ISSN:0975-9646

Recognition of spoken English phrases using visual features extraction and classification

Salma Pathan[#1], **Archana Ghotkar** [#2]
[#]*Department of Computer Engineering*
Pune Institute of Computer Technology
Pune, India

Abstract: **Lip reading plays an important role in speech recognition under noisy conditions or for listeners with hearing impairment. Lip reading means identification of words or letters said from the movement of lips. Lip reading has applications in security, gaming, human-computer interactions, and deafness research. In this paper, a solution for automatic lip tracking and recognition of phrases of English language spoken by speakers using the only visual information obtained from lip movements is proposed. Here, four main lip points are detected left, upper, right, and lower. The visual features selected are the angles between these lip points. Applying support vector machine classification to these features the said phrase is recognised. The proposed system has been found to have overall accuracy of 65.6% for phrase recognition with speaker lip movements as the only input and without using any speech recognition system in parallel.**

Keywords: **Lip Reading, Feature extraction and Classification, Support vector machines.**

I. INTRODUCTION

Lip reading is a method used to understand or interpret speech without hearing it, especially used by people with hearing impairment. A person with a hearing impairment depends on lip reading to communicate with others and to engage in social activities. He tries to understand what is said only by watching lip movements. Also, visual speech information is important for speech recognition in noisy area. Audio information is affected by loud noise and crosstalk noise among speakers. So in area where noise cannot be avoided it will be beneficial to use visual information along with audio information. This is challenging because of the visual appearance vary with speaker to speaker and can contain very less information as compared to a sound signal therefore identification of robust features in visual domain is still centre of attraction of many researchers.

Recent advances in the fields of image processing, pattern recognition has led to increase in research for automating lip reading. Visual speech recognition (VSR) or speech reading could open the door for many novel applications where in absence of audio the speech could be recognized. VSR can be used in many applications such as speaker recognition, human-computer interaction (HCI), sign language recognition, language recognition, audio-visual speech recognition (AVSR), and security purposes when audio is not available or audible. VSR aims to recognize spoken word(s) by using only the visual information that is produced during speech. Hence, VSR is a research area with the visual domain of speech and involves image processing, artificial intelligence, object detection, pattern recognition, etc.

In the proposed work, a camera is required to capture the video while a speaker is saying the phrase. Using this recorded video, the points on lips are obtained by tracking face and then locating lips. From these points the features which are angles between these points are calculated and then applying Support vector machines classification the phrase said is recognized.

The objective of this paper is to recognize different English phrases using visual features such as geometrical shape of the lip. Section II describes related work, Section III describes proposed work, Section IV database, experimental result, and Section V discusses limitations and Section VI conclusions and future work.

II. RELATED WORK

The performance of lip reading techniques depends on visual feature selection. There were many methods developed to extract visual features. Visual features can be extracted by either appearance based methods or geometrical methods or combination of both. In geometric approach, information is obtained from the mouth region such as the shape of mouth, height, width, and area of mouth region and in appearance-based approaches the pixel values of the mouth region are considered, and they apply to both grey and coloured images. First system using geometrical based feature was developed by Petajan *et al.* [1]. In their system simple thresholding of the mouth image is used to highlight the lip area, and then measurements of mouth height, width, and area were taken. Potamianos *et al.* [2] compared PCA, DWT and DCT trans-form techniques for digit recognition using HMM with 6 states and found that result of DCT is more accurate as compared to other techniques. Mattews *et al.* [3] describes and evaluates two methods of visual feature extraction for integration into an audio-visual speech recognizer. Video-only recognition results are presented for multi-speaker, word-level, isolated letters recognition, using HMMs for speech modeling, and using low resolution grayscale video. The best results presented are 41.9% word accuracy, using active appearance model (AAM) features, and 26.9% using active shape model (ASM) features. Potamianos *et al.* [4] uses both multi-speaker and speaker-independent scenarios. In a digit recognition task using studio recorded video, they obtained 61.47% word accuracy in a speaker-independent task and 76.42% using a multi-speaker task. Cox *et al.* [5] presents results for visual speech-recognition only. They also used letters of the alphabet as test-data but at higher camera resolution and using color. Their results show that an accuracy of above 80% is achieved for all speakers in a multi-speaker testing scenario, but in speaker-independent tests, the accuracy drops dramatically to below 10%, and in some cases to around chance level. This paper illustrates the strong speaker dependency of the AAM features and cites this as the reason for the poor speaker-independent performance. Alghathbar *et al.* [6] adapted approach called wrapping snakes, where the image forces are modified based on the snake's location and orientation was presented. The modified wrapping snakes encouraged to continue along features which they have been already partially found. Saitoh [7] used AAM, automatic utterance section detector, and phrase classifier using dynamic programming matching. The combined parameter (called c-parameter) obtained with lip AAM is used as the feature. This parameter contains both shape parameter and texture parameter of the target region, obtained accuracy for Japanese words over telephonic conversation of 94.4% for speaker dependent experiments.

III. Proposed Approach

Here, our aim is to recognize the phrase said only by selection and extraction visual features from the movement of lips. The system block diagram is shown in Fig.1. The speaker is asked to say the phrase which is recorded using a camera. From this video frames are extracted, and for each frame face points are detected. Face points include various points on eyes, chin, lips etc. The points on lip are used for features extraction.

The first step in developing the lip reading system involves recognizing the speaker's face in every video frame using a facial recognition algorithm in OpenCV2. The face detector is based on the detection of features called Haar-like features, which encode the existence of oriented contrasts between regions in the image. Details of the face detection system can be found in [8]. Fig 2 shows the face detection which lip points marked. After detecting the speaker's mouth region, key points were placed on the left, right, upper and lower points which allowed for numerical feature extraction based on the changes in the positions of the speaker's lips over time. These points are used to calculate the features which are angles between these points. The first feature is p1 the angle between left, upper and right points and second feature is p2 which is angle between upper, left and lower points. These two feature vectors are given input to support vector machine classifier and the said phrase is recognized.

In order to extract the visual features in the face image acquired from the previous step, an accurate extraction of the lip area is essential. In our approach landmark model is used which gives the points on lips. The points of interest (key points) used in our approach for lip reading is shown in Fig 3.The upper, lower, left and right key points are detected. These points are used to calculate the angles between them. The feature vector p1 and p2 are calculated as follows

$$Angle\ p1 = atan\left[\frac{m2-m1}{1+m2m1}\right]$$

$$Angle\ p2 = atan\left[\frac{m3-m1}{1+m1m3}\right]$$

where
m1= Δ[Left(x,y), Upper(x,y)]
m2= Δ[Left(x,y),Lower(x,y)]
m3 =Δ [Upper(x,y),right(x,y)]

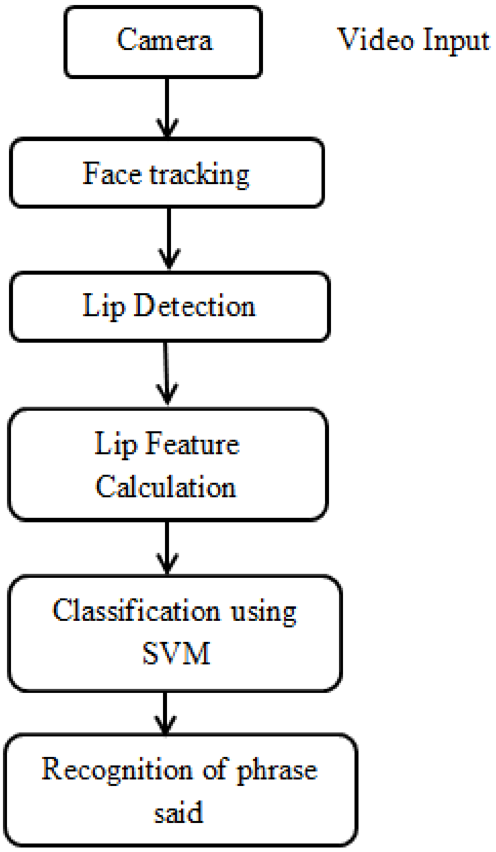


Fig.1 System Block Diagram

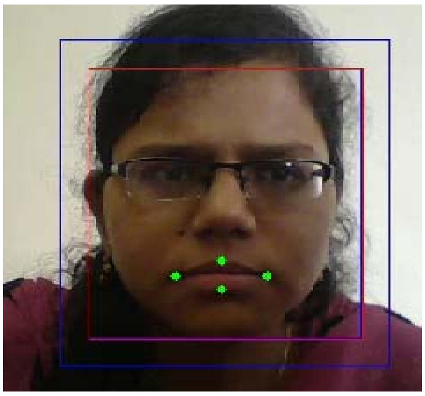


Fig.2 Face Detection

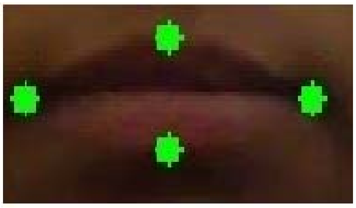
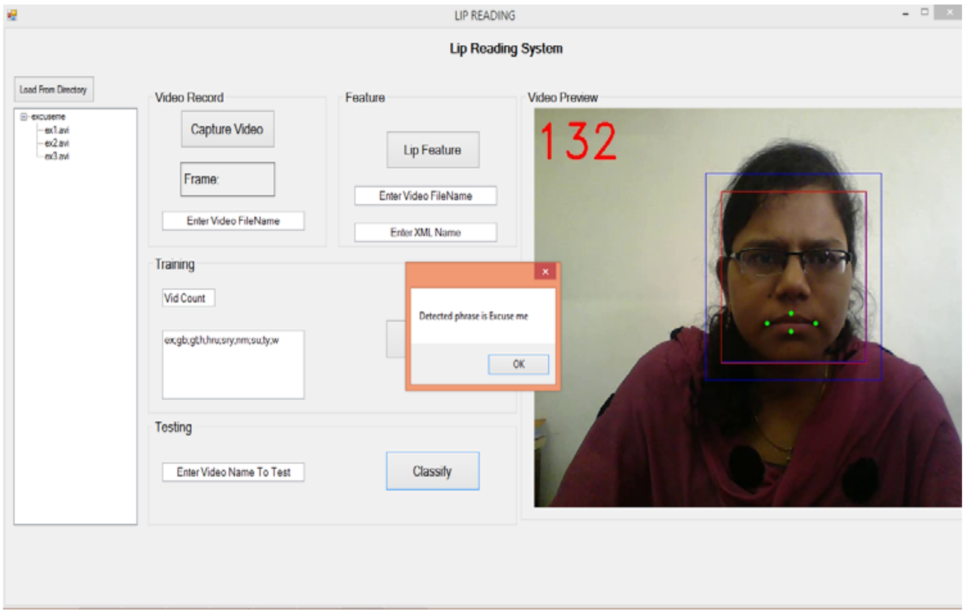


Fig.3 Lip points (left, upper, right, lower)



ex,gb,gt,h,hru,sry,nm,su,ty,w

Detected phrase is Excuse me

Fig.4 Screenshot of lip reading system

The feature vectors calculated in previous step are trained using support vector machines. Support vector machine (SVM) classifier for linear and separable case is capable to find the optimal separating hyper plane between classes in sparse high-dimensional spaces with relatively few training data. SVM maximizes the distance of the separating hyper plane from the closest training data point called the support vectors. SVMs were selected due to the ability of SVMs to find a globally optimum decision function to separate the different classes of data. The SVMs were trained with 300 training samples and were tested using the 300 remaining samples (3 samples from each phrase) for all 10 speakers.

As shown in Fig. 4 the lip reading system, it can record a video using capture button. Extract visual features of that video, store these feature vectors in xml file. Train the support vectors for ten different phrases said by ten different speakers using 3 xml files for each phrase and each speaker. For testing, load the video files to be tested and using classify button the recognized phrase will be shown in message box.

IV. Database And Experimental Results

The speech was recorded using resolution of 640 × 480 pixels, with 20 frames per second rate setting. Face lighting and distance between speaker and camera throughout experiment was maintained in similar level. To evaluate the performance of the proposed automatic lip-reading algorithm, a database consisting of 10 isolated English phrases were recorded from ten different speakers repeating 6 times each frame. The English phrases recorded were ***Hello, Excuse me, I am sorry, Thank you, Good bye, See you, Nice to meet you, You are welcome, How are you, Have a good time***. Out of the 600 videos, 300 videos were used for training and remaining for testing. The subject dependent tests are summarised using a confusion matrix shown in Table I. The confusion matrix summarizes the result for 30 videos of each phrase which has been tested. From this, it can be seen that all classes except one class with accuracy rates above 60%. The "welcome" class had the lowest accuracy. The overall accuracy of the system is 65.6%. The highest accuracy for recognition between various speakers is for speaker P1. The recognition rate for each phrase is shown in Fig.5. Accuracy for each speaker is shown in Fig. 6

ex	Excuse me
gb	Goodbye
gt	Have a good time
h	Hello
hru	How are you
imsry	I'm sorry
nm	Nice to meet you
su	See you
ty	Thank you
w	You are welcome

	ex	**gb**	**gt**	**h**	**hru**	**imsry**	**nm**	**su**	**ty**	**w**
ex	22	2	1	1	0	1	0	0	1	2
gb	4	21	0	1	0	1	0	0	1	2
gt	0	0	20	0	0	2	3	1	3	1
h	2	0	1	20	0	4	0	2	0	1
hru	1	3	0	0	19	1	1	3	2	0
imsry	2	3	0	1	0	20	2	0	2	0
nm	1	1	0	0	1	0	22	1	3	1
su	1	0	2	4	1	3	0	18	1	0
ty	0	2	3	0	0	4	1	1	18	1
w	2	0	0	3	2	2	1	2	1	17

Table I Confusion matrix for speaker dependent test

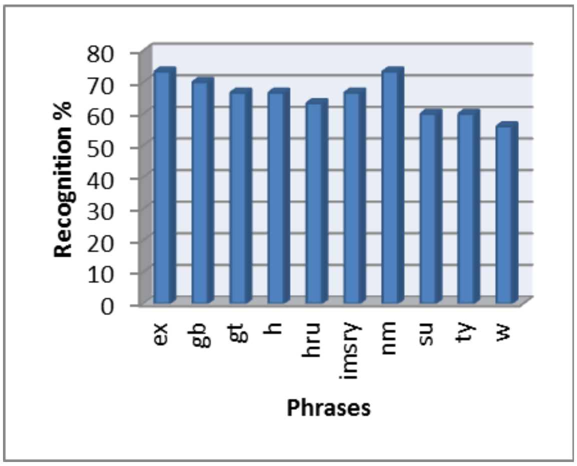


Fig.5 Recognition Rate for each phrase

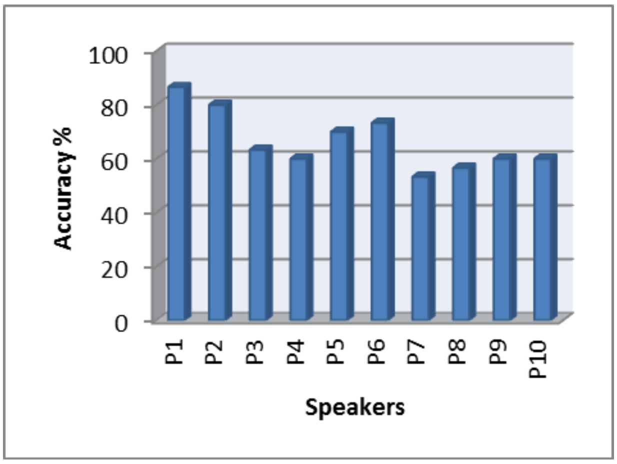


Fig.6 Recognition Rate for each person

V. Discussion

Some limitations include difficulties in lip segmentation in low-light conditions, curve fitting, and feature extraction and training. Factors such as lighting, the speaker's skin color, variations in the pronunciation of common words due to accents, variations in the appearance of the speaker's face, and potential errors in the lip segmentation algorithm all may have affected the lip reading system's performance. The recognition rate depends on the feature selection and classifier. The angles remain same or vary very less for many phrases having similar lip movements, thus resulting in low accuracy and poor performance. Another limitation of the system is it can recognize only 10 different phrases. The accuracy of the system can be improved by adding more robust visual features like height, width, area of lip region etc.

VI. Conclusion

The researchers are still working to find the most accurate and robust processing model for lip tracking. The entirely different methods can be transformed in each other with accuracy that is sufficient to preserve the obtained recognition rates. The developers of lip-reading systems may now start to concentrate on the recognition models themselves and assume that they will be able to plug in any mouth tracking method that appears to be the most robust one at a later point in time. Selection of visual features and their classification plays important role in performance of lip reading systems .Here, in proposed method angles were the features and using SVM as a classifier accuracy of 65.6% was obtained. In future additional features will be added and tested with different classifier for improving accuracy and also increasing the vocabulary for recognition.

References

[1] E. Petajan, B. Bischoff, D. Bodoff, An improved automatic lip reading system to enhance speech recognition, CHI' 88 (1988) 19–23.

[2] G. Potamianos, H. Graf, E. Cosatto, An image transform approach for HMM based automatic lip reading, in: International Conference on Image Processing, 1998, pp. 173–177.

[3] I. Matthews, T. Cootes, J. Bangham, Extraction of visual features for lipreading, in: IEEE Trans. on Pattern Analysis and Machine Vision, 2002, pp. 198–213.

[4] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, Recent advances in the automatic recognition of audiovisual speech, *Proc.IEEE*, vol. 91, no. 9, pp. 1306–1326, Sep. 2003.

[5] S. Cox, R. Harvey, Y. Lan, J. Newman, and B.-J. Theobald, The challenge f multispeaker lip-reading, in *Proc. Int. Conf. Auditory-Vis.Speech Process. (AVSP)*, 2008, pp. 179–184.

[6] Alghathbar, Khaled, & Mahmoud, Hanan A. Block-based motion estimation analysis for lip reading user authentication systems, WSEAS Transactions on Information Science and Applications, 6(5), 2009, pp 829–838.

[7] Takeshi Saitoh, Real-time Lip Reading System for Fixed Phrase and Its Combination, First Asian Conference on Pattern Recognition (ACPR), 2011 pp 461 – 464.

[8] P. Viola, and M. Jones, Rapid Object Detection Using a Boosted Cascade of Simple Features, *CVPR*, Kauai, 2001