PROFILE VIEW LIP READING

Kshitiz Kumar, Tsuhan Chen and Richard M. Stern

Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213 USA
Email: {kshitizk,tsuhan,rms}@ece.cmu.edu

ABSTRACT

In this paper, we introduce profile view (PV) lip reading, a scheme for speaker-dependent isolated word speech recognition. We provide historic motivation for PV from the importance of profile images in facial animation for lip reading, and we present feature extraction schemes for PV as well as for the traditional frontal view (FV) approach. We compare lip reading results for PV and FV, which demonstrate a significant improvement for PV over FV. We show improvement in speech recognition with the integration of Audio and Visual features. We also found it advantageous to process the visual features over a longer duration than the duration marked by the endpoints of the speech utterance.

Index Terms— Speechreading, Visual feature extraction, Audiovisual speech recognition, Profile view

1. INTRODUCTION

Lipreading has been shown to improve speech recognition accuracy in noisy environments [1][2] and also is a component technology for multimedia phones for hard-of-hearing people [3]. Most past research in lip reading has been confined to frontal face lip reading, largely forgetting that features from the profile view (PV) can be important for lip reading as well. PV lip reading also has applications in speech recognition using cell phones, as a camera mounted on cell phone can detect PV features while user is talking, while traditional frontal view (FV) features are not easily extracted.

We provide motivation for PV lip reading from earlier research in speech visual synthesis in [4][3]. A media transformation from speech to 3-D wire frame model was shown in [4]. This 3-D model enabled frontal as well as side view, thus improving overall lip reading visualization. Improvement with side view was later conclusively reported in [3], where the authors obtained significant improvement in representation of visual cues with simultaneous PV and FV images. Our inspiration for PV lip reading is this enhanced visualization with side images. Side images have thus helped humans do better lip reading in the form of improved visual cues and in our present work, we plan to extend the benefit of side images for humans to machines in better lip reading.

While some recent studies [5][6] have considered PV lip reading, these studies were based on digit recognition, which is visually less confusable [2]. In addition, both of these studies used image based features, which are essentially information deficient when compared to shape based features for lip reading. Specifically, optical flow of mouth image was used in [5], and DCT features were used in [6]. Optical flow still captures some information in the form of lip movement information but it is prone to rotation and illumination. In case of DCT based features, usually a few top-energy DCT coefficients are picked up but they bear no physical significance to lip reading task. DCT features have been shown to perform worse than geometrical features in [7]. Our work is novel in that it develops PV lip reading using meaningful geometrical features of lip height, width and protrusion, rather than heuristic features. In particular, our work is the first to report results with lip protrusion features which are absent in FV and hence specific to PV images. Interestingly, we observed that lip protrusion features can provide better accuracy than lip height or width features.

In Section 2, we describe our Audio-Visual data. In Section 3, we discuss feature extraction steps for the PV and Section 4 does the same for the FV. Section 5 presents in which we observe that the word error rate (WER) obtained with PV is lower than that obtained with the FV. We also describe results obtained by combining PV and FV features. We also exploit the fact that lip cues start earlier and end later than audio cues.

2. DATA COLLECTION

Since this is the first attempt to develop PV recognition for the present task, a new database needed to be collected, which we call the "CMU Audio-Visual Profile and Frontal View" (CMU AVPFV) database. This database consists of simultaneously-recorded profile and frontal view audio and visual data in a soundproof IAC studio. Video was recorded in VGA(640*480) resolution at 30 fps. Our vocabulary consists of 150 words from the Modified Rhyme Test (MRT) [8], which is widely used in speech intelligibility testing. In MRT list, all the words are of the pattern consonant-vowel-consonant. This composition of MRT vocabulary helps highlight the confusability of a consonant keeping the rest of word fixed. We collected data from 10 subjects, with each subject repeating the 150-word MRT list 10 times.

3. PROFILE VIEW FEATURE EXTRACTION

In this initial analysis of profile view lip reading, we are especially interested in the development of lip features which can be valuable for speech recognition and at the same are simple to work with. We considered lip protrusion and lip height parameters as potentially valuable, as protrusion parameters describe forward and backward movement of the lips while height parameters describe upward and downward movement. We extracted four features from the subjects

images, two for lip protrusions (one each for the upper and lower lip) and, similarly, two for lip height.

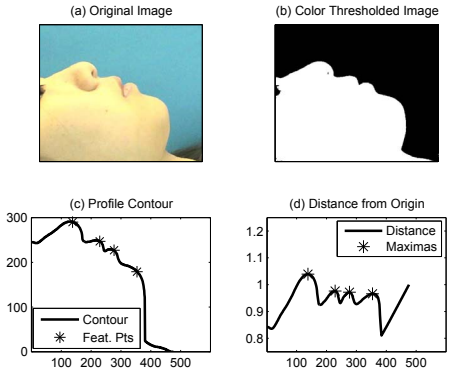

Fig. 1. Profile View feature extraction steps

Figures 1 and Fig. 2 illustrate the the steps involved in PV feature extraction. To obtain our feature vectors, we first aim to isolate the profile contour from the subject's images. We define the profile contour as the boundary between the subject's PV face region and background. Isolating this profile contour is important as it carries all the information for measuring geometrically the desired protrusion and height features.

We employ color thresholding in the red channel of the RGB image to obtain the profile contour. The human face is usually rich in red color. We chose a blue background screen for the data collection process making the background deficient in red color, which in turn enabled us to perform red channel thresholding on the image. This thresholding produces a binary image, and the profile contour is defined to be the boundary between black and white pixels. The red-thresholded image will in general have noise in the form of a few patches of white pixels corresponding to the background region in the RGB image, which we remove by removing all white regions having area less than the connected white region with greatest area. (Our assumption, which is justified empirically) is that this region represents the face region. Fig. 1(b) is an example of the result of color thresholding and removing white pixels in the fashion described above. The profile contour, shown in Fig. 1(c), is easily identified as the first boundary between black and white pixels encountered when traversing the image from its top side.

We note that the profile contour has all the information necessary to measure the lip protrusion and lip height parameters geometrically. We now describe our methods for extracting lip height and lip protrusion features from the PV. In extracting the lip height parameters, we first define lip center as center point in lip and lip corner as its corner point, see Fig. 3. Next define lip center line as the line joining lip centers and then a height reference point as point of intersection of lip center line and perpendicular from lip corner to lip center line. Note that in profile face, only one lip corner is visible. The lip height parameter for the upper and lower lip respectively is defined to be the Euclidian distance between their center positions and the height reference point. We conclude that for the height parameters, we need to detect the lip centers and lip corner.

In defining lip protrusion we first need to establish a static reference frame with respect to which we can measure the forward-backward movement of the lips. We cannot use the bounding box of the image as a reference frame because the subject may be moving in this frame, thus adding noise to the actual forward/backward movement of lips. We propose the line joining the tip of the nose and the center of the chin as a static frame for finding protrusion parameters. This nose-chin frame is robust with respect to motion of the subject's face because nose, lips and chin are part of the face and they undergo the same transformation with respect to a subject's facial translation and rotation. We define lip protrusion parameters as the Euclidian perpendicular distance between the center lips and the nose-chin line defined above. Hence we have reduced our feature extraction problem to one of first detecting the tip of nose, centers of the lips and chin and the lip corner, and then obtaining the required distances.

We refer to the tip of the nose, and the centers of the lips and chin as feature points, and we initially hypothesize them to be the local maxima in the profile contour. Specifically, the first four local maxima of the profile contour in the horizontal direction are assumed to be the nose, upper and lower lip, and chin feature points. The lip corner is assumed to be the local minimum between the two lips. We emphasize that these locations are only initial hypotheses because in general all feature points in a profile contour may not have a local maxima. For example, the chin is not a local maximum in Fig. 1(c). Similarly, the centers of the chin and lips may also be locally concave without having a local maxima. Thus searching for feature points as local maxima in the profile contour is inappropriate. For this reason we propose a distance transformation which maps the original profile contour to a new contour, which we later show to be better suitable for detecting feature points. More formally, consider a function $y = f(x); \{x, y\} \in \mathbb{Z}$, where x and y form the co-ordinates of the profile contour. Next consider:

$$x\prime = x/\max(x), \quad y\prime = y/\max(y).$$

We define the distance transformation T to be

$$T = \sqrt{x\prime^2 + y\prime^2}.$$

Fig. 1(d) shows the T-transformed profile contour. There, we see that chin is easily detectable as local maxima which is not so in Fig. 1(c). The transformed profile contour is found to be better oriented for detecting our feature points as local maxima and it works extremely well. In cases, where we still do not observe four local maxima in Fig. 1(d), we assign feature labels to observed and unobserved labels based on assignments in the previous frame.

After detecting the feature points, we perform geometrical measurements to obtain height and protrusion parameters. As described above, the height parameters are the distances between the centers of the lips and the height reference point. In Fig. 2, L_1L_2 is the center lip line; point H, where $MH \perp L_1L_2$, forms the height reference point. L_1H and L_2H constitute the height parameters. For protrusion parameters, we construct the nose-chin line (the dash-dot line in Fig. 2). L_1B_1 and L_2B_2, both perpendicular to the nose-chin line, constitute the two protrusion parameters. A four-dimensional vector is formed from the height and protrusion parameters for PV lip reading. We refer to this feature extraction method as $Method\ 1$.

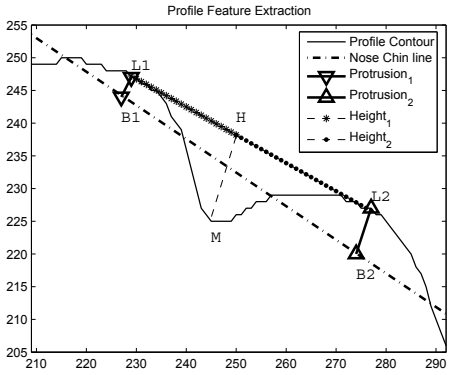

Fig. 2. Profile View final feature extraction steps.

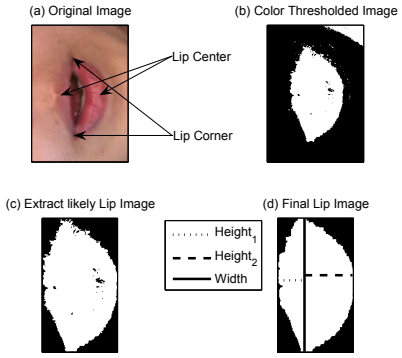

Fig. 3. Frontal View geometrical measurements

3.1. Improved Profile View Feature Extraction

In this subsection, we present an improvement over the Profile View feature extraction Method 1 described in the preceding section. The nose and chin points are kept the same as in Method 1 but lip detection is improved. Lip centers are obtained as local minima points of a distance function constructed for points in between nose and chin. The distance function for a particular point is the perpendicular distance between its corresponding point on the profile contour and the nose-chin line. We refer to this type of feature extraction as $Method\ 2$. This method is motivated by the fact that lip centers protrude outwards, thus forming a local extremum with respect to the nose-chin line. In Method 1, the lip centers may not necessarily be the local maxima of the profile contour but they can still be detected as local extremum points in terms of their distance to the nose-chin line.

4. FRONTAL VIEW FEATURE EXTRACTION

The steps used in frontal view feature extraction are summarized in Fig. 3. We follow a red-detection approach followed by a series of morphological operations for detecting lips in the face image. Features for red detection were also reported in [9]. A binary image is constructed around lips using blue-channel thresholding in RGB image. From the binary image, we identify the connected component which appears to be closest to the lips. In making this decision, constraints based on lip orientation, area, and distance from the image bounding box are employed. A correct decision in the initial image frame is extremely important, because in the first frame we do not exactly know which connected components corresponds to the lips. The binary image may have outlier white pixels from the nose or chin, or even hairs visible from behind the neck of the subject. But once we correctly isolate the lip region in first frame, we only need to look for that region and a small neighborhood around that region for lips in the next frames.

Three features are extracted from the lips, after detecting their locations: one for lip width and two for the lip heights. First, define lip corner line as the line joining the lip corner points; call lip width as its length. Define lip heights as the shortest distances from lip centers to lip corner line. In our current feature extraction method, we are not concerned about detecting the exact lip contour. We only need to detect correctly the bounding box for the lips and then decide where the lip corners lie in the bounding box. With the lip bounding box and lip corner points, we can measure lip width and lip heights, as is shown in Fig. 3(d). We approximate lip width as vertical length of the bounding box and lip heights as distances from the two vertical sides to lip corner line.

5. RESULTS

We present in this section our speaker-dependent isolated word lip reading results in Table 1. We used the Sphinx-3 system [10] for Hidden Markov Model (HMM) training and decoding. An automatic segmentation algorithm was used to identify the beginning and endings of the sounds, with some manual post-processing to reject dictation errors and other non-vocabulary sounds.

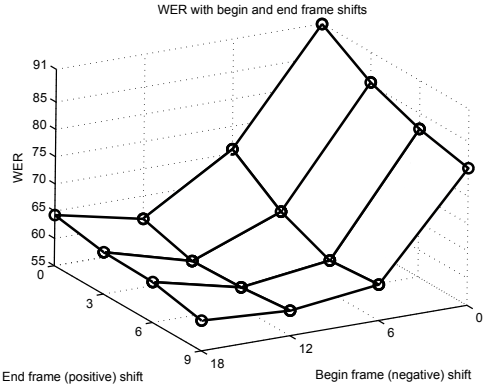

Fig. 4. Profile View error of Subject 2 with begin and end frame shifts

Because the facial musculature operates more slowly than the articulators for speech, we found it advantageous to consider processing the visual features over a longer duration than that marked by the endpoints of the speech utterance. Fig. 4 describes the WER observed for Subject 2 as a function of temporal offset between the beginning and end points of the visual features and those of the speech itself. We found that a best WER of 57.7% using visual features was obtained for this speaker by beginning 12 frames earlier (400 ms) and ending 6 frames later (200 ms) than the begin and end points of

the audio signal. In comparison, the WER obtained if the analysis of the visual features began and ended with the audio would be 90.6%. Results for other speakers were obtained using these same temporal offsets.

Table 1 describes the WER obtained using the two sets of PV features, FV features, and a combination of the two. We note that the WER obtained using FV features is consistent with earlier results (*e.g.* [1][2][11]). The results of Table1 also indicate that PV Method 2 outperforms PV Method 1. This should be expected because in Method 2 we have a better estimate for lip positions and hence the height and protrusion parameters. We also obtained combined WERs in the range of $51 - 55\%$ by concatenating the PV and FV features into a single feature vector.

Next, we present results using decision fusion [2], which is also known as late integration. Consider $c \in C$, where C is the set of words, assume that $O = \{Audio, PV, AV\}$, is the set of features and, let λ be the set of weights. Under our approach the joint word conditional feature probability factors are:

$$P(O|c) = \prod_{o \in O} P(o|c)^{\lambda_o}; \qquad \sum_{o \in O} \lambda_o = 1.$$

The weights λ are selected so as to minimize the overall error on a training set. Fig. 5 shows our decision fusion results for different combinations of features. In this figure "PV" indicates PV Method 2. We optimize λ differently for different SNRs. Since this is a speaker-dependent experiment, we optimize λ separately for each speaker as well. Fig. 5 plots the average WER over the subjects as in Table 1. Comparing the results for the profile and frontal views, we note that profile view lip reading provides substantially smaller WERs than frontal view processing. This result can be attributed to protrusion features in profile view, which play a significant role in the generation of lip cues. Furthermore, we show that lipreading can improve speech recognition, even under relatively high SNR conditions.

Table 1. Word Error Rate, WER (%)

	PV Mth. 1	PV Mth. 2	FV	PV Mth. 2 + FV
Subject1	55.64	52.50	60.50	51.45
Subject2	64.62	57.67	65.50	55.67
Subject3	61.15	55.67	76.83	55.04

6. FUTURE WORK

In this paper we have discussed lip feature extraction based on chromatic separation. We will incorporate other lip modeling techniques in the future. Readers are directed to [12] for further discussion about methods such as deformable templates, active shape modeling and spline fitting. We can extend and adapt these methods from FV to PV.

7. CONCLUSIONS

We introduced profile view lip reading as a scheme for speaker dependent isolated word speech recognition. We presented feature extraction schemes for both the profile and frontal views. We presented results for decision fusion over Audio, PV and FV features. We showed that profile view lip reading provides significantly lower

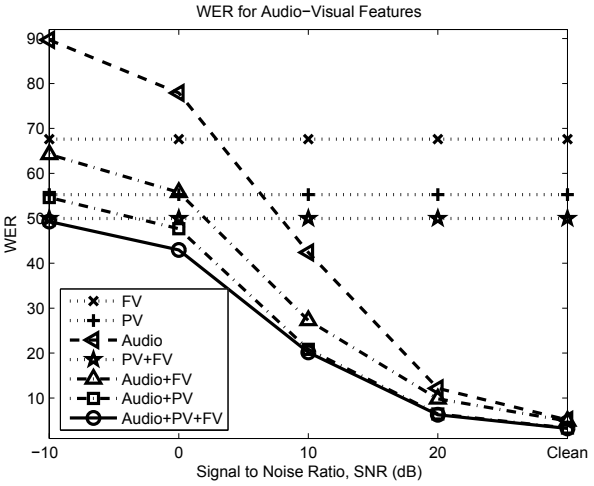

Fig. 5. Decision Fusion of Audio and Visual Features

WERs than are observed using frontal view lip reading. Our results also indicate that best results are obtained when visual analysis begins before and ends after the conventional end points of the speech utterance. Our lip reading database is available for research purposes.

8. REFERENCES

[1] T. Chen, "Audiovisual speech processing. lip reading and lip synchronization," *IEEE Signal Processing Mag.*, vol. 18, pp. 9–21, January 2001.

[2] G. Gravier A. Garg G. Potamianos, C. Neti and A. W. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proc. of the IEEE*, vol. 91, 2003.

[3] F. Lavagetto, "Converting speech into lip movements: A multimedia telephone for hard of hearing people," *in IEEE Trans. on Rehab. Eng.*, vol. 3, March 1995.

[4] K. Aizawa S. Morishima and H. Harashima, "An intelligent facial image coding driven by speech and phoneme," *in Proc. IEEE ICASSP*, pp. 1795–1798, 1989.

[5] K. Iwano S. Furui T. Yoshinaga, S. Tamura, "Audio-visual speech recognition using lip movement extracted from side-face images," *Proc. Auditory Visual Speech Processing (AVSP)*, pp. 117–120, 2003.

[6] G. Potamianos P. Lucey, "Lipreading using profile versus frontal views," *IEEE Multimedia Signal Processing Workshop*, pp. 24–28, October 2006.

[7] C. Savariaux M. Heckmann, K. Kroschel and F. Berthommier, "Dct-based video features for audiovisual speech recognition," *Proc. Int. Conf. Spoken Lang. Process.*, pp. 1925–1928, 2002.

[8] MRT, `http://www.meyersound.com/support/papers/speech/mrtlist.htm`.

[9] T. W. Lewis and D. M. W. Powers, "Lip feature extraction using red exclusion," *in Pan-Sydney Workshop Visualization, Selected Papersl*, vol. 2, pp. 61–67, 2000.

[10] CMU Sphinx Open Source Speech Recognition Engines, `http://cmusphinx.sourceforge.net/html/cmusphinx.php`.

[11] D. G. Stork and M. E. Hennecke, "Speechreading: An overview of image processing, feature extraction, sensory integration and pattern recognition techniques," *in Proc. 2nd Int. Conf. Automatic Face Gesture Recognition*, 1996.

[12] J. Luettin and N. A. Thacker, "Speechreading using probabilistic models," *Comput. Vis. Image Understand.*, vol. 65, pp. 163–178, February 1997.