VISUAL SPEECH FEATURE EXTRACTION FOR IMPROVED SPEECH RECOGNITION

X. Zhang, R. M. Mersereau, M. Clements *C. C. Broun*

ABSTRACT

Mainstream automatic speech recognition has focused almost exclusively on the acoustic signal. The performance of these systems degrades considerably in the real world in the presence of noise. On the other hand, most human listeners, both hearing-impaired and normal hearing, make use of visual information to improve speech perception in acoustically hostile environments. Motivated by humans' ability to lipread, the visual component is considered to yield information that is not always present in the acoustic signal and enables improved accuracy over totally acoustic systems, especially in noisy environments. In this paper, we investigate the usefulness of visual information in speech recognition. We first present a method for automatically locating and extracting visual speech features from a talking person in color video sequences. We then develop a recognition engine to train and recognize sequences of *visual parameters* for the purpose of speech recognition. We particularly explore the impact of various combinations of visual features on the recognition accuracy. We conclude that the inner lip contour features together with the information about the visibility of the tongue and teeth significantly improve the performance over using outer contour only features in both speaker dependent and speaker independent recognition tasks.

1. INTRODUCTION

In the field of automatic speech recognition (ASR), mainstream research has focused almost exclusively on the acoustic signal and has ignored the visual speech cues. While purely acoustic-based ASR systems yield excellent results in a laboratory environment, the recognition error rate can increase dramatically in the real world in the presence of noise such as in a typical office environment with ringing telephones, noise from fans and human conversations. Noise robust methods using feature-normalization algorithms, microphone arrays, representations based on human hearing and other approaches have only limited success in these environments. Indeed, multiple speakers are very hard to separate acoustically.

To overcome this limitation, automatic speechreading systems, through their use of visual information to augment acoustic information, have been considered. The first automatic speechreading system was developed by Petajan in 1984 [1]. He showed that an audio-visual system outperforms either modality alone. During the following years various automatic speechreading systems have been developed [2] which demonstrated that the visual speech information yields information that is not always present in the acoustic signal and enabled improved recognition accuracy over conventional ASR systems, especially in environments corrupted by acoustic noise and multiple talkers. Audio and visual sources of information have been shown to serve complementary functions in speechreading. While an audio speech signal is represented by its acoustic waveform, visual speech signal usually refers to the accompanying lip movements, tongue and teeth visibility and other relevant facial features.

In this paper we investigate the usefulness of visual speech information in speech recognition. We present an automatic visual feature extraction algorithm and provide recognition results based on visual only information. Although fusion of acoustic and visual modalities is possible, in this study we focus exclusively on the visual aspect.

This paper is organized as follows. Section 2 gives a review of previous work on extraction of visual speech features. Section 3 presents our visual front-end for lip feature extraction. In Section 4, we examine the problem of speech recognition using visual speech information. Finally, Section 5 concludes the paper.

2. PREVIOUS WORK

The choice for a visual representation of lip movement has led to various approaches to visual speech feature extraction. At one extreme, the entire image of the talking person's mouth is used as a feature [3, 4]. In this case no information is lost, but it is left to the recognition engine to determine the relevant features in the image. This approach tends to be very sensitive to changes in illumination, position, and speaker [5]. With other approaches, only

a small set of parameters describing the relevant informa-r tion of the lip movement is used for the recognition. In this r approach, model-based methods such as deformable tem-r plates, "snakes" and active shape models [2] are commonly r used. Traditionally, they are performed using gray-scale im-r ages. The difficulty with these approaches usually arises when the contrast is poor along the lip contours, which oc-r curs quite often under natural lighting conditions. In partic-r ular, edges on the lower lip are hard to distinguish because r of shading and reflection. r

An obvious way of overcoming the inherit limitation r of the intensity-based approach is to use color, which can greatly simplify lip identification and extraction. Lip feature r extraction using color information has gained interest in re-r cent years with the increasing processing power and storage r of hardware making color image analysis more affordable. r

In this work, we present an approach that extracts lip r features using color video sequences. Previous work re-stricts the visual speech features to the lip outer contour r only. However, it is known from human perceptual stud-r ies that more visual speech information is contained within r the lip inner contour. Besides, the presence/absence of the r teeth and the tongue inside the mouth is also important to human lipreaders [6]. We, therefore, aim at extracting both r outer and inner lip contour parameters, as well as detecting r the presence/absence of teeth and tongue. r

3. LIP FEATURE EXTRACTION

Fig. 1 depicts the procedures involved in the visual process-r ing. The first stage of the visual analysis involves lip region r localization. In our previous work [7], we demonstrated that hue is an effective descriptor in characterizing the lips i e-i cause of its invariance to luminance and human skin color and its discriminative properties. Using hue and saturation r information, combined with motion cues, we are able to re-r liably detect the mouths of a talking person from a video r sequence [8]. r

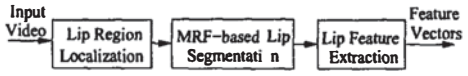

Fig. 1. Visual processing. s

To derive the lip dimensions within a video sequence, i we make use of both color and edge information of an im-r age. r These are combined within a Markov random field r (MRF) framework, which has been shown to be suitable r for the problem of spatial statistical modeling. r Details r f r MRF-based lip segmentation can be found in [8]. r

Segmentation results with different persons and differ-r ent lip opening situations are demonstrated in Fig. 2. We i bserve that the highlighted pixels fairly well match the r true lip area. Based on the segmented lip image, we are t ready to extract the key feature points on the lips. We first r compute the horizontal lip line, which has the longest hor-r izontal span in the segmented image, and extract the two r feature points — the left and right outer corners on the lip r line. Then we derive the vertical lip line assuming left/right symmetry of the lips. We detect four feature points along r the vertical lip line —the uppper/lower outer/inner lip. To r increase the accuracy r f the identified feature points, we in-r corporate intensity gradient information. If the gradient r f r the detected point is below a preset value, we start search-r ing for the largest gradient in its vicinity, and replace the r old value with it. Finally, given the constraints of the outer r corners and the upper/lower inner lip, we locate the inner r lip corners. Fig. 3 shows the extracted key feature points.

Fig. 2. Segmented lips overlayed on the original image.

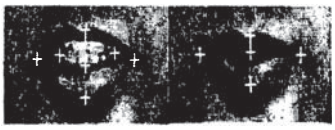

Fig. 3. Measured feature points on the lips. r

Based on the extracted key feature points, we can derive r the geometric dimensions of the lips. The following features r are used in our study: mouth width (w_2), upper/lower lip i width (h_1, h_3), lip opening height/width (h_2, w_1), and the r distance between the horizontal lip line and the upper lip r (h_4). An illustration r f the geometry is shown in Fig. 4.

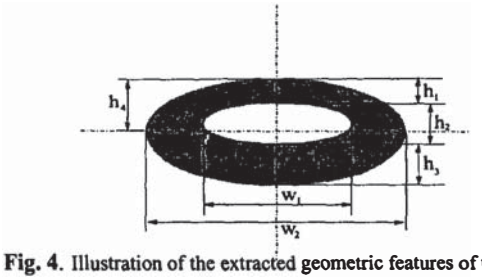

Fig. 4. Illustration of the extracted geometric features of the lips.

Besides the geometric dimensions of the lips, we also s detect the visibility of the tongue and teeth. For detecting r the tongue, we search for the "lip" labels along the vertical r lip line within the inner lip region. Two cases need to be dif-i ferentiated, as shown in Fig. 5. In the first case, the tongue

is separated from the lips by the teeth. Tongue detection is trivial in this case. In the second case however, the tongue merges with the lips. From the segmented image, we have a lip closure case. Here we use the gradient of the intensity to detect the inner upper/lower lip. In the case that $h_2 \neq 0$, we search for intensity gradient values along the vertical lip line. If the gradients of two points exceeding a preset value are found, they are identified as upper/lower inner lip, and the indicator of the tongue is set to 1 (presence), otherwise 0 (absence).

a) (b)

Fig. 5. (a) Tongue is separated from the lips. (b) Tongue merges with the lips.

The teeth are also easy to detect since their H value is distinctly different from the hue of the lips. This is a big advantage compared to gray-level based approaches which may confuse skin-lip and lip-teeth edges. Teeth are detected by forming a bounding box around the inner mouth area and testing pixels for white tooth color: $S < S_0$, where $S_0 \approx 0.35$. The indicator of the teeth is either 1 or 0.

We applied the feature extraction algorithm on the Carnegie Mellon University database [9] with ten test subjects. The database includes head-shoulder full frontal face color *video sequences* of a person talking. The test subjects have various skin complexions with no particular lipstick. The feature extraction algorithm works well for the data sets. In a few cases, a few pixels of inaccuracy are observed.

4. SPEECH RECOGNITION

In this section we describe the modeling of the extracted lip features for speech recognition using hidden Markov models. HMMs have been successfully used by the speech recognition community for many years. These models provide a mathematically convenient way of describing the evolution of time sequential data.

In speech recognition, we model the speech sequence by a first-order Markov state machine. The Markov property is encoded by a set of transition probabilities with $a_{ij} = P(q_t = j | q_{t-1} = i)$, the probability of moving to state j at time t given the state is at time $t-1$. The state at any given time is unknown or hidden. It can however be probabilistically inferred through the observations sequence $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_T\}$, where $\mathbf{o}_t$ is the feature vector extracted at time frame t and T is the total number of observation vectors. Since the feature vector for the tongue/teeth is represented in binary form, we employ a discrete HMM. The form of the output distributions in a discrete HMM is given by the following:

$$b_i(\mathbf{o}_t) = P_i(v(\mathbf{o}_t)), \tag{1}$$

where $v(\mathbf{o}_t)$ is the output of the vector quantizer given input vector $\mathbf{o}_t$ and $P_i(v)$ is the probability of state i generating symbol v.

An HMM representing a particular word class is defined by a parameter set $\lambda = (A, B, \pi)$, where π is the vector of initial state probabilities, $A = \{a_{ij}\}$ the matrix of state transition probabilities, and $B = \{b_i(\mathbf{o}_t)\}$ the vector of state dependent observation probabilities. Given a set of training data (segmented and labeled examples of speech sequences), the HMM parameters for each word class are estimated using a standard EM algorithm. Recognition requires evaluating the probability that a given HMM would generate an observed input sequence. This can be approximated by using the Viterbi algorithm. For this, given a test token $\mathbf{O}$, we calculate $P(\mathbf{O}|\lambda_i)$ for each HMM, and select λ_c where $c = \arg\max_i P(\mathbf{O}|\lambda_i)$.

We perform the speech recognition task using the audio-visual database from Carnegie Mellon University [9]. This database includes ten test subjects (three females, seven males) speaking 78 isolated words repeated 10 times. In our experiment, we use the data set for seven weekdays — *Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday.*

We conducted tests for both speaker dependent and independent tasks using visual parameters only. The eight visual features used are: $w_1, w_2, h_1, h_2, h_3, h_4$ corresponding to Fig. 4, and the presence/absence of the teeth/tongue. For comparison, we also provide test results on partial feature sets. In particular, we limited the features to the geometric dimensions of the inner contour (w_1, h_2), and outer contour $(w_2, h_1 + h_2 + h_3)$. The role of the use of the tongue and teeth parameters was also evaluated. For the HMM, we used ten states and the recognition system was implemented using the HTK Toolkit.

For the speaker dependent task, the test was set up by using a leave-one-out procedure, i.e., for each person, nine repetitions were used for training and the tenth for testing. This was repeated ten times. The recognition rate was averaged over the ten tests and a gain over all ten speakers. For the speaker independent task, we use different speakers for training and testing, i.e., nine subjects for training and the tenth for testing. The whole procedure was repeated ten times, each time leaving a different subject out for testing. The recognition rate was averaged over all ten speakers.

The experimental results for the two modes are shown in the following table. Rows correspond to various combinations of visual features used. The numbers in the brackets give the total number of features used in each test. The Δ

refers to the delta features — the difference of each feature between successive frames. The second and third columns give th average results in the speake dependent (S.D.) and speaker independent (S.I.) mode, respectively. For the speaker independent task, feature vectors were preprocessed by normalizing against the average mouth width w_2 of each speaker to account for the difference in scale between different speakers. All ecognition rates are given in %.

Features	S.D.	S.I.
all (8)	72.572	40.29
above+Δ (16)	78.285	48.43
all except tongue/teeth (6)	67.429	36.43
above+Δ (12)	73.285	44.29
outer/inner contour (4)	66.285	36.71
above+Δ (8)	72.429	46.28
outer contour (2)	59.856	25.85
above + Δ (4)	65.428	32.71
inner contour (2)	61.713	36.85
above+Δ(4)	65.144	41.86

Table 1. Recognition rate fo the speech recognition tasks using database [9].

We observe that the geometric dimensions of the lip outer contour, as used in many previous approaches, are not adequate for recovering the speech information. While the use of the lip inner contour features achieves almost the same cognition ate as that of the lip outer contour in the S.D. mode, it outperforms the former by a significant 11% in the S.I. task, and suggests it provides a better speaker independent characteristic. The contribution of the use of tongue/teeth is 5.1% in the S.D. and 3.8% in the S.I. task. The delta features yield additional improved accuracy by providing extra dynamic information. Overall best - sults are obtained by using all relevant features, achieving 78.285% fo S.D. and 48.43% for S.I. task. These compare favorably with using outer contour only features by 12.8% for S.D. and 15.7% for S.I., respectively.

5. SUMMARY AND CONCLUSIONS

In this paper we described a method of automatic lip feature extraction and its application to speech recognition. Our algorithm first reliably locates the mouth region, then subsequently segments the lip from its surroundings by utilizing a Markov random field framework. The lip key points that define the lip position are detected and the relevant visual speech parameters are derived and form the input to the recognition engine. In our speech recognition xperiments, we applied hidden Markov models to model the xtracted features. Experiments from both speaker dependent and speaker independent tasks indicate that the lip features of the outer contour alone are not sufficient for recovering the relevant speech information. y incorporating the inner lip contour features and the information about the visibility of the tongue and teeth, significant improvements of 12.8% for speaker dependent case and 15.7% for speaker independent case can be achieved.

Acknowledgments

We would like to acknowledge the use of audio-visual data [9] from the Advanced Multimedia Processing Lab at the Carnegie Melon University.

6. REFERENCES

[1] E. D. Petajan, *Automatic lipreading to Enhance Speech a Recognition, a* Ph.D thesis, Univ. of Illinois, Urbana-Champaign, 1984.

[2] D. G. lStork and M. E. Hennecke, *Speechreading by a Humans and Machines*, vol. 150 of *NATO dSI Series F*, Springer Verlag, 1996.

[3] J. R. l Movellan, "Visual speech recognition with stochastic networks," in *Advances in Neural Informa-a tion Processing Systems*, vol. 7. MIT Press, 1995, editor G. Tesauro, D. Touretzky, and T. Leen.

[4] G. lPotamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual LVCSR," in *roc. IEEE ICASSP*, 2001.

[5] P. L. Silsbee and A. C. Bovik, "Computer lipreading for improved accuracy in automatic speech cognition," *IEEE Transaction on Speech and Audio Processing*, vol. 4, no. 5, p. 337–351, 1996.

[6] A. Q. Summerfield, "Lipreading and audio-visual speech perception," *Philosophical Transactions of the a Royal Society of London, Series B*, vol. 335, pp. 71–78, 1992.

[7] X. Zhang and R. M. Mersereau, "Lip feature extraction towards an automatic speechreading system," in *roc IEEE ICIP*, 2000.

[8] X. Zhang, C. C. Broun, R. M. Mersereau, and M. Clements, "Automatic speechreading with applications to human-computer-interfaces," Submitted to EURASIP Journal on Applied Signal Processing, Special issue on Audio-Visual Speech Processing, 2002.

[9] "URL: amp.ece.cmu.edu/intel/feature_data.html," .