SPEAKER IDENTIFICATION AND SPEECH READING USING LIP MOTION

Rakesh Parekh[1], Nausheen Azam[2], Shruthika Nair[3], Asmita Deshmukh[4]
[1,2,3]B.E. Student, [4]Assitant Professor, Computer Engineering Departement
K.C. College of Engineering and Management Studies,
Thane, India
[1]rocky.rp180@gmail.com,[2] nausheen1dec@gmail.com,[3] shruthsnr@yahoo.com, [4]asmitadeshmukh7@gmail.com

Abstract- There have been many applications made for security which consists of password which needs to be entered by user with their hands. The existing system uses lip motion on the large scale. This paper proposes to use the motion of the user's lip instead of entering a password. It consists of few algorithms that detects the motion of the user's lip by creating a grid around the lip region.

KEYWORDS*:*
Applications, security, algorithms

I. INTRODUCTION

Lip information has been extensively used in audio visual speech reading and speaker identification applications, since lip motion is highly correlated with audio signal. Hence there exists 3 alternative representation of lip information: 1) lip texture 2) lip geometry and 3) lip motion features [1]. Out of these 3 representations lip texture may degrade the recognition performance since it is sensitive to acquisition conditions. The second, lip geometry, usually requires tracking of the lip contour and fitting contour model parameters and computing geometric features such as horizontal/vertical openings, contour perimeter, lip area, etc. Thus the last option is using explicit lip motion feature, which are comparatively easy to compute and robust to lightning variations between the training and tests data sets. Determination of the best lip motion features for speech-reading and speaker identification is the focus of this work. In audio-visual speech recognition (speech-reading), lip texture information is widely used. The principal component analysis (PCA) has been applied to raw lip intensity image to reduce its dimension, and the reduced vector is used as the visual feature [1]. Another possibility is to use discrete cosine transform (DCT) coefficients of the gray-scale lip image. Then apply linear discrimination analysis (LDA) to the final feature vector formed by concatenating a number of consecutive feature vectors centered at the current frame so as to capture dynamic speech information.

IJERT

II. PROPOSED SYSTEM

Discriminative analysis of lip motion for speaker identification is basically a very large scale procedure by means of which the lip motion of individuals are recorded , the speaker is identified and if he is found to be an authorized entity the individual is allowed to access the data In our project we are focusing on speaker identification for Home security. Here we are going to create silent password. We are planning to do this using both from a video file or from live streaming. The main thing is that he camera should focus on the lip. It should track the lip movement. The key idea is that we are going to take a threshold value, using that particular value we are going to check whether the password is wrong or right. We will track the lip movement, if the movement is beyond threshold thenit will indicate that the password is wrong and if the threshold is in the range then it will indicate that the password is right.

A. Extraction of Grid-Based Motion Features

We first consider dense motion estimation over a uniform grid of size $G_x x\ G_y$ on the extracted lip region image. We use hierarchical block matching to estimate the lip motion with quarter-pel accuracy by interpolating the original lip image using the 6-tap

Wiener and bilinear filters [1]. The motion estimation procedure yields two $G_x x\ G_y$ 2-D matrices V_x and V_y, which contain the x- and y– components of the motion vectors at grid points, respectively. The motion matrices, V_x and V_y , are separately transformed via 2-D-DCT. The first M DCT coefficients along the zig-zag scan order, both for x and y directions, are combined to form a feature vector f of dimension $2M$. This feature vector representing the dense grid motion will be denoted by f_{GRD}.

Transforming the motion data into DCT domain has two advantages. First, it serves as a tool to reduce the feature dimension by filtering out the high frequency components of the motion signal.

These high frequency components are mostly due to noise and irrelevant to our analysis since it is unnatural to have very abrupt motion changes between neighboring pixels of the lip region, where the motion signal is expected to have some smoothness. Second, DCT de-correlates the feature vector so that the discriminative power of each feature component can independently
be analyzed.

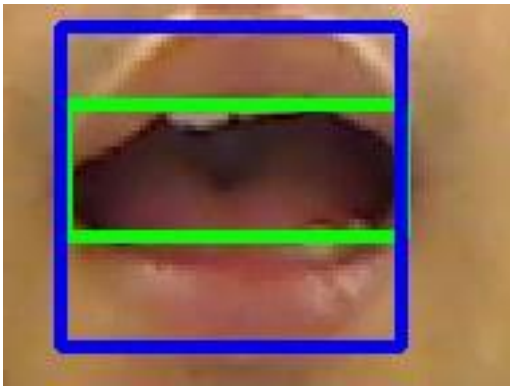


Fig 1. Image of a lip region

Here are the steps we need to do to create a silent password -
1.Take video of lip movement or record using webcam.
2.Now use a detection algorithm algorithm from **here**
3.Now what you need to do is that you need to set the threshold value.

Threshold value:

Now the threshold value can be set or generated in many ways.

1.Calculate the no of frames between one lip touching frame to the next one.

2. In live streaming case the threshold should be given a bit large for less errors.

3. For this case the calculated frame no is the threshold.
4.Thus in this way the threshold value is calculated.

5.Now simple calculation,i.e. if detected result is beyond threshold then it is rejected, otherwise it is accepted.

FLOWCHART FOR EXTRACTING LIP REGION:

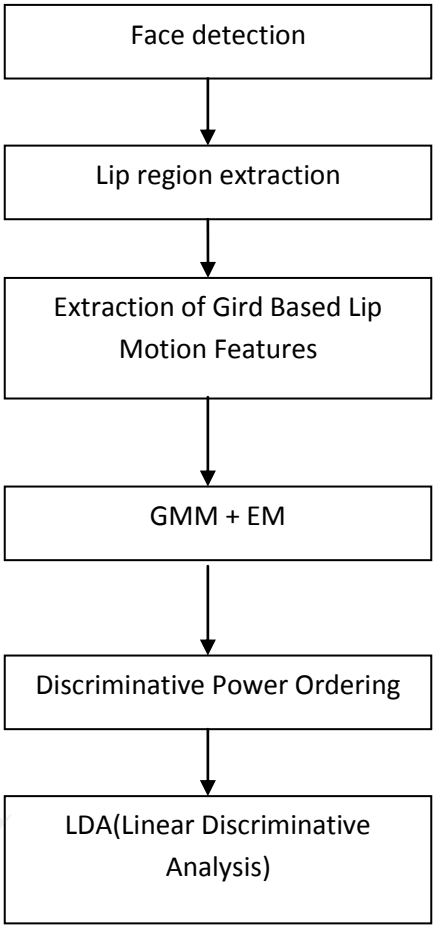


Fig 2. Flow of extacting Lip region and processing.

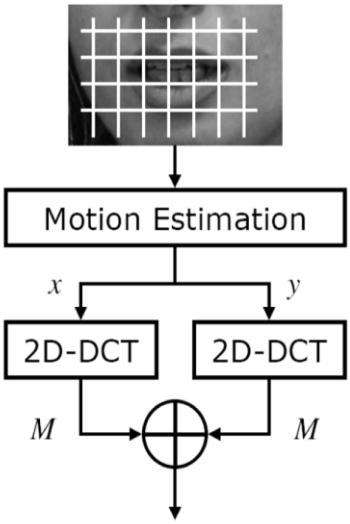
In the above flowchart first the preprocessing steps is performed. The preprocessing steps involves

i. Face detection
ii. Lip region Extraction

In Face detection process the user's face is first tracked using the webcam. The face detection process is the elementary step but it doesn't involves completely in the project.

The LIP REGION EXTRACTION is the main focus of our project. The purpose of the preprocessing module is to register lip regions in successive frames by eliminating global head motion so that the extracted motion features within the lip region
correspond to speaking act only. Hence, each frame of the sequence is aligned with the first frame using a two-dimensional (2-D) parametric motion estimator.

IJERT



$$f_{\mathrm{GRD}} = \{ f_x^1, f_y^1, f_x^2, f_y^2, \ldots, f_x^M, f_y^M \}$$

The next process that is carried out is BAYESIAN DISCRIMINATIVE FEATURE SELECTION.

Let f_k denote the th component of a feature vector .
Given an observation f_k , the maximum *a posteriori* (MAP) estimator selects the class with the MAP probability $P(\lambda_i|f_k)$ which can be written in terms of class conditional probability distributions

$$P(\lambda_i\ f_k) = \frac{P(f_k|\lambda_i)P(\lambda_i)}{P(f_k)} = \frac{P(f_k|\lambda_i)P(\lambda_i)}{P(f_k|\lambda_i)P(\lambda_i) + \sum_{j\neq i} P(f_k|\lambda_j)P(\lambda_j)} = \left[1 + \frac{\sum_{j\neq i} P(f_k\ \lambda_j)P(\lambda_j)}{P(f_k|\lambda_i)P(\lambda_i)}\right]^{-1}.$$

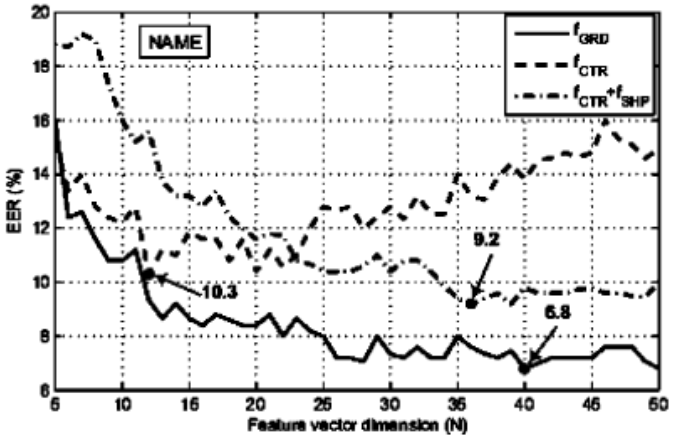
This ratio can be interpreted as the ratio of intra-class and interclass
probabilities, and when maximized, it can serve as a measure
of discrimination between the class and all other classes
for the corresponding feature component f_k .

B. Evaluaton of various Lip Motion Feature:

1) Speaker Identification: Name Scenario:

In the name scenario implementation, the D_n database is partitioned into two disjoint sets,$\{D_{n1}$ and $D_{n2}\}$, each having five repetitions from each subject in the database. The subset D_{n1} is used for training, and D_{n2} is used for testing. Since there are 50 subjects and five repetitions for each true and imposter client tests, the total number of trials for the true accepts and true rejects is respectively $N_a = 250$
and $N_r = 250$.

The three lip motion feature representations, f_{GRD}^N, f_{CTR}^N and $f_{CTR}^N + f_{SHP}$, are tested on the database. Fig. 3. Displays the EER performances with varying feature dimension .
We observe that the grid-based motion features, , achieve 6.8% EER, and outperform the contour-based features.



2) *Speaker Identification: Digit Scenario:*

In the digit scenario, the D_d database is partitioned into two disjoint sets,$\{ D_{d1}$ and $D_{d2}\}$, each having five repetitions of the same 6-digit number from each subject in the database. Again the D_{d1} is used for training and D_{d2} is used for testing [2].
Note that, in the digit scenario, no imposter recordings are performed since every subject utters the same 6-digit number. Hence, the imposter
clients are generated by the *leave-one-out* scheme, where each subject becomes the imposter of the remaining R-1 subjects in the population. Having R=50 subjects and five testing repetitions, the resulting total number of trials for the true accepts and true rejects (imposters) becomes respectively $N_a = 250$ and $N_r = 250$.

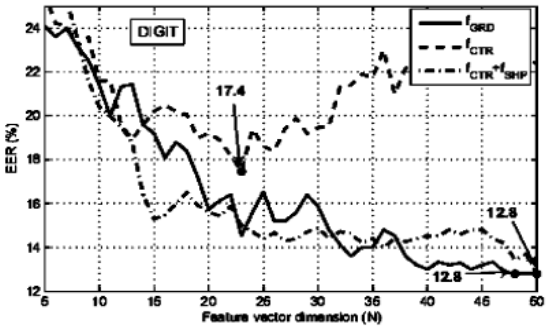


Fig 3. shows the EER performances for different lip motion representations with varying feature dimension N.

We observe that the grid-based motion features, f_{GRD}^N, and the contour- based motion with shape features, $f_{CTR}^N + f_{SHP}$, achieve the same minimum 12.8% EER, and outperform the contour based only features f_{CTR}^N.

IJERT

Note that the EER performance of speaker identification under the name scenario is significantly better than that of the digit scenario. This is as expected since in
the name scenario each speaker in the database utters a different person-specific phrase, making the identification task easier.

3) Speech-reading Scenario:

The speech reading database D_s is constructed as a subset of speaker identification database
for the name scenario. It includes 35 different phrases, i.e., R=35. Each phrase is a name from the name database with twelve repetitions. The number of source speakers for each phrase varies, we have at least four and at most seven speakers. The D_s database is partitioned into two disjoint sets D_{s1} and D_{s2}, one for training and the other for testing, each having six utterance repetitions [2]. Fig. 5(c) displays the recognition rates for different lip motion representations with varying feature dimension N.

TABLE I
EVALUATION OF THE PROPOSED TWO-STAGE DISCRIMINATION ANALYSIS FOR SPEAKER IDENTIFICATION AND SPEECH-READING

Feature Set	EER (%)		Recog. Rate (%)
	Name	Digit	Speech
$\boldsymbol{f}_{GRD}^{N}$	6.8	12.8	67.62
$\tilde{\boldsymbol{f}}_{GRD}^{N}$	6.5	12.2	**72.86**
$LDA(\boldsymbol{f}_{GRD}^{N})$	5.6	5.8	67.14
$LDA(\tilde{\boldsymbol{f}}_{GRD}^{N})$	**5.2**	**5.2**	67.62
$\boldsymbol{f}_{CTR}^{N}$	10.3	17.4	69.52
$\tilde{\boldsymbol{f}}_{CTR}^{N}$	9.8	17.6	70.00
$LDA(\tilde{\boldsymbol{f}}_{CTR}^{N})$	12.0	18.88	60.95
$\boldsymbol{f}_{SHP}$	18.9	23.5	51.43
$\boldsymbol{f}_{CTR}^{N} + \boldsymbol{f}_{SHP}$	9.2	12.8	70.48
$\tilde{\boldsymbol{f}}_{CTR}^{N} + \boldsymbol{f}_{SHP}$	9.4	13.8	69.52
$LDA(\tilde{\boldsymbol{f}}_{CTR}^{N} + \boldsymbol{f}_{SHP})$	10.4	8.8	61.90

In Table I, we observe that the best performances are obtained using the grid-based motion features for both speaker identification and speech-reading.

III. CONCLUSION AND FUTURE WORK

In this paper, we have investigated different lip motion representations and proposed a two-stage discriminative lip feature selection method for speaker identification and speech-reading.

We have shown by experiments that, for speaker/speech recognition:

- explicit lip motion is useful in addition to lip intensity and/or geometry;
- grid-based dense lip motion features are superior and more robust compared to contour-based lip motion features.

Application for this project is applied for deaf and dumb people who are not able to speak. This application is applied on a large scale.

Apart from this application we are planning to apply this in home security and lockers system. As this application is on a small scale the success ratio increases.

REFERENCES

[1] H. Ertan Çetingül, **Yücel Yemez, A. Murat Tekalp** "Discriminative analysis of Speaker identification and Speech Reading using lip motion" in IEEE transactions on Image processing,2006.

[2] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," in Proc. IEEE Conf. Acoustics, Speech and Signal Processing, 1994, pp. 669 – 672.

IJERT