

# Lip Reading Using DWT and LSDA

Sunil Sudam Morade

PhD Student,  
Department of Electronics Engineering,  
SVNIT, Surat, India.  
ssm.eltx@gmail.com

Suprava Patnaik

Professor,  
Department of E and TC Engineering,  
Xavier Institute of Engineering,  
Mahim, Mumbai, India.  
suprava\_patnaik@yahoo.com

**Abstract—** In lip reading, selection of feature play crucial role. Goal of this work is to compare the common feature extraction modules. Proposed two stage feature extraction technique is exceedingly discriminative, precised and computation efficient. We have used, Discrete Wavelet Transform (DWT) to decorrelate spectral information and extract only the salient visual speech information from lip portion. In the second stage the Locality Sensitive Discriminant Analysis (LSDA) is used to further trim down the feature dimension while preserving the required identifiable ability. A competent feature extraction module result a novel automatic lip reading system. We have compared performance of classical Naive Bayes with the popular SVM classifier. The CUAVE database is used for experimentation and performance comparison. Experimental results show that DWT+LSDA feature mining is better than DWT with PCA or LDA. The performance of Naïve Bayes classifier is exceedingly augmented with DWT+LSDA.

**Keywords—** LSDA; LDA; DWT; SVM; Lip reading

## I. INTRODUCTION

Lip reading is a technique of identifying the speech by visually interpreting the movements of the lips and tongue. Although primarily lip reading was used by deaf and hard-of hearing people, in the recent past computerized lip reading has become one of the most actively researched areas of computer vision because of its crime finding potential and invariance to acoustic noisy environment. The two fundamental steps of lip reading are feature extraction and feature classification. Further feature extraction is made through two basic ways: 1) lip contour guided geometrical model and 2) active appearance guided DCT or DWT models. Performance of lip contour guided geometrical model highly depends on the contour tracking accuracy. It is not suitable for real time application due to accuracy and complexity trade off. Also the geometrical model highly depends on the contour tracking accuracy. It is not suitable for real time application due to accuracy and complexity trade off. Also the geometrical feature guided model fails to include lip inner cavity information. On the other side active appearance model extract features by using gray

scale intensity transformation and is weak in preserving minute geometrical variations. Major challenge faced by the active appearance model are human produces less visual variation as compared to acoustic phonetics, human practice the phonemes right from the childhood however not the visemes. Lip reading accuracy highly depends on ambient light, ability to identify identical visemes associated with different utterances, skill to deal with poor motion or visual difference exhibited by individuals. Initially E. Petajan experimented on lip-reading to enhance speech recognition [1].

In this paper focus is on active appearance model as it has the ability to include cavity information. Notion is to take support from cavity features like percentage appearance of teeth and tongue, particularly when lip motion can't be monitored accurately or even if monitored doesn't help much in estimation. Foremost constrain of appearance model is feature vector size. State of the art literatures deal with DCT or DWT as the foremost step of active appearance model. DCT enlightens about spectral information. Shifting in the DC coefficient is a guideline of teeth visibility, variance among the AC coefficients is used to train the classifier. DCT works well for a limited dictionary, for example vowel detection or decimal points.

Above described transformations are often combined with a dimension reduction stage like Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA), in order to design a realizable system, requiring feasible number of computation for training as well as testing. PCA is a method aims to reduce the data size while minimizing the error between the actual and recovered data. LDA aims to have better classification and preserved the discriminating features. However demand of lip reading systems is fore mostly to concentrate on the lip shape and then do the discrimination analysis. This paper presents a Locality Guided Discriminate Analysis of DWT for lip reading application. Soul contribution is robust feature extraction. Performance of ancient Naive Bays Classifier has been investigated and produced to show the strength of hauled out feature.

The speech reading system proposed by Bregler [2] et al. used Eigen lips as feature vectors. Potamianos et al. compared three linear image transforms namely PCA, DWT and DCT transform techniques [3]. R. Seymour et al. [4] used comparison of image transform features in visual speech recognition of clean and corrupted videos. They evaluated fast

discrete curvlet transform (FDCT), DCT, PCA and LDA methods.

Frame work of proposed lip reading model is described in section-II. Section-III deals with the detail description of LSDA. Experimentation results and description of test corpus is given in section-IV. Finally section-V is based on our conclusion and scope for future work.

## II. PROPOSED LIP READING FRAMEWORK

A typical lip reading system consists of four major stages: video frame normalization, lip localization, feature extraction, and the final step is classifier. Fig. 1 shows the major steps used in the proposed lip reading process. One major challenge in a complete English language lip reading system is the need to train the whole of the English language words in the dictionary or to train (at least) the distinct ones. However same can be effective if it is trained on a specific domain of words, e.g. numbers, postcodes, cities, etc. Present experimentation is limited to digit utterance.

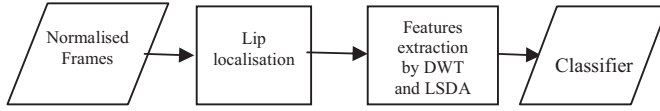


Fig. 1. Lip reading process

### A. Video Segmentation and Lip Contour Localization

There are large inter and intra subject variations in speed of utterance and this results in difference in the number of frames for each utterance. We have used audio analysis, using Pratt software to segment the time duration and the associated video frames of each digit which is uttered. On an average 16 frames are sufficient for utterance of any digit between 0-9. Out of 16 frames we have selected 10 significant frames. Mean square difference  $\sigma_i$ , as given in equation (1) is computed for all the frames and is arranged in decreasing order and initial 10-frames are selected for feature extraction. This step resembles the dynamic time warping operation of speech analysis. Outcome is an optimal alignment of utterances. The number of frames for each utterance is made same such that the feature vectors size remains same for each utterance.

$$\sigma_i = \left[ \frac{1}{M \times N} \sum_0^M \sum_0^N \{I_{(i)}(x, y) - I_{(i+1)}(x, y)\} \right]^2 \quad (1)$$

Where,  $I_i(x, y)$  stands for the  $(x, y)$  spatial location of  $i^{\text{th}}$  video frame and each frame is of size  $M \times N$ .

Lip detection or segmentation is very difficult problem due to the low gray scale variation around the mouth. Chromatic or color pixel based features, especially red domination of lips, have been adopted by most researchers to segment lips from the primarily skin background. However, color representation can be influenced by background lights, and red blobs in the speaker's clothing can cause segmentation failures. We have used Adaboost algorithm for face and mouth detection. A sample result is shown in Fig. 2(a) and Fig 2(b).

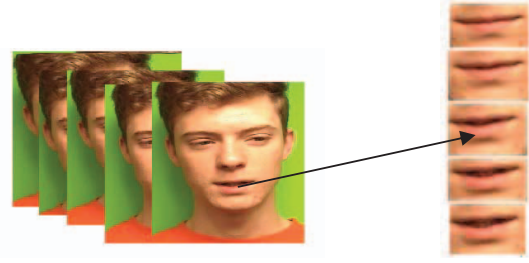


Fig. 2. (a) Detection of face and lip area for CUAVE database s02m(b) Lip portion

Viola and Jones, invented this algorithm in 2004 based on Adaboost classifier to rapidly detect any object including human face. They presented a face detector which uses a holistic approach and is much faster than any contemporaries. Adaboost classifier cascades the Haar like features and not pixels features, hence fast and work accurately for human face detection [5]. Adaboost forms a product of Haar-like operators at each image location and at several scales and then to use the results to train weak classifiers. Single strong object classifier is then formed by cascading these weak classifiers. The advantage of having weak classifiers operating in cascade is that early processing can isolate regions of likely object locations, to bear on these regions in subsequent operations. ROI is a rectangle image containing speaker's mouth and has fixed size.

### B. Frame Normalization and DWT

Image transform methods attempt to transform image pixels of video frame into a new space which separates redundant information and provides better discrimination. Before applying transformation on lip, ROI it is rotated for orientation alignment with respect to a static reference frame and is passed through an LPF to remove high frequency noise.

DWT is used in view of reducing the feature dimension. Goal is to select only those coefficients which play the dominant role in the representation of lip motion. The wavelet transform can be interpreted as a multiscale differentiator or edge detector that represents the singularity of an image at multiple scales and three different orientations — horizontal, vertical, and diagonal. If the singularity is within the support of a wavelet basis function, then the corresponding wavelet coefficient is large. Contrarily, the smooth image region is represented by a cascade of small wavelet coefficients across scale. In standard wavelet decomposition based approach, each level of filtering splits the input image into four parts via pair of low-pass and high-pass filters with respect to column vectors and row vectors of the image array. Then the low-spatial frequency sub-image is selected for further decomposition. After 3-levels of decomposition the lowest spatial-frequency approximation sub-image, a matrix of size  $M/8 \times N/8$  is extracted as the feature vector by matrix to column operation.

## III. PCA, LDA AND LSDA FOR DIMENSION REDUCTION

L.Yaling et al. proposed DCT and LSDA based Feature Extraction for lip reading [7]. In this section, we will introduce

the feature extraction method briefly. DWT is used widely for image compression. Popularly the low frequency sub-band coefficients of DWT are given prior consideration and primarily considered as visual features for lip reading. Applying more decomposition levels will produce smaller feature vector but at the cost of lessening the associated discriminating property. Other commonly used dimension reduction operators are Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA). The proposed lip reading framework is a two step feature extraction technique. The approximation coefficient obtained by applying DWT coefficients are further reduced in size by LSDA to obtain more precised and highly discriminating set of feature.

PCA is the optimal statistical approach used to minimise the mean square error and is obtained when a vector is projected onto the basis function in the principal direction. PCA represents the data in a new coordinate system in which basis vectors, also called the Eigen vectors, follow modes of greatest variance in the data. In image analysis applications, standard practice of obtaining PCA is 2D image array are converted into row or column vectors. By rearranging pixels column by column to a 1D vector, relations of a given pixel to pixels in neighbouring rows are not taken into account. Another disadvantage is in the global nature of the representation; small change or error in the input images influence the whole Eigen representation. In PCA the optimality criterion is to maximize the spread of the resulting feature vectors over all the samples irrespective to number of classes.

#### A. Mathematical formulation of PCA

For good classification we want to maximize the distance between the feature vectors and the Euclidean distance between two vectors is:

$$(\vec{c}_i - \vec{c}_j)^T (\vec{c}_i - \vec{c}_j) \quad (2)$$

In PCA we want to derive a linear transformation  $\Phi$  that maximizes this distance over all the pairs of feature vectors. We want to maximize:

$$\sum_{i=1}^K \sum_{j=1}^K (\vec{c}_i - \vec{c}_j)^T (\vec{c}_i - \vec{c}_j) \quad (3)$$

where K is the number of data points.

Thus in PCA we want to maximize:

$$s_1(\Phi) = \sum_{i=1}^K \sum_{j=1}^K (\vec{c}_i - \vec{c}_j)^T (\vec{c}_i - \vec{c}_j) \quad (4)$$

$$= \sum_{i=1}^K \sum_{j=1}^K (\Phi \vec{f}_i - \Phi \vec{f}_j)^T (\Phi \vec{f}_i - \Phi \vec{f}_j) \quad (5)$$

$s_1(\Phi)$  is the total square distance of all features to each other.  $\Phi$  is the mapping vector. A trivial solution to this maximization problem is one that has components of mapping function  $\Phi$  approaching infinity resulting in expansion of volume. A practical and logical idea is to bind the components of  $\Phi$  to be within a certain range or precisely not permitting for change in scale. A simple way for controlling the range of values of the components of  $\Phi$  is to try to keep its norm as close to unity. We can combine these two optimization goals

into a single optimization criterion using a Lagrange multiplier  $\lambda$ .

$$s_1(\Phi) = \sum_{i=1}^K \sum_{j=1}^K (\Phi \vec{f}_i - \Phi \vec{f}_j)^T (\Phi \vec{f}_i - \Phi \vec{f}_j) - (\|\Phi\|_2 - 1) \quad (6)$$

The 1<sup>st</sup> term controls the spread of the feature points. The 2<sup>nd</sup> term controls that of  $\Phi$ . In other words, we are looking for a linear transformation vector  $\Phi$ , among all possible  $\Phi$ s that maximizes  $s_1(\Phi)$ . That gives:

$$\Phi = \arg \max \sum_{i=1}^K \sum_{j=1}^K (\Phi \vec{f}_i - \Phi \vec{f}_j)^T (\Phi \vec{f}_i - \Phi \vec{f}_j) - (\|\Phi\|_2 - 1) \quad (7)$$

1<sup>st</sup> term can also be written as:

$$s_1(\Phi) = \sum_{i=1}^K \sum_{j=1}^K (\vec{f}_i - \vec{f}_j)^T \Phi^T \Phi (\vec{f}_i - \vec{f}_j) = \sum_{i=1}^K \sum_{j=1}^K g_{ij}^T \Phi^T \Phi g_{ij} \quad (8)$$

Where  $g_{ij} = (\vec{f}_i - \vec{f}_j)$  Using the matrix arithmetic lemma that for a symmetric matrix M:

$$x^T M y = \text{trace}(M x y)$$

and assigning

$$M_{ij} = (\vec{f}_i - \vec{f}_j)^T (\vec{f}_i - \vec{f}_j)$$

and

$$\begin{aligned} \sum_{i=1}^K \sum_{j=1}^K \text{trace} \left( M_{ij} \sum_{k=1}^M \vec{\phi}_k \vec{\phi}_k^T \right) &= \sum_{k=1}^M \vec{\phi}_k^T \sum_{i=1}^K \sum_{j=1}^K M_{ij} \vec{\phi}_k \\ &= \sum_{k=1}^M \vec{\phi}_k^T Q \vec{\phi}_k \end{aligned}$$

Our earlier defined optimization function can be written as

$$s_1(\Phi) = \sum_{k=1}^M \vec{\phi}_k^T Q \vec{\phi}_k - \lambda (\sum_{k=1}^M \vec{\phi}_k^T \vec{\phi}_k - 1) \quad (9)$$

Partial derivative of above equation leads to typical Eigen value and Eigen vector problem, that is

$$\frac{\partial s_1(\Phi)}{\partial \vec{\phi}_k} = 0 \Rightarrow 2Q \vec{\phi}_k - 2\lambda \vec{\phi}_k = 0 \quad (10)$$

$$Q \vec{\phi}_k = \lambda \vec{\phi}_k \text{ where}$$

$$Q = \sum_{i=1}^K \sum_{j=1}^K M_{ij} = \sum_{i=1}^K \sum_{j=1}^K (\vec{f}_i - \vec{f}_j)^T (\vec{f}_i - \vec{f}_j) \quad (11)$$

The matrix  $\Phi$  that maximizes the spread of feature is computed by building the covariance matrix Q, computing its Eigen vectors via singular value decomposition and then using the most significant few Eigen vectors. Eigen vectors of Q become rows of  $\Phi$ .

Thus, for given a signal, PCA look for the attributes which can explain the observed covariance/co-dependence in a set of variables.

## B. Mathematical formulation of LDA

H. Jun et al. used LDA based feature extraction method in DCT domain for lip reading [5]. For good classification results not only the co-dependence but also we often want two complementary conditions to be satisfied. 1) Feature vectors of the same class to be clustered tightly together, to form compact clusters. In other words, within the same class we want small intra class distance. 2) Feature vectors from different classes to be spread far apart from each other, condition to be easily separable. In other words, between the classes we want large inter-class distance.

The measure of intra-class distance is

$$s_2(\Phi) = \sum_{c=1}^P \sum_{i=1}^k \sum_{j=1}^k (\bar{c}_i^c - \bar{c}_j^c)^T (\bar{c}_i^c - \bar{c}_j^c) \\ = \sum_{c=1}^P \sum_{i=1}^k \sum_{j=1}^k (\Phi^c \bar{f}_i - \Phi^c \bar{f}_j)^T (\Phi^c \bar{f}_i - \Phi^c \bar{f}_j) \quad (12)$$

and inter class distance is

$$s_3(\Phi) = \sum_{m=1}^p \sum_{n=1}^p \sum_{i=1}^k \sum_{j=1}^k (\bar{c}_i^m - \bar{c}_j^n)^T (\bar{c}_i^m - \bar{c}_j^n) \\ = \sum_{m=1}^p \sum_{n=1}^p \sum_{i=1}^k \sum_{j=1}^k (\Phi^m \bar{f}_i - \Phi^n \bar{f}_j)^T (\Phi^m \bar{f}_i - \Phi^n \bar{f}_j) \quad (13)$$

where 'P' is the number of classes and 'k' is the number of data points together, to form compact clusters. Ideally we would like to have both minimal intra-class and maximal interclass distance. We could combine these two criteria in a single minimization function using a Lagrange multiplier and minimize the resulting equation,

$$s_4(\Phi) = s_2(\Phi) - \lambda s_3(\Phi) \quad (14)$$

or, alternatively, the intra- and inter-class distance can be combined using ratios, with an aim to maximize the ratio:

$$s_5(\Phi) = s_3(\Phi) / s_2(\Phi) \quad (15)$$

This ratio is used in LDA. Resulting transformation is known as Fisher Transformation. Let  $X_1, X_2, \dots, X_c$  be the classes in the database. For each class if there are  $k$  samples  $x_j$ ,  $j=1, 2, \dots, k$ . Mean of the class  $\mu_i$  can then be computed as

$$\mu_i = \frac{1}{k} \sum_{j=1}^k x_j \quad (16)$$

Mean of all the classes in the database  $\mu$ , is then obtained as:

$$\mu = \frac{1}{c} \sum_{i=1}^c \mu_i \quad (17)$$

As a measure of intra-class variation, practice is to compute the within-class scatter matrix:

$$S_W = \sum_{i=1}^c \sum_{x_k \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T \quad (18)$$

and between the class scatter matrix

$$S_B = \sum_{i=1}^c N_i (\mu_i - \mu)(\mu_i - \mu)^T \quad (19)$$

Product  $S_W^{-1} S_B$  is equivalent to (11). Then the eigenvectors and Eigen values of this product results the required mapping function  $\Phi$ .

## C. Mathematical formulation of LSDA

In this section we introduce the LSDA. LSDA is proposed by Deng Cai, etc in [8], which attempts to study both discriminating and geometrical structure. The framework is based on construction of two graphs, within the class graph  $G_w$  and between the class graph  $G_b$ . A mapping is then carried out so that connected points of  $G_w$  stay as close to each other as possible while connected points of  $G_b$  stay as separated as possible. Fig.3 below is the demonstration of the above concept. The dots being from the same class stay in close proximity after the mapping. The goal is to maximize the margin.

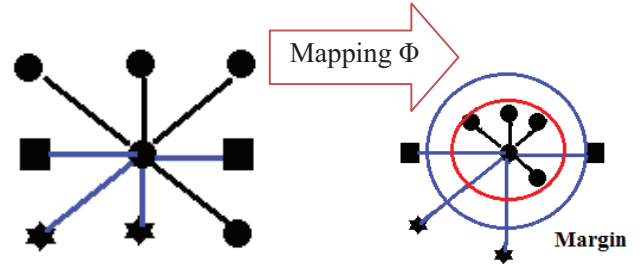


Fig. 3. Margin between class of dots and the rest

For each data point  $f_i$  the set of neighbours  $N(f_i)$  is partitioned into two subsets,  $N_b(f_i)$  and  $N_w(f_i)$ .  $N_w(f_i)$  contains neighbours sharing the same class label with  $f_i$  and  $N_b(f_i)$  contains neighbours sharing the different labels.

Mathematically,

$$N_w(f_i) = \{f_i^j | l(f_i^j) = l(f_i), 1 \leq j \leq k\} \quad (20)$$

$$N_b(f_i) = \{f_i^j | l(f_i^j) \neq l(f_i), 1 \leq j \leq k\} \quad (21)$$

where 'l' is the class level.

Let  $Y = (Y_1, Y_2, \dots, Y_m)^T$  be the mapping outcome and is obtained by projecting input feature set onto basis vector 'a', that is:  $Y^T = a^T f$

The objective function is to optimize the following two conditions

$$\min \sum_{i,j} (y_i - y_j)^2 W_{w,ij} \quad (22)$$

$$\max \sum_{i,j} (y_i - y_j)^2 W_{b,ij} \quad (23)$$

Where  $W_{w,ij}$  and  $W_{b,ij}$  are binary valued weight metrics defined as:

$$W_{b,ij} = \begin{cases} 1, & f_i \in N_b(f_j) \\ 0, & \text{otherwise} \end{cases} \quad (24)$$



$$W_{w,ij} = \begin{cases} 1, & f_i \in N_w(f_j) \\ 0, & \text{otherwise} \end{cases} \quad (25)$$

The objective functions can be reduced to

$$\max_a (a^T f W_w f^T a) \quad (26)$$

$$\max_a (a^T f L_b f^T a) \quad (27)$$

Where  $L_b = D_b - W_b$  and is called the Laplacian matrix.  $D_b$  is the diagonal matrix, where  $D_{b,ij} = \sum W_{b,ij}$ . The matrix  $D_w$ , provides a natural measure, bigger is its value implies that the class containing  $f_i$  has high density around  $f_i$ . Therefore, a constraint is imposed as follows:

$$a^T f D_w f^T a = 1 \quad (28)$$

Finally the two objective functions reduce to a single optimization problem by including a cost factor  $\alpha$ . The projection vector 'a' that satisfies the above requirement is given by the maximum Eigen value solution to the generalized Eigen value problem:

$$f(\alpha L_b + (1 - \alpha) W_w) f^T a = \lambda f D_w f^T a \quad (29)$$

Mapping is obtained by  $f_i \rightarrow y_i = A^T f_i$ , where

$A = [a_1, a_2, a_3, \dots, a_d]$ , and Eigen vectors  $a_1$  to  $a_d$  are ordered according to their Eigen values  $\lambda_1 > \lambda_2 \dots > \lambda_d$ .

#### IV. PROPOSED LIPREADING METHODOLOGY

Final Stage of lip reading system is feature classification and assigning a class number to the associated video sequence. There are varieties of classifiers available. Selection of appropriate classifier is crucial. The two classifiers investigated and compared in this work are the classical Naive Bayes (NB), a probabilistic method and the recently introduced Support Vector Machine (SVM), a projection optimization approach. Surprisingly SVM is not a clear winner.

Primarily SVM aims to find the decision surface that maximizes the margin between the data points of the two classes. Two class linear SVM optimization problem aims to estimate parameters  $\alpha_i$  while maximizing the function

$$= \sum_{i=1}^k \alpha_i - \frac{1}{2} \sum_{i=1}^k \sum_{j=1}^k c_i c_j \alpha_i \alpha_j \langle f_i, f_j \rangle \quad (30)$$

Subject to  $\sum_{i=1}^k c_i \alpha_i = 0$ , where  $c_i = [1, -1]$  is the binary valued class number for feature points  $f$ . Class prediction function  $\Phi$  of the test data point  $f'$  is given by

$$\Phi(f) = \text{sign} \left( \sum_{i=1}^k \sum_{j=1}^k c_i \alpha_i \langle f_i, f' \rangle + b \right)$$

where 'b' is the bias or off-set value.

In Naive Bayes classifier if  $P(c_i)$  is the prior probability of class  $c_i$  and  $P(f_x^k | c_i)$  is the conditional probability to observe the attribute  $f_i^k$  given the class  $c_i$ , a data point  $f_x$

with attributes  $(f_x^1, \dots, f_x^2, \dots, f_x^d)$  is assigned class ' $c_i$ ' while maximizing  $\Phi(f')$ :

$$\Phi(f') = \arg \max (P(c_i) \prod_{i=1}^d P(f_x^i | c_i)) \quad (32)$$

A flow chart of the steps involved in our simulation technique is shown in Fig. 4. Three major execution steps of the algorithm are: 1) pro-processing, 2) feature extraction and dimension reduction, and 3) feature classification. The two step salient feature extraction step is the core contribution of our work. After DWT, we are taking only low frequency components (LL) of the image for further dimensionality reduction by using LSDA. Then the final feature vectors of all the train images are stored in the training database along with class level. DWT attempts to transform image pixels of significant lip frame into a new space which separates redundant information and provides better discrimination.

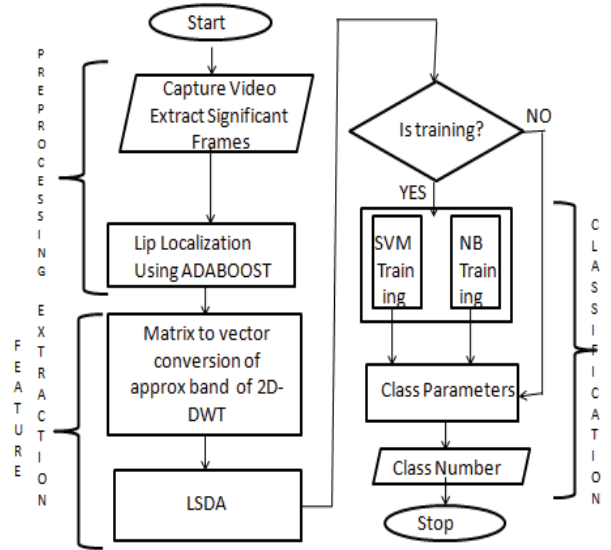


Fig. 4. Flowchart for DWT-LSDA lip reading implementation

#### V. CORPUS AND RESULT

##### A. CUAVE database

We performed our simulation on CUAVE database. CUAVE database is a standard data base and its video frame rate is 30frames/sec. It contains mixture of white and black features. Database digits are continuous and with pause also. Data is recorded with sequential and random manner. Some videos are taken from side view. Total 36 videos are in data base, out of which, 19 are male and 17 are female. It has been shown that this does not significantly affect the collection of visual features for lip reading [9].

## B. Feature extraction

Before applying transformation on lip ROI it is rotated for orientation alignment with respect to a static reference frame, lip area localized to size 32 x 20 and passed through an LPF to remove high frequency noise. In proposed experimentation 2D DWT with three decomposition levels are applied to lip area. In DWT-db4 30 (5 x 3) Coefficients are generated for per frame. Total Coefficients are calculated for 10 normalized frames. This results in a feature vector of size 150 x 1.

Seven users and each one uttering each digit 5 times produces 350 x 150 dimensional training dataset Feature vectors are leveled with 10 different classes each class corresponding to a digit. After applying PCA, LDA and LSDA technique feature size reduces to highly discriminating much smaller vectors of size 15 x 1.

## C. Results

This section deals with the results. Table 1 shows confusion matrix. Nine has confusion with six though they acoustically very different. Zero has confusion with six. Table 2 shows that four is most recognised digit. Table 3 indicates DWT and PCA with naïve base classifier perform best.

TABLE 1

CONFUSION MATRIX FOR SAMPLES OF DIGIT UTTERANCE FOR DWT WITH LSDA WITH SVM CLASSIFIER

Digits	0	1	2	3	4	5	6	7	8	9
0	30	0	0	1	0	0	3	1	0	0
1	0	30	0	1	1	1	0	0	0	2
2	0	0	32	0	0	1	1	2	0	0
3	0	1	0	34	0	0	0	0	0	0
4	0	0	0	0	35	0	0	0	0	0
5	1	1	0	1	0	31	1	0	0	0
6	2	0	0	1	0	0	31	1	0	0
7	4	0	1	0	0	1	0	28	0	1
8	0	0	0	0	0	1	0	0	33	1
9	1	0	1	0	0	1	3	0	1	28

TABLE 2

CLASSIFICATION RATE FOR DIFFERENT DIGIT FOR SVM AND NAIVE BAYES

Classifier	Average Recognition Rate(%) for 0 – 9 digits									
	0	1	2	3	4	5	6	7	8	9
SVM	85	85.7	91.4	97	100	88	88	80	94	80
Naive Bayes	91	85	94	97	100	91	94	88	97	88.6

TABLE 3  
COMPARISON OF RESULT WITH PCA, LDA AND LSDA

TYPE OF TRANS.	SVM	NAIVE BAYES
DWT+PCA	70.56	60.66
DWT+LDA	88.33	89.14
DWT+LSDA	90.28	91.85

## VI. CONCLUSION

In this paper, we have compared DWT +PCA, DWT +LDA, DWT +LSDA features for lip reading. DWT+LSDA feature performance is better. SVM and Naive Bayes are trained for feature classification. A subset from CUAVE data base consisting of 7 speakers with front view for digits 0-9, uttered ten times is used for performance comparison. Among the digits, '4' is found as most discriminative and has been always acknowledged. SVM are by default binary classifiers. To deal with multiclass problems practice is to deal with many SVM classifiers (one less than the class number), operated in cascade and one versus rest strategy is followed. However multi-class Naive Bays has no such constrain.

There are some ways we can further advance the work of this paper. Further, experimentation to use shape normalized separate inner appearance traits along with geometric visemes of lip is needed. Motion amplification may make the performance robust and noise invulnerable.

## REFERENCES

- [1] E. D. Petajan, "Automatic lip-reading to enhance speech recognition", Ph.D. Thesis University of Illinois, 1984.
- [2] C. Bergler and Y. Konig, "Eigenlips" For robust speech recognition," in *Proc. IEEE Int. Conference on Acoustics, Speech and signal processing*, 1994.
- [3] G. Potamianos, H. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lip reading," *International Conference on Image Processing*, 173–177, 1998.
- [4] R. Seymour, D. Stewart, and Ji Ming, "Comparison of image transform-based features for visual speech recognition in clean and corrupted videos," *EURASIP Journal on Video Processing*, Vol. 2008, 1- 9, 2008.
- [5] P. Viola, M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple features", *IEEE Int. Conference*, 511-517, 2001.
- [6] H. Jun, Z. Hua, I. Jizhong, "LDA based feature extraction method in DCT domain in lip reading", *computer Engineering and application*, 45(32), 150-152, 2009.
- [7] L.Yaling, Y. Wenjuan, D. Minghui, "Feature Extraction Based on LSDA for lip reading", *Proceedings of IEEE International conference*, 2010.
- [8] Deng Cai, X. He, K. Zhou, J.Han, H. Bao, "Locality discriminant analysis", *International joint conference on artificial Intelligence Hyderabad Morgan kauffmann Publishers*, 2007.
- [9] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "CUAVE: a new audio-visual database for multimodal human computer- interface research", *Proceedings of IEEE International conference on Acoustics, speech and Signal Processing*, 2017-2020, 2002.