers would be using a mobile device with a small camera installed

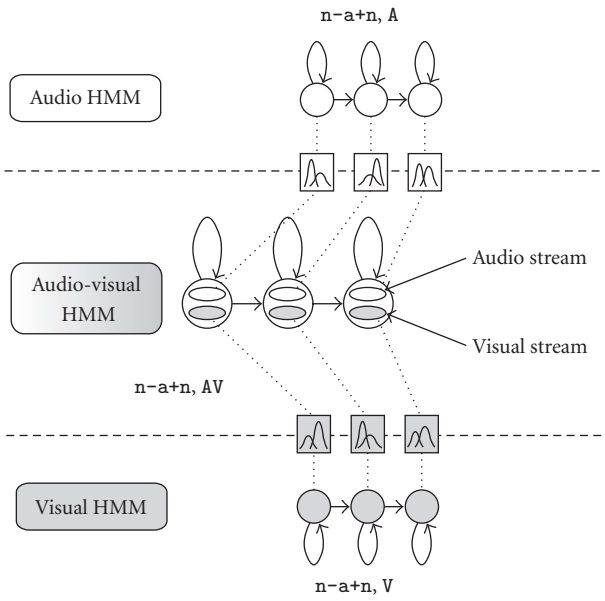


Figure 8: An example of the integration process using a mixture-tying technique to build audio-visual HMMs.

near a microphone, speech and lip images were recorded by a microphone and a DV camera located approximately 10 cm away from each speaker's right cheek. The speakers were requested to shake their heads as little as possible.

5.2. 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 20 dB. The total number of states in the audio-visual HMMs was 91. In all the HMMs, the number of mixture components for each state was set at two. Each component was modeled by a diagonal-covariance Gaussian distribution. 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. The recognition grammar was constructed so that all digits can be connected with no restrictions.

5.3. Experimental results

5.3.1. Comparison of various visual feature vectors

Table 1 shows digit recognition accuracies obtained by the audio-only and the audio-visual methods at various SNR conditions. Accuracies using only LCGFs or LMVFs as visual information are also shown in the table for comparison. "LCGF + LMVF" indicates the results using combined four-dimensional visual feature vectors. The audio and visual stream weights used in the audio-visual methods were optimized *a posteriori* for each noise condition; multiple experiments were conducted by changing the stream weights, and the weights which maximized the mean accuracy over all the 38 speakers were selected. The optimized audio stream weights (λ_a) are shown next to the audio-visual recognition accuracies in the table. Insertion penalties were also optimized for each noise condition.

In all the SNR conditions, digit accuracies were improved by using LCGFs or LMVFs in comparison with the results obtained by the audio-only method. Combination of LCGFs and LMVFs improved digit accuracies more than using either LCGFs or LMVFs, at all SNR conditions. The best improvement from the baseline (audio-only) results, 10.9% in absolute value, was obtained at the 5 dB SNR condition.

Digit accuracies obtained by the visual-only method using LCGFs, LMVFs, and the combined features "LCGF + LMVF" were 24.0%, 21.9%, and 26.0%, respectively.

5.3.2. Effect of the stream weights

Figure 9 shows the digit recognition accuracy as a function of the audio stream weight (λ_a) at the 5 dB 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) result, and others indicate the results obtained by audio-visual methods. For all the visual feature conditions, improvements from baseline are observed over a wide range of the stream weight. The range over which accuracy is improved is the largest when the combined visual features are used. It was found that the relationship between accuracies and stream weights at other SNR conditions was similar to that at the 5 dB SNR condition. This means that the method using the combined visual features is less sensitive to the stream weight variation than the method using either LCGF or LMVF alone.

5.3.3. Combination with audio-HMM adaptation

It is well known that noisy speech recognition performance can be greatly improved by adapting audio HMM to noisy speech. In order to confirm that our audio-visual speech recognition method is still effective, even after applying the audio-HMM adaptation, a supplementary experiment was performed. Unsupervised noise adaptation by the MLLR (maximum likelihood linear regression) method [16] was applied to the audio HMM. The number of regression classes was set to 8. The audio-visual HMM was constructed by integrating the adapted audio HMM and nonadapted 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 Table 1, it can be observed that accuracies are largely improved by MLLR adaptation. It can also be observed that visual features further improve the performance. Consequently, the best improvement from the non-adapted audio-only result, 30%(= 58.4%-28.4%) in absolute value at the 5 dB SNR condition, was observed when using the adapted audio-visual HMM which included the combined features.

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

SNR (dB)	Audio-only (baseline)	Audio-visual (optimized λ_a)		
		LCGF	LMVF	LCGF + LMVF
∞ (clean)	99.3%	99.3% (0.60)	99.3% (0.95)	99.3% (0.85)
20	91.5%	92.3% (0.55)	92.2% (0.60)	92.6% (0.70)
15	75.6%	79.1% (0.35)	78.7% (0.55)	79.9% (0.55)
10	51.9%	57.5% (0.30)	56.7% (0.60)	59.4% (0.45)
5	28.4%	36.4% (0.20)	34.7% (0.40)	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)		
		LCGF	LMVF	LCGF + LMVF
∞(clean)	99.5%	99.5% (0.90)	99.5% (0.90)	99.5% (0.90)
20	97.0%	97.4% (0.60)	97.2% (0.90)	97.2% (0.90)
15	91.5%	93.3% (0.55)	93.3% (0.55)	93.4% (0.70)
10	69.4%	77.2% (0.30)	76.9% (0.45)	79.5% (0.35)
5	39.5%	53.1% (0.20)	52.6% (0.30)	58.4% (0.30)

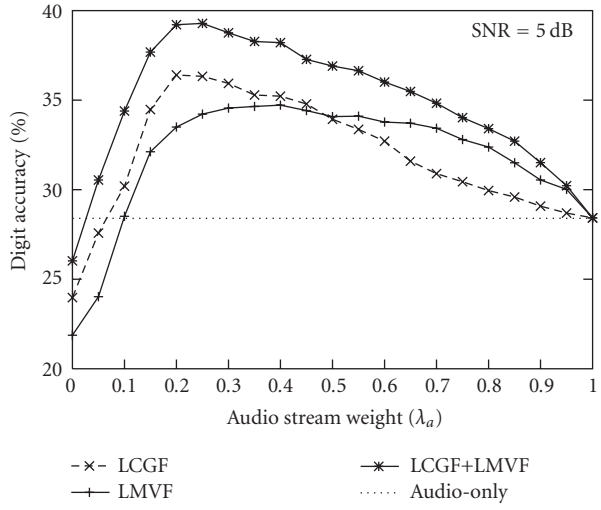


Figure 9: Digit recognition accuracy as a function of the audio stream weight (λ_a) at 5 dB SNR condition

5.3.4. Performance of onset detection for speaking periods

As another supplementary experiment, we compared audio-visual HMMs and audio HMMs in terms of the onset detection capability for speaking periods in 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 3 shows the onset detection errors in various SNR conditions. MLLR adaptation is not applied in this experiment. The optimized audio and visual stream weights decided by the experiments in Section 5.3.1 were used. Comparing the results under audio-only and audio-visual conditions, it can be found that the LMVFs, having significantly smaller detection errors than the audio-only condition, are effective in improving the onset detection. Therefore, the recognition error reduction by using the LMVFs can be attributed to the precise onset information prediction. On the other hand, the LCGFs do not yield significant improvement for onset detection in most of the SNR conditions. Since the LCGFs can also effectively increase recognition accuracies, they are considered capable of increasing the capacity to discriminate between phonemes. The increase of noise robustness in audio-visual speech recognition by combining LCGFs and LMVFs is therefore attributed to the integration of these two different effects.

5.3.5. Performance comparison of audio-visual speech recognition methods using frontal-face and side-face images

In our previous research on audio-visual speech recognition using frontal-face images [9], LMVFs were used as visual features and experiments were conducted under similar conditions to this paper; Japanese connected-digits speech contaminated with white noise was used for evaluation. Reference [9] reported that error reduction rates achieved using LMVFs were 9% and 29.5% at 10 and 20 dB SNR conditions,

Table 3: Comparison of the onset detection errors (ms) of speaking periods in various SNR conditions.

SNR (dB)	Audio-only (baseline)	Audio-visual		
		LCGF	LMVF	LCGF + LMVF
20	40.0	40.4	34.5	35.5
15	52.6	51.3	44.0	42.2
10	72.7	63.6	61.2	57.3
5	97.4	96.8	85.1	98.5

respectively. Since the error reduction rates achieved using LMVFs from side-face images were 8.8% (5 dB SNR) and 10% (10 dB SNR), it may be said that the effectiveness of LMVFs obtained from side-face images is less than that obtained from frontal-face images, although they cannot be strictly compared because the set of speakers was not the same for both experiments. Lucey and Potamianos compared audio-visual speech recognition results using profile and frontal views in their framework [17], and showed that the effectiveness of visual features from profile views was inferior to that from frontal views.

It is necessary to evaluate the side-face-based and frontal-face-based methods from the human-interface point of view, to clarify how much the ease-of-use advantages of the side-face-based method described in the introduction could compensate for the method's performance inferiority to frontal-face-based approaches.

6. CONCLUSIONS

This paper has proposed audio-visual speech recognition methods using lip information extracted from side-face images, focusing on mobile environments. The methods individually or jointly use lip-contour geometric features (LCGFs) and lip-motion velocity features (LMVFs) as visual information. This paper makes the first proposal to use LCGFs based on an angle measure between the upper and lower lips in order to characterize side-face images. Experimental results for small vocabulary speech recognition show that noise robustness is increased by combining this information with audio information. The improvement was maintained even when MLLR-based noise adaptation was applied to the audio HMM. Through the analysis on the onset detection, it was found that LMVFs are effective for onset prediction and LCGFs are effective for increasing the phoneme discrimination capacity. Noise robustness may be further increased by combining these two disparate features.

In this paper, all evaluations were conducted without considering the effects of visual noises. It is necessary to evaluate the effectiveness/robustness of our recognition method on a real-world database containing visual noises. Our previous research on frontal-face images [11] showed that lip-motion features based on optical-flow analysis improved the performance of bimodal speech recognition in actual running cars. The lip-angle extraction method investigated in this paper might be more sensitive to illumination conditions, speaker variation, and visual noises. Therefore, this method also needs to be evaluated on a real-world database. Feature normalization techniques, in addition to the maximum-based method used in this paper, also need to be investigated in real-world environments. Developing an automatic stream-weight optimization method is also an important issue. For frontal images, several weight optimization methods have been proposed [8, 18, 19]. We have also proposed weight optimization methods and confirmed their effectiveness by experiments using frontal images [20, 21]. It is necessary to apply these weight optimization methods to the side-face method and evaluate the resulting effectiveness. Future works also include (1) evaluating the lip-angle estimation process using manually labeled data, (2) evaluating recognition performance using more general tasks, and (3) improving the combination method for LCGFs and LMVFs.

ACKNOWLEDGMENT

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

REFERENCES

[1] C. Bregler and Y. Konig, ""Eigenlips" for robust speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '94)*, vol. 2, pp. 669–672, Adelaide, SA, Australia, April 1994.

[2] M. J. Tomlinson, M. J. Russell, and N. M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, vol. 2, pp. 821–824, Atlanta, Ga, USA, May 1996.

[3] G. Potamianos, E. Cosatto, H. P. Graf, and D. B. Roe, "Speaker independent audio-visual database for bimodal ASR," in *Proceedings of ESCA Workshop on Audio-Visual Speech Processing (AVSP '97)*, pp. 65–68, Rhodes, Greece, September 1997.

[4] C. Neti, G. Potamianos, J. Luettin, et al., "Audio-visual speech recognition," Final Workshop 2000 Report, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, Md, USA, October 2000.

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

[6] Y. Zhang, S. Levinson, and T. S. Huang, "Speaker independent audio-visual speech recognition," in *Proceedings of IEEE International Conference on Multi-Media and Expo (ICME '00)*, pp. 1073–1076, New York, NY, USA, July-August 2000.

[7] S. M. Chu and T. S. Huang, "Bimodal speech recognition using coupled hidden Markov models," in *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP '00)*, vol. 2, pp. 747–750, Beijing, China, October 2000.

[8] C. Miyajima, K. Tokuda, and T. Kitamura, "Audio-visual speech recognition using MCE-based HMMs and model-dependent stream weights," in *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP '00)*, vol. 2, pp. 1023–1026, Beijing, China, October 2000.

[9] K. Iwano, S. Tamura, and S. Furui, "Bimodal speech recognition using lip movement measured by optical-flow analysis," in *Proceedings of International Workshop on Hands-Free Speech Communication (HSC '01)*, pp. 187–190, Kyoto, Japan, April 2001.

[10] S. Tamura, K. Iwano, and S. Furui, "Multi-modal speech recognition using optical-flow analysis for lip images," *Journal of VLSI Signal Processing—Systems for Signal, Image, and Video Technology*, vol. 36, no. 2-3, pp. 117–124, 2004.

[11] S. Tamura, K. Iwano, and S. Furui, "Improvement of audio-visual speech recognition in cars," in *Proceedings of the 18th International Congress on Acoustics (ICA '04)*, vol. 4, pp. 2595–2598, Kyoto, Japan, April 2004.

[12] T. Yoshinaga, S. Tamura, K. Iwano, and S. Furui, "Audio-visual speech recognition using new lip features extracted from side-face images," in *Proceedings of COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction (Robust '04)*, Norwich, UK, August 2004.

[13] T. Yoshinaga, S. Tamura, K. Iwano, and S. Furui, "Audio-visual speech recognition using lip movement extracted from side-face images," in *Proceedings of International Conference on Audio-Visual Speech Processing, ISCA Tutorial and Research Workshop (AVSP '03)*, pp. 117–120, St. Jorioz, France, September 2003.

[14] A. C. Bovik and M. D. Desai, "Basic binary image processing," in *Handbook of Image and Video Processing*, A. C. Bovik, Ed., pp. 37–52, Academic Press, San Diego, Calif, USA, 2000.

[15] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1–3, pp. 185–203, 1981.

[16] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[17] P. Lucey and G. Potamianos, "Lipreading using profile versus frontal views," in *Proceedings of the 8th IEEE Workshop on Multimedia Signal Processing (MMSP '06)*, pp. 24–28, Victoria, BC, Canada, October 2006.

[18] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98)*, vol. 6, pp. 3733–3736, Seattle, Wash, USA, May 1998.

[19] S. Nakamura, H. Ito, and K. Shikano, "Stream weight optimization of speech and lip image sequence for audio-visual speech recognition," in *Proceedings of 6th International Conference on Spoken Language Processing (ICSLP '00)*, vol. 3, pp. 20–24, Beijing, China, October 2000.

[20] S. Tamura, K. Iwano, and S. Furui, "A stream-weight optimization method for audio-visual speech recognition using multi-stream HMMs," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 1, pp. 857–860, Montreal, Quebec, Canada, May 2004.

[21] S. Tamura, K. Iwano, and S. Furui, "A stream-weight optimization method for multi-stream HMMs based on likelihood value normalization," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 1, pp. 469–472, Philadelphia, Pa, USA, March 2005.