13—7 A Set of Mesh Features for Automatic Visual Speech Recognition

Hyun-Hwa Oh, Young-Mi Jeoun, Sung-Il Chien [1]
School of Electronic and Electrical Engineering, Kyungpook National University

Abstract

This paper proposes three types of mesh features to analyze the lip movements of the speakers, which show much of spatio-temporal variability. These features are extracted from the region of interest around the lips, which is divided into the same shape meshes, the overlapping meshes considering the boundary problem, or the variable shape meshes properly reflecting the characteristic lip movements. The left and right corner points of lips are automatically detected from an image sequence by using a coarse-to-fine processing based on the binarization and connected component analysis, thereby effectively determining the region of interest to be partitioned into mesh features. The recognition performances of these mesh features are empirically evaluated through a series of experiments on visually recognizing seven Korean vowels by implementing a lipreading system based on the discrete hidden Markov models. The experimental results show that the proposed mesh features are quite efficient for characterizing the lip shapes and movements. Specially, the mesh feature by variable shape meshes achieves a noticeable recognition performance for lipreading.

1 Introduction

Humans, especially the hearing impaired, utilize visual information, lipreading or speechreading, for the improved accuracy of speech recognition, because the lip movements provide valuable information for the visual categorization of speech. In the McGurk effect [1], it was shown that human perception of acoustic speech is affected by the visual cues of the lip movements. Therefore, several successful attempts [2-6] on utilizing the visual information of speech have been made to construct a noise-tolerant speech recognition system in highly noisy environment.

In order to improve the performance of lipreading, it is quite necessary to extract the robust visual speech features, which are considered to well represent the dynamic characteristics of the lip movements during the utterance of a speaker. Most of the previous approaches have utilized the image-based method [2][3] or the model-based method [4][5] to extract the visual speech information from the given image sequences. In the first approach, the image intensities are measured and employed as a feature vector after performing a certain preprocessing usually including filtering and dimension reduction for an input image. It is quite sensitive to the variations in illumination; if the intensity or direction of illumination changes, all the pixel values vary, thereby resulting in the degradation of the recognition performance. In the second approach, a model representing the lip contours is built and the spatio-temporal parameters of the model are estimated for the visual analysis of speech signal. Usually this method requires the small set of parameters to describe the lip model but encounters substantial difficulties in estimating exact parameter values.

In this paper, we introduce a set of visual speech features based on the mesh for efficient lipreading of Korean vowels. The effectiveness of mesh feature in dealing with spatial complexity of signals has been successfully proven in many pattern recognition based applications, especially in character recognition [9]. Three types of mesh features are extracted from the region of interest (ROI) in each frame of a lip image sequence, according to the mesh shapes: the same shape of basic meshes, overlapping meshes to overcome the boundary problem occurring on the boundary of the basic meshes, and variable shape meshes specially designed to reflect directional characteristics of the lip movements.

The ROI is determined using the distance between the left and right lip corner points, and the lip thickness of a speaker. Thus, the exact lip corner points and thickness should be firstly detected for automatic extraction of the mesh feature representing properly the spatio-temporal information of the lip movements. We propose an efficient method of finding lip corners by a coarse procedure including binarization and connected component analysis (CCA) [7] to estimate initial position of the corner points and a fine-tuning procedure to locate their exact positions. The lip thickness is simply determined by searching a certain region bounded by the estimated lip corners. In recognition experiments on seven Korean vowels, we evaluate the performances of the mesh features by implementing a multi-speakers lipreading system based on the discrete hidden Markov models (HMMs).

2 Database for Lipreading

Our ultimate interest is to build a multi-speakers lipreading system, which is also able to supplement the discrimination capability of the automatic speech recognition (ASR) system. So, our audio-visual speech database for the Korean vowels has been simultaneously collected under the natural room lighting and pronunciation condition. However, since this paper focuses on investigating the effectiveness of the visual features based on mesh for Koran vowel lipreading, only the visual database describing the lip shapes and movements in the Korean utterance is considered now.

The Korean language is composed of 21 vowels (V) and 19 consonants (C). The Korean syllable types are V, CV, VC, and CVC. It is quite notable that particularly for the Korean language, the lip shapes of a speaker depend dominantly on the vowel constituting a syllable except the consonants of /b/, /p/, and /m/. Thus, in this paper, we concentrate our focus on recognizing these vowels visually. We have clustered Korean vowels into seven groups according to their lip shapes and defined these groups as the Korean vowel visemes, which are /a/, /æ/, /ə/, /o/, /u/, /ɨ/,

[1] Address: 1370 Sangyeok-dong, Buk-gu, Daegu, 702-701, Korea. E-mail: sichien@ee.knu.ac.kr

and /i/. Here, the viseme is the smallest unit of visible speech [8].

Our database consists of 700 image sequences obtained from 10 speakers (7 males and 3 females), each repeating the isolated seven Korean vowels ten times, respectively. A sequence starts with the lips being closed and continues until it returns to the original place after utterance. The number of frames composing a sequence varies according to speakers and vowels. All image sequences are acquired at 30 frames/sec and the resolution is 320×240 pixels with 8-bit gray levels. Figure 1 shows three examples of lip image sequences representing /a/, /o/, and /i/.

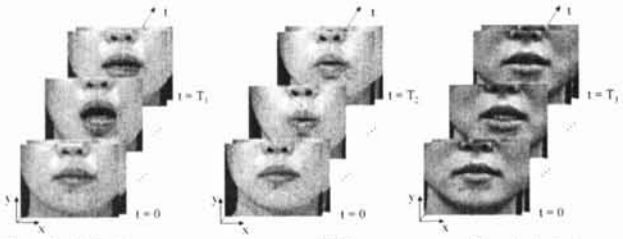


Fig. 1. Lip image sequences of Korean vowels, /a/, /o/, and /i/.

3 Detection of Lip Corner Points and ROI

For automatic extraction of the mesh features, the exact lip region, ROI should be firstly localized on every frame of an image sequence. We find that region using lip corner points, the maximum extent of lip opening, and thickness which are detected by utilizing the gray-level intensity of the lips. A detailed procedure is described below.

3.1 Detection of Lip Corner Points

As the first step for detecting the lip corner points, we define the initial ROI including the lips at the first frame of the given image sequence. Considering that the gray-level intensity of the lips is generally lower than that of the facial skin, the initial ROI is determined from the horizontal and vertical projections of the binary image for the first frame. The binary image is produced by Otsu method after histogram equalization of the gray level image.

We obtain several labeled regions by applying the CCA method to the binary image of the initial ROI on the first frame. As shown in Fig. 2a, the left and right points of the biggest labeled region are appointed as the initial positions for the corresponding lip corner points. Actually, these positions can be deviated from the real lip corner points, since the region labeling technique depends on the illumination condition and threshold level for binarization. Therefore, a fine-tuning procedure is introduced to locate exact corner points. We put the $n \times n$ square window around the initial position and then perform above-mentioned binarization and region labeling again within the region. In our experiments, n is set to 60. Finally, as shown in Fig. 2b, exact corner points are acquired from the two readjusted regions. These corner points are not always on the same horizontal axis because a speaker's head is slightly tipped. Thus, we calculate the tip angle θ, as shown in Fig. 2b, and rotate the image with respect to the center point (x_c, y_c) of lips to locate corner points on the same horizontal axis.

3.2 Extraction of ROI from Lip Image Sequences

Now, we determine the optimum size of the ROI containing the lip region but excluding the surrounding facial region as much as possible even when the lips are fully opened. The width w of the ROI is fixed as the distance between two lip corner points detected at the first frame. The height h of the ROI is determined by the lip thickness and the maximum extent of the lip opening, which vary from person to person.

Under the natural illumination condition and without any special marker on the lips, the outer edge of the lower lip is generally seen to be less clear than that of the upper lip. Thus, the lip thickness of a speaker is determined using the ratio of the thickness of the upper lip to the lip width on the first frame. If the ratio is within the upper and lower bounds statistically determined, the lips are classified as the normal one. If the ratio excesses the upper bound or is under the lower bound, then the lips are determined as thick or thin, respectively. Let h_u and h_l are the heights of the upper and lower parts of the ROI, respectively, the height h of ROI is given by

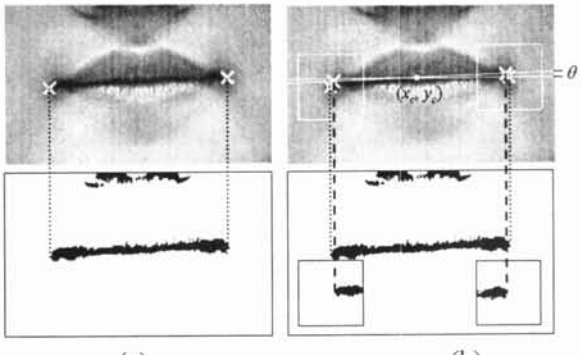
$$h = h_u + h_l$$

where, $h_l = w/2 - c$,

$$h_u = \begin{cases} h_l - \Delta y_{down} & \text{for thin lip} \\ h_l & \text{for normal lip} \\ h_l + \Delta y_{up} & \text{for thick lip.} \end{cases} \quad (1)$$

Here, $\Delta y_{down} = w/a_1$ and $\Delta y_{up} = a_2 / w$ and c, a_1 and a_2 are constant values. We adjust the upper height h_u depending on the lip thickness of a speaker by Eq. (1). Specifically, as for the thin lips, h_u is reduced as much as Δy_{down} to prevent the facial and nose regions intruding into the ROI. In the case of the thick lips, expanding of h_u as much as Δy_{up} prevents part of the upper lip being removed from the ROI when the lips are fully opened.

There is not much movement of the lips between two consecutive frames, as the image sequence of our database is obtained at 30 frames/sec. Thus, two lip corner points of the current frame are found by using the binarization and CCA within the local regions around those points detected on the previous frame. We compensate the rotational variation of the lip region in each frame, which is mainly caused by the speaker's head movement during the utterance. Finally, the sequence of features is consistently extracted by positioning the ROI on the center of the lip corners in each frame without any changes in height and width which have been determined at the first frame. Figure 3 shows the results of the detection of lip corner points and the ROI on the image sequence of /o/ in which frames are rotated by the tip angles θ_i, respectively.



(a) (b)

Fig. 2. Lip corner points detection within initial ROI on the first frame; (a) Step 1: roughly detected lip corner points, (b) Step 2: exactly detected corner points using fine-tuning procedure, center point (x_c, y_c) and tip angle θ of lips.

Frame: 1 30 35 40 44 46

Fig. 3. An example of showing successive detection of lip corner points (represented by black cross) and ROI (represented by rectangle) on image sequence of Korean vowel, /o/.

4 Extraction of Mesh Features

Now, we introduce the sequence of mesh feature vectors extracted from the corresponding image sequence as an input to analyze the lip movements of a speaker, which shows much of spatio-temporal variability. In our experiment, three types of mesh features are designed and extracted from the ROI of a lip image according to the shapes of a mesh.

In the first case as shown Fig. 4a, the upper and lower heights of the ROI are equally divided into four segments, respectively, and the width of the ROI is also equally divided into eight segments. Accordingly, the ROI consists of the same shape of 64 basic meshes. The mesh features are extracted from the gradient magnitude map of such ROI, since the gradient of gray-level image is more insensitive to an illumination. The mesh feature vector $\mathbf{f}_{\mathrm{BM}}^{(i)}$ is defined by

$$\mathbf{f}_{\mathrm{BM}}^{(i)} = [\,\bar{g}_1^{(i)}\ \bar{g}_2^{(i)}\ \cdots\ \bar{g}_M^{(i)}\,]^T, \quad i=1, 2, \ldots, T$$
$$\text{where,}\quad \bar{g}_j^{(i)} = \frac{1}{N_b}\sum_{I=1}^{N_b}\|\nabla I_I\|, \tag{2}$$

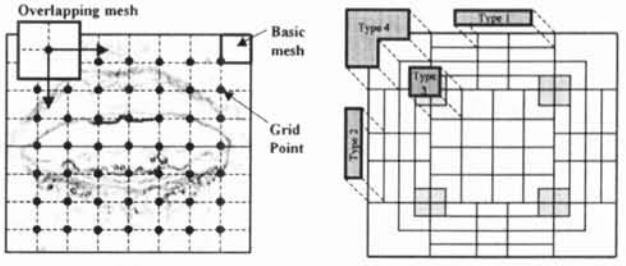
where T is the total number of frames and $\bar{g}_j^{(i)}$ is the average value of gradient magnitude $\|\nabla I_I\|$ within the jth basic mesh at the ith frame. Also, M and N_b denote the number of basic meshes in the ROI and the number of pixels within the basic mesh, respectively.

In each frame of an image sequence, two lip corner points are not always located at the exact same positions because the gray-level intensity of each frame changes frequently due to the speaker's lip movements or head-movement during the utterance. Hence, $\bar{g}_j^{(i)}$ in Eq. (2) also can vary severely from frame to frame when the lip contour exists on the boundary of basic meshes. To overcome such boundary problem, we obtain the additional mesh features by applying the bigger overlapping mesh consisting of 2×2 basic meshes to each grid of the ROI as shown in Fig. 4a. We now define the augmented feature vector $\mathbf{f}_{\mathrm{OM}}^{(i)}$ at ith frame by adding $\mathbf{f}_{\mathrm{BM}}^{(i)}$ with the average values of the gradient magnitude inside the overlapping meshes.

$$\mathbf{f}_{\mathrm{OM}}^{(i)} = [\,\bar{g}_1^{(i)}\ \bar{g}_2^{(i)}\ \cdots\ \bar{g}_M^{(i)}\ \bar{g}_{M+1}^{(i)}\ \cdots\ \bar{g}_O^{(i)}\,]^T,$$
$$\bar{g}_j^{(i)} = \frac{1}{K}\sum_{I=1}^{K}\|\nabla I_I\|, \tag{3}$$
$$\text{where}\quad K = N_b \quad \text{if } 1 \le j \le M$$
$$K = N_o \quad \text{if } M+1 \le j \le O.$$

Here, N_o is the number of pixels within the overlapping mesh and O is 113, which is the total number of basic meshes and overlapping meshes within the ROI.

It is found that the horizontal central parts of the upper and lower lips mainly move in the vertical direction and the left and right corner parts of the lips move largely in the horizontal direction during the utterance [10]. Now, we design another mesh structure imposed on the ROI by using four different shapes of meshes as shown in Fig. 4b to represent properly such characteristic lip movements. Type 1 mesh is designed to reflect the vertical moving style of the lips more faithfully and is located at the upper horizontal and lower horizontal central parts of the ROI. On the contrary, type 2 mesh is put on the left vertical and right vertical central parts of the ROI to consider the horizontal movements of lip corner regions. For the inner and diagonal parts of the ROI, type 3 mesh is an aforementioned basic mesh itself. Lastly, type 4 mesh is located at four corner parts of the ROI which are expected to contain relatively less visual information related to the lip movements when compared to other parts. Finally, we average the gradient magnitude inside each type mesh at every frame and define them as feature vector $\mathbf{f}_{\mathrm{VM}}^{(i)}$ whose dimension is 60.



(a) (b)

Fig. 4. Mesh features from ROI; (a) basic mesh, grid points, and overlapping mesh, (b) variable shape meshes.

5 Experimental Results

Now, effectiveness of the proposed mesh features is evaluated by implementing a HMM based lipreading system to recognize seven Korean vowels. An HMM can be expressed compactly as $\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$. Here $\boldsymbol{A}$ represents the state transition probability matrix, $\boldsymbol{B}$ the observation symbol probability matrix, and $\boldsymbol{\pi}$ the initial state probability vector. The parameters of an HMM have been estimated using the Baum-Welch algorithm. The Viterbi algorithm has been used to evaluate the trained HMM at the time of recognition. A vector quantization procedure was also conducted in advance using the LBG algorithm to build the codebook consisting of 64 fixed prototype vectors. Thus, each feature vector is converted into one of 64 symbols for a discrete HMM, which follows immediately. For more detailed references on the theory, computation, and applications of HMM, the readers are referred to [11].

We constructed three sets of HMMs according to the types of mesh features and compared their performances using the testing database described in Section 2. Twelve states were commonly assigned to all HMMs. Clearly, our experimental results could not be directly compared with those reported elsewhere in the literature because of the difference of the databases used. Nevertheless, from the result shown in the Table 1, it can be found that the proposed mesh features show fairly good recognition rates for multi-speakers lipreading when compared with results from the previous research [3-6]. Here, if the recognition result for a given input is determined only by the HMM with the maximum probability, the recognition rate is recorded as Top 1. If we consider two HMMs which produce the first and second maximum probabilities, Top 2 is considered as the recognition performance. Note that the recognition rate of HMMs of $\mathbf{f}_{\mathrm{OM}}$ is about 2% better in terms of Top 1 than that of HMMs of $\mathbf{f}_{\mathrm{BM}}$, though they show similar recognition performance by Top 2. Such re-

sult demonstrates that the overlapping mesh provides a quite effective means of overcoming the deterioration in performance mainly due to the boundary problem occurring between the basic meshes. More improvement in recognition performance could be achieved by employing the feature vector $\mathbf{f}_{VM}$ designed to emphasize characteristic lip movements in the horizontal and vertical direction. Thus, we can conclude that the $\mathbf{f}_{VM}$ extracted from the ROI by four types of variable meshes shows relatively good discriminating power to cope with the spatio-temporal characteristics of the lip image sequences for Korean vowels.

Table 2 shows a confusion matrix for the recognition result of $\mathbf{f}_{VM}$. It can be seen from this table that most of classification errors occur internally within three groups which are /a/-/æ/-/ə/, /o/-/u/, and /ɨ/-/i/. In addition, it is occasionally found that vowel /i/ is misclassified as /æ/. These grouping and misclassifications of seven vowels have also happened when you recognize these vowels visually. Usually, vowel utterance requires much simpler lip movements when compared to isolated or continuous words utterance, which causes severe shape confusion among the vowels, specially, when those vowels are pronounced from the multi-speakers. Thus, it is analyzed that most of these errors were the results of such confusion combined with a 2D spatial distortion mainly due to the unconstraint utterance condition allowing naturalness to each speaker.

Table 1. Recognition rates of three mesh features for lipreading

Mesh features	Recognition rate (%)	
	Top 1	Top 2
$\mathbf{f}_{BM}$	58.6	79.7
$\mathbf{f}_{OM}$	60.8	79.7
$\mathbf{f}_{VM}$	62.6	81.4

Table 2. Confusion matrix for recognition result of $\mathbf{f}_{VM}$ with four types of meshes in Top 1

Vowels	/a/	/æ/	/ə/	/o/	/u/	/ɨ/	/i/	Recognition Rate (%)
/a/	36	10	3	1	0	0	0	72.0
/æ/	13	29	4	1	0	1	2	58.0
/ə/	12	1	29	2	2	1	3	58.0
/o/	2	1	1	37	8	1	0	74.0
/u/	1	0	3	13	32	0	1	64.0
/ɨ/	1	1	5	0	2	28	13	56.0
/i/	2	12	1	1	2	4	28	56.0
Average								**62.6**

6 Conclusions and Future Work

Three types of mesh features have been proposed to analyze the visual information of speech signal. These features are extracted by partitioning the ROI into same shape of basic meshes, overlapping meshes to cut down the boundary problem appearing between two basic meshes, or variable shape meshes to reflect the typical directional trend of lip movements. An efficient coarse-to-fine procedure has been also introduced to detect exact two lip corner points, whereby the optimum position and size of the ROI are calculated even if the slight rotational variations in the lip region exist. In lipreading experiments on seven Korean vowels, we evaluated the recognition performance of the proposed mesh features by implementing a multi-speakers lipreading system based on the discrete HMMs. The experimental results show that these mesh features, specially the variable shape mesh feature achieves the impressive recognition performance when compared to the results of previous research works. In the future work, we will apply these mesh features to the isolated or continuous word recognition and will combine our visual technique with acoustic speech recognition system to enhance the recognition performance in a noisy environment.

Acknowledgement

This work was supported by grant No.(R01-1999-00233) from the Basic Research Program of the Korea Science & Engineering Foundation.

References

[1] H. McGurk and J. MacDonald, "Hearing Lips and Seeing Voices," Nature, vol. 264, pp. 746-748, 1976.

[2] P. Duchnowski, M. Hunke, D. Busching, U. Meier, and A. Waibel, "Toward Movement-Invariant Automatic Lip-Reading and Speech Recognition," Proc. of IEEE Int'l Conf. on Acoustics, Speech and Signal Processing, vol. 1, pp. 109-112, 1995.

[3] G. Potamianos, H.P. Graf, and E. Cosatto, "An Image Transform Approach for HMM Based Automatic Lipreading," Int'l Conf. on Image Processing, vol. 3, pp. 173-177, 1998.

[4] M.E. Hennecke, K.V. Prasad, and D.G. Stork, "Automatic Speech Recognition System Using Acoustic and Visual Signals," Proc. of 29th Asilomar Conf. on Signals, Systems and Computers, vol. 2, pp. 1214-1218, 1995.

[5] J. Luettin and N.A. Thacker, "Speechreading using Probabilistic Models," Computer Visions and Image Understanding, vol. 65, no. 2, pp. 163-178, 1997.

[6] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, and R. Harvey, "Extraction of Visual Features for Lipreading," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 2, pp. 198-213, 2002.

[7] R.C. Gonzalez and R.E. Woods, Digital Image Processing, Addison Wesley, 1992.

[8] C. Fisher, "Confusions Among Visually Perceived Consonants," Journal of Speech and Hearing Research, vol. 11, pp. 796-804, 1968.

[9] E. Sackinger, B.E. Boser, J. Bromley, Y. LeCun, and L.D. Jackel, "Application of the ANNA Neural Network Chip to High-Speed Character Recognition," IEEE Trans. Neural Networks, vol. 3, no. 3, pp. 498-505, 1992.

[10] H.-H. Oh, I.-C. Kim, and S.-I. Chien, "Experiments on Various Visual Speech Features for Korean Vowel Lipreading," Proc. of IASTED Int'l Conf. on Signal and Image Processing, pp. 225-230, Honolulu, Hawaii, USA, Aug. 2001.

[11] X.D. Huang, Y. Ariki, and M.A. Jack, Hidden Markov Models for Speech Recognition, Edinburgh University Press, Edinburgh UK, 1990.