Visual Lip Reading using 3D-DCT and 3D-DWT and LSDA

Sunil S. Morade
PhD Student
Dept of Electronics Engg SVNIT,Surat
Asso.Professor, KKWIEER, Nashik

Suprava Patnaik
Professor
Dept. E and Tc Engg.,
Xavier Institute of Engineering,
Mahim, Mumbai, India.

ABSTRACT

Human uses visual information while trying to understand speech, especially in noisy conditions or in situations where the audio signal is not available. Lip reading is the technique of a comprehensive understanding the underlying speech by processing on the movement of lips. However, the recognition of lip motion is a difficult task since the region of interest (ROI) is nonlinear and noisy. In proposed method lip reading system we have used two stage feature extraction model which is precised, discriminative and computation efficient. The first stage 3D Discrete Wavelet Transform (3D-DWT) or 3D Discrete Cosine Transform (3D-DCT) is used and the second stage is Locality Sensitive Discriminant Analysis (LSDA) to trim down the feature dimensions. These features make a novel lip reading system with small feature vector size. In addition to the novel feature extraction technique, the performance of Naive Bayes and SVM classifier is compared. CUAVE database of 0 to 9 utterances in English is used for experimentation. Results of 3 dimension transform with LSDA are compared with 2 dimension transform with LSDA. Experimental results show that 3D-DWT+LSDA feature mining are compared with 3D-DWT with PCA or LDA. 3D-DWT+LSDA result is also compared with 3D-DCT + LSDA.

Keywords

LSDA, LDA, 3D-DWT, 3D-DCT, SVM, Naive Bayes, Lip reading.

1. INTRODUCTION

Automatic lip reading is an active research area. Lip reading is a speech reading process from visual information of lip and areas around the mouth. Automatic acoustic speech recognition systems tend to perform poorly in noisy environment. Audio information is affected by acoustic noise and crosstalk among speakers. Whereas visual modality provides information which is robust and therefore there have been numerous studies going on to assess and improve the performance of visual speech recognition. Two fundamental steps are involved in lip reading one is feature extraction and other is feature classification. Lip features are extracted by a geometrical model and image transform method. Lip geometric model will depend on contour along the lip portion, so variation in extraction of lip contour affect the different parameters such as width, height and area. Because of accuracy and complexity geometrical model is not suitable for real time application. Also in this model cavity information is not taken into account. In this paper we focus on image transform model. On the other side image transform model extracts feature by using gray scale intensity transformation and is weak in preserving minute geometrical variations. State of the art literatures deal with DCT or DWT as the foremost step of appearance model. Important constraints of the image transform techniques is the vector size of feature vector. Major challenge faced by the lip reading model are human produces less visual variation as compared to acoustic phonetics, human practice the phonemes right from the childhood however not the visemes. E. Petajan [1] experimented on lip-reading to enhance speech recognition. Matthews et al. [2] worked on DCT, DWT and PCA image transform methods. The image transform methods have been compared with Active Appearance Model (AAM) and found that Image transform methods have highest recognition accuracy as compared to AAM.

In this paper focus is on visual appearance model as it has the ability to include cavity information. Idea 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. 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. Environmental changes such as change in lighting condition, scale, rotation etc. affect the results of lip reading. DWT offers time scale analysis, hence can amend for change in scale and lighting while preserving the spectral information. Above described transformations are often combined with a dimension reduction stage like Linear Discriminant Analysis (LDA) or Principal Component Analysis (PCA) , in order to design a realizable system, requiring feasible number of computation for training as well as testing. In this paper LSDA is combined with 3D-DCT or 3D-DWT. Both the results are compared with classifier Naive Bayes and SVM. The speech reading system proposed by Bregler et al. [3] used Eigen lips as feature vectors. In experimentation Potamianos et al. [4]compared three linear image transforms namely PCA, DWT and DCT transform techniques. Potamianos et al. [5]have used DCT and LDA transform for real time lip reading application. They conclude that computation of feature extraction require very less time as compare to other processing such as face detection, lip detection and preprocessing of image.

Seymour et al. [6] compared different image transform methods for clean and corrupted video. They added three types noise namely: compression, blurred and Jitter in clean video. They concluded that DCT feature extraction method performance is better for first two noises of corrupted videos but worst for last. They have tested their result on digit as database and found that for their database “one” recognized most correctly while “nine” is more confused with “six”.

Wang et al. [7] used different region of interest (ROI) as a visual features in lip reading process. Authors discussed about different ROI processing methods expressed its impact on recognition accuracy. N. Puviarasan et al. [8] used DCT and DWT methods for visual feature extraction. They used data base of hearing impaired person and observed that DWT with HMM gives better result. A. Shaikh et al. [9] used optical flow information as a feature vector for lip reading. The vocabulary used in their experiment was viseme. Visemes are the basic visual movements associated with phonemes. They tested the result of lip reading using SVM classifier with kernel function consisting of Gaussian Radial Basis function. They used classification performance parameters such as specificity, sensitivity and accuracy to test classifiers.

Meyor et al. [10] used DCT transform technique for pixel information of continuous digit recognition and proposed different fusion techniques for audio and video feature data. They found that Word Error Rate (WER) is more for continuous digit recognition. L. Rothkrantz et al. [11] presented a lip geometry estimation (LGE) method and it was compared with geometry and image intensity based techniques such as geometrical model of lip, specific points on mouth contour and raw image. Authors found LGE method competitive with some strong points on its favor.

Heckman et al. [12] investigated selection of the DCT coefficients influences the recognition scores in a hybrid ANN/HMM audio-visual speech recognition system. In their experiment, they found that 30 DCT coefficients are sufficient for recognition. While selecting the coefficient, they found that energy feature of coefficients gives a slightly better performance.

Frame work of proposed lip reading model is described in section-2. Section-3 deals with feature extraction using DCT and DWT with PCA, LDA, and LSDA. Section 4 deals with classifiers. Section 5 deals with lip reading methodology. Experimentation results and description of test corpus is given in section-5. Finally section-8 is based on our conclusion and scope for future work.

2. PROPOSED LIP READING FRAMEWORK

2.1 Proposed lip reading system

A typical lip reading system consists of four major stages: video frame normalization, Face and lip detection, feature extraction, and the final step is classifier. For feature extraction both methods i.e. DCT +LSDA and DWT +LSDA are used. 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.

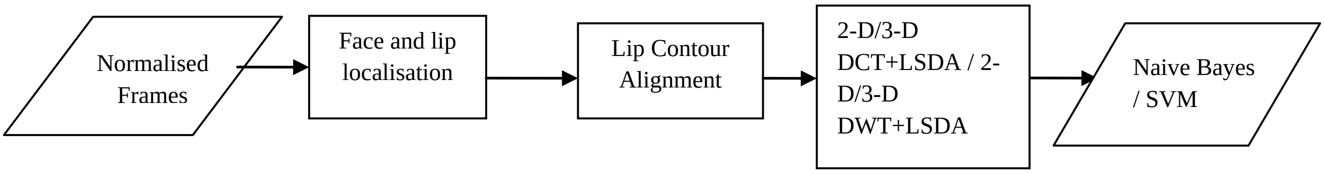
Normalised Frames
Face and lip localisation
Lip Contour Alignment
2-D/3-D DCT+LSDA / 2-D/3-D DWT+LSDA
Naive Bayes / SVM

Fig. 1 Lip reading process using 2-D and 3-D transform

2.2 Video Segmentation and Lip Contour Localization

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. Viola and Jones [13], invented this algorithm in 2004 based on Adaboost classifier to rapidly detect any object including human face. Procedure to align lip from different frames is to find two extreme locations from left and right locations in the x direction from the lip contour of Fig. 2(a). Rotate face by an angle θ. This is shown in Fig. 2(b).Using algorithm developed by Viola and Jones mouth detection result is shown in Fig. 2(c) and is passed through an LPF to remove high frequency noise.

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 σ_i, 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 remain same for each utterance.

3. FEATURE EXTRACTION

3.1 Discrete Cosine Transform (DCT) for visual lip analysis

2D –DCT features are extracted by using equation (2). The values B_{pq} are called the DCT coefficients of A at location (p, q). The DCT technique of image transform is used to transform lip area into DCT coefficients. Only 28 coefficients per frame are used out of 22 x 32 DCT coefficients. To select DCT coefficients, upper triangular mask is preferred over rectangular mask because it gives lower frequency component information. To select DCT coefficients consists of DC and 27 AC coefficients.

$$B_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{mn} \cos\frac{\pi(2m+1)}{2M} \cos\frac{\pi(2n+1)}{2N}) \ldots (2)$$

$$\alpha_p = \left\{ \begin{matrix} \frac{1}{\sqrt{M}}, & p = 0 \\ \sqrt{2/m}, & 1 \le p \le M-1 \end{matrix} \right\}$$

$$\alpha_q = \left\{ \begin{matrix} \frac{1}{\sqrt{N}}, & q = 0 \\ \sqrt{2/n}, & 1 \le q \le M-1 \end{matrix} \right\}$$

3.2 Discrete Wavelet Transform (DWT) for visual lip analysis

While computing 2D-DWT feature vector, the goal is to select only those coefficients which play the dominant role in the representation of lip motion. In standard 2D 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 few levels of decomposition the lowest spatial-frequency approximation sub-image, is extracted as the feature vector.

Total average power by each coefficient for a digit is given by equation (4). The average power calculated using DCT coefficient is P_{avec} and similar way the average power is calculated for DWT coefficient. Average normalised power in DCT and DWT coefficient is compared. Result shows that DWT coefficients preserve energy that is 88.84 % which is more as compared to DCT coefficients which preserve energy 53.98 %. Result shows that DWT coefficients are more significant as compared to DCT coefficients.

$$P_{dct} = \sum_{q=0}^{M} \mathrm{B}_{\mathrm{q}}^{2} \ldots (3)$$

Total Average power by DCT coefficient for a digit is given by (4)

$$P_{avec} = \frac{1}{10} \sum_{nd=0}^{9} \mathrm{P}_{\mathrm{nd}} \ldots . (4)$$

Original lip area is shown in Fig. 3(a) and Fig 3(b) is the normalized coefficient of three levels. Small rectangle in Fig 3(b) indicates the selected coefficients of LL3 band are used as feature vector.

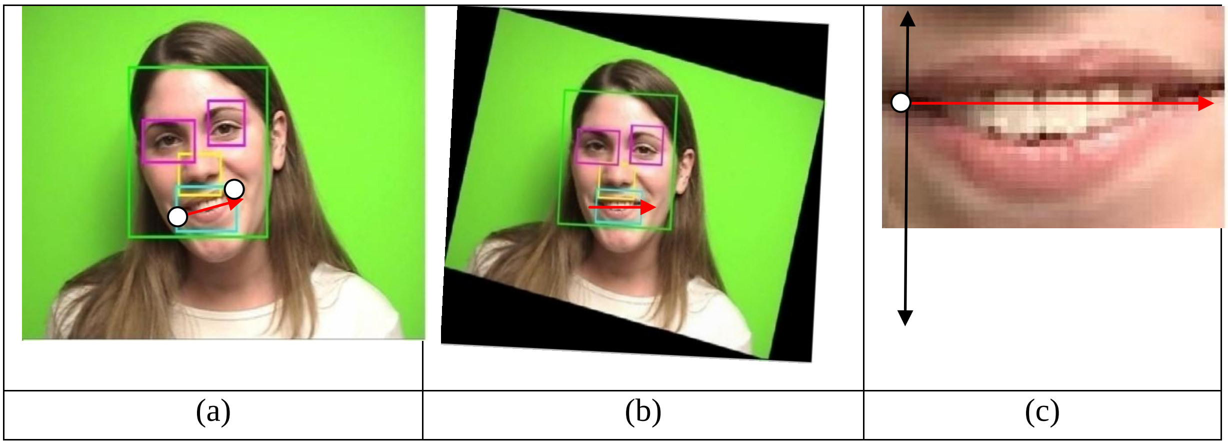

Fig.2 Cropped lip images (a) Original Face image (b) Face shifted by an angle of 11 degree (c) Cropping of lip bounding box

.

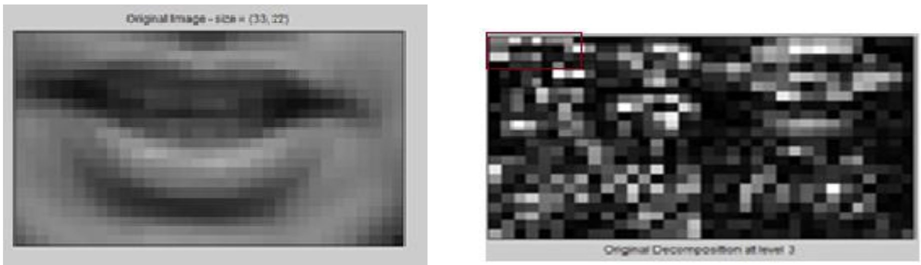

Fig. 3(a) Original Lip image

(b) Display of three levels coefficients

3.3 3D-DWT for lip feature extraction

Previous result feature vectors generated using 3D-DWT with Demy is better than 3D-DWT with other wavelets [14]. So these feature vectors are applied for three discrimination techniques. The 3D-DWT is like a 1-D DWT in three directions. lip reading is a video processing application. To use the wavelet transform for volume and video processing, a 3-D version of filter banks are implemented. In 3D-DWT, the 1D analysis filter bank is applied in turn to each of the three dimensions [15]. Mathematical expression for single dimension DWT is based on equation (5).

$$W_l(n,j) = \sum_{m=0}^{2n} w(m, j-1) * h(2n-m) \qquad (5)$$

$$W_h(n,j) = \sum_{m=0}^{2n} w(m, j-1) * g(2n-m) \qquad (6)$$

where $\mathrm{W}(\mathrm{n},\mathrm{j})$ is wavelet output. h(n) and g(n) are the filter impulse response of low pass and high pass filter, j is the current level, n is the current input index and $\mathrm{w}(\mathrm{n}, \ \mathrm{j}-1)$ is the input signal.

3.4 3 D-DCT for visual lip analysis

1-D transform is used in speech and music, while 2D is used in image processing. In order to extract dynamic features of lip motion 3D-DCT is considered. Y. Fan et al. [16] compare the energy distribution of 3D-DCT and 2D-DCT. The author concluded that in the coefficients of 3D-DCT contains more concentration of energy with less number coefficient than 2D-DCT. K. Min et al. [17] used 3D-DCT for lip feature extraction and 3D-HMM as classifier.

4. PCA, LDA AND LSDA FOR DIMENSION REDUCTION

L.Yaling et al. proposed DCT and LSDA based feature extraction for lip reading [18]. 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 Discriminant Analysis (LDA). The proposed lip reading framework is a two step feature extraction technique. The approximation coefficient obtained by applying DWT and DCT coefficients are further reduced in size by LSDA to obtain more precised and highly discriminating set of feature.

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 neighboring rows are not taken into account. Another disadvantage is in the global nature of the representation; small change or error in the input images influences 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. 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:

$$\left(\vec{c_i} - \vec{c_j}\right)^T \left(\vec{c_i} - \vec{c_j}\right) \ldots. \qquad (7)$$

In PCA we want to derive a linear transformation Φ that maximizes this distance over all the pairs of feature vectors.

$$Q\overrightarrow{\boldsymbol{\phi}_k} = \boldsymbol{\lambda}\overrightarrow{\boldsymbol{\phi}_k}$$

where

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{k} M_{ij} \quad = \sum_{i=1}^{k}\sum_{j=1}^{k} \left(\vec{f_i} - \vec{f_j}\right)^T \left(\vec{f_i} - \vec{f_j}\right) \ldots..(8)$$

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

4.1 Mathematical formulation of LDA

H. Jun et al. used LDA based feature extraction method in DCT domain for lip reading [19]. For good classification results not only the co-dependence but also 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. This ratio is used in LDA. Let $X1, X2, \ldots, Xc$ be the classes in the database. For each class if there are k samples $x_j, j = 1,2, \ldots, k.$ Mean of the class μ_i can then be computed as

$$\mu_i = \frac{1}{k}\sum_{j=1}^{k} x_j \ldots..(9)$$

Mean of all the classes in the database μ, is then obtained as:

$$\mu = \frac{1}{c}\sum_{i=1}^{c} \mu_i \ldots..(10)$$

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

$$S_W = \sum_{i=1}^{c}\sum_{x_k \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T \ldots..(11)$$

and between the class scatter matrix

$$S_B = \sum_{i=1}^{c} N_i(\mu_i - \mu \quad)(\mu_i - \mu \quad)^T \ldots..(12)$$

Product $S_W{}^{-1}S_B$ is equivalent to equation (8). Then the eigenvectors and eigen values of this product results the required mapping function Φ.

4.2 Mathematical formulation of LSDA

In this section we introduce the Locality Sensitive Discremination Analysis (LSDA). LSDA is proposed by Deng Cai, et al. [20], which attempts to study both discriminating and geometrical structure. The framework is based on construction of two graphs, within the class graph Gw and between the class graph Gb a mapping is then carried out, so that connected points of Gw stay as close to each other as possible while connected points of Gb stay as separated as possible. In Fig 5.7 Dots, stars and blocks corresponds to feature vectors from three different classes. The dots being from the same class stay in close proximity after the mapping. The goal is to maximize the margin ‘m’.

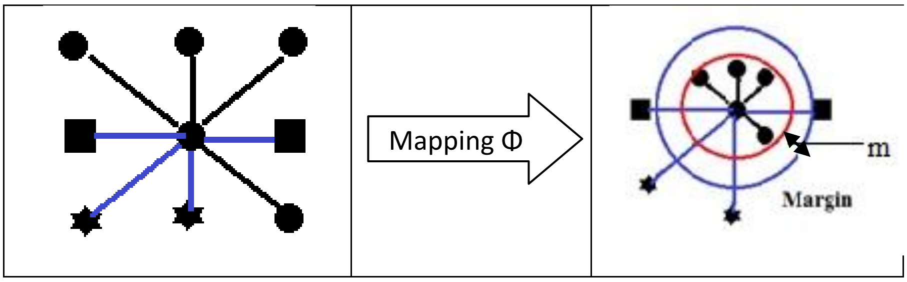

Fig 4 Margin between class of dots and the rest

For each data point fi the set of neighbours N(fi) is partitioned into two subsets, Nb(fi) and Nw(fi). Nw(fi) contains neighbours sharing the same class label with fi and Nb(fi) contains neighbours sharing the different labels. Mathematically,

$$N_w(f_i) = \{f_i^j \,|\, l(f_i^j) = l(f_i), 1 \leq j \leq k\} \ldots . (13)$$

$$N_b(f_i) = \{f_i^j \,|\, l(f_i^j) \neq l(f_i), 1 \leq j \leq k\} \ldots . (14)$$

where ' l ' is the class level.

Let Y = $(Y_1, Y_2, \ldots Y_m)^T$ 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

$$min \sum_{i,j} (y_i - y_j)^2 \, W_{w,ij} \; \ldots . . (15)$$

$$max \sum_{i,j} (y_i - y_j) \;\; ^2 \, W_{b,ij} \; \ldots . (16)$$

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

$$W_{b,ij} = \begin{cases} 1, & f_i \in N_b(f_i) \\ 0, & otherwise \end{cases} \quad (17)$$

$$W_{w,ij} = \begin{cases} 1, & f_i \in N_w(f_j) \\ 0, & otherwise \end{cases} \quad (18)$$

The objective functions can be reduced to

$$\max_a (a^T \, f W_w f^T a) \quad (19)$$

$$\max_a (a^T \, f L_b f^T a) \quad (20)$$

Where L_b = Db - Wb and is called the Laplacian matrix. Db is the diagonal matrix, where

The matrix Dw, 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:

$$a^T f D_w f^T a = 1 \quad (21)$$

Finally the two objective functions reduces to a single optimization problem by including a cost factor α. The projection vector 'a' that satisfies the above requirement is given by the maximum eigen value solution to the generalized eigenvalue problem:

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

Mapping is obtained b $f_i \rightarrow y_i = A^T f_i$, where $A = [a_1, a_2, a_3 . . , a_d]$ and eigen vectors a_1 to a_d reordered according to their eigen values.

5. CLASSIFICATION

5.1 Support Vector Machines (SVM)

Data separation is completely possible by using nonlinear separation but it is not using linear separation. In SVM for nonlinear separation between classes mapping function Φ is used. Using Φ lower dimension input space is transferred to higher dimension. Mapping is projecting the original set of variables x in higher dimensional feature space Φ. Kernel functions is given by (22) scalar product $\Phi^T(x_i)\Phi(x_j)$. Applying kernels we do not even have to know what the actual mapping. A kernel is a function k such that the learning algorithm is a hyperplane in a feature space. Thus by choosing kernel $k(\boldsymbol{x}, \boldsymbol{x_i})$, we can construct an SVM that operates in an infinite dimension space.

$$x \in R^d \Rightarrow \boldsymbol{\Phi}(x)$$

$$\boldsymbol{\Phi}(x) \equiv (\Phi_1(x), \Phi_2(x), \ldots \ldots , \Phi_n(x)) \in R^n$$

$$k(\boldsymbol{x_i}, \boldsymbol{l_j}) = \boldsymbol{\Phi^T}(x_i)\boldsymbol{\Phi}(l_j) \quad (22)$$

Kechman in his literature discussed the mapping and different kernel function [21]. SVM maximizes the distance of separating plane from the closest training data point. Linear kernel is defined by (22). Polynomial kernel is defined by (23) where d is degree of polynomial.

$$k(\boldsymbol{x}, \boldsymbol{l}_i) = (\boldsymbol{x^T l}_i) \qquad (22)$$

$$k(\boldsymbol{x}, \boldsymbol{l}_i) = (\boldsymbol{x^T l}_i + 1)^d \qquad (23)$$

In equation (24) we need to find suitable Lagrange multipliers α to get the following function reach its maximum value.

$$L_d(\alpha) = \sum_1^l \alpha_i - \frac{1}{2} \sum_1^l y_i \alpha_i \alpha_j \Phi_i{}^T \, \Phi_j \quad (24)$$

Where $\quad k(x_i, l_j) = \alpha_j \Phi_i{}^T \, \Phi_j$

For classification, class decision is based on the f (x) value is given by (25).

$$f(x) = \sum_{i=1}^N y_i \alpha_i \, k(x, l_i) + b) \quad (25)$$

where k kernel function, b scalar bias, α langrage's multiplier, y is output and support vector is obtained from training data.

5.1.1 Sequential Minimal Optimization (SMO)

Sequential Minimal Optimization (SMO) is a SVM learning algorithm which is conceptually simple, easy to implement, and have faster and better scaling properties than a standard SVM algorithm. John Platt [22] proposed this algorithm for efficiently solving the optimization problem which arises during the training of SVM. Though SVMs are popular, two major weaknesses made their use limited. First the training of SVM is slow, especially for large problems. Second, SVM training algorithms are complex, subtle and sometimes difficult to implement. E. Osuna et al. [23] has suggested two modifications in Platt's SMO algorithm so that the SMO algorithm is speed up to train SVM in many situations. Because no matrix algorithms are used in SMO, it is less susceptible to numerical precision problems. For the real-world test sets, SMO can be a factor of 1200 times faster for linear SVMs and a factor of 15 times faster for non-linear SVMs . Because of its ease of use and better scaling with training set size, SMO has become the standard SVM training algorithm. In this experiment SMO is used for training SVM with 2nd degree polynomial kernel function.

5.2 Naive Bayes

In his paper T. M. Mitchell compared Gaussian Naïve Bayes with logistic regression [24]. One of the difficulties in training an algorithm is deciding which features to include and also simply the number of feature vectors involved. The size of training data increases exponentially with the number of features used, in other words the more features a system employs the more training data it requires. The Naive Bayes classifier greatly simplifies learning by assuming that features are independent given class. Although independence is generally a poor assumption, in practice Naive Bayes often competes well with more sophisticated classifiers. To design learning algorithms based on Bayes rule. Consider a supervised learning problem in which we wish to approximate an unknown target function $f : X \rightarrow Y$, or equivalently $P(Y|X)$. To begin, we will assume Y is a Boolean-valued random variable, and X is a vector containing n Boolean attributes. In other words, $X = \langle X_1, X_2, \dots, X_n \rangle$, where Xi is the Boolean random variable denoting the ith attribute of X. Applying Bayes rule, we see that $P(Y = y_i|X)$ can be represented as

$$P(Y = y_i|X = x_k) = \frac{P(X = x_k|Y = y_i)P(Y = y_i)}{\sum_j P\left(X = x_k\middle|Y = y_j\right)P(Y = y_j)} \quad (26)$$

where y_m denotes the mth possible value for Y, x_k denotes the kth possible vector value for X, and where the summation in the denominator is over all legal values of the random variable Y. One way to learn $P(Y|X)$ is to use the training data to estimate $P(X|Y)$ and $P(Y)$. We can then use these estimates, together with Bayes rule above, to determine $P(Y|X = x_k)$ for any new instance x_k.

The Naive Bayes algorithm is a classification algorithm based on Bayes rule, that assumes the attributes X1,…,Xn are all conditionally independent of one another, given Y. The value of this assumption is that it dramatically simplifies the representation of $P(X|Y)$, and the problem of estimating it from the training data. More generally, when X contains n attributes which are conditionally independent of one another given Y, we have

$$P(X_1 \dots X_n|Y) = \prod_{i=1}^{n} P(X_i|Y) \quad (27)$$

Multiclass problem is solved by doing the comparison between the two classes. Probability $P1$ and $P0$ depends on mean and variance. Variance σ of LSDA and LDA coefficients is more as compare to other transform methods as shown in Fig. 6, so classification rate is more for Naïve Bayes. As σ increases w_0 and w_i decreases $P1$ is more in equation (28). So probability difference of getting output 1 and 0 is more, classification rate in case of Naive Bayes is more with feature vector of LSDA and LDA.

$$P1 = P(Y = 1|X) = \frac{1}{1 + exp\{w_0 + \sum_1^n w_i X_i\}} \quad (28)$$

Where the weights $w_1, \dots \dots . w_n$ are given by

$$w_i = \frac{\mu_{i0-}\mu_{i1}}{\sigma^2}$$

$$w_0 = \ln\frac{1-\pi}{\pi} + \sum_{\mathrm{i}}^{\mathrm{n}} \frac{{\mu_{i1}}^2 - {\mu_{i0}}^2}{2{\sigma_i}^2}$$

$$P0 = P(Y = 0|X) = 1 - P(Y = 1|X) \quad (29)$$

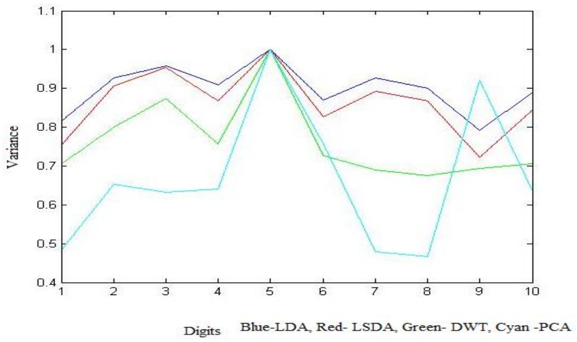

Fig. 6 Plot of normalized variance of coefficients vs. digit

6. PROPOSED LIP READING SYSTEM

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.

Following steps are used in Algorithm

1. Capture the video & separate it into frames
2. Mark the frames for the duration of utterance of digits.
3. Extract 10 significant frames for each digit utterance.
4. Detect the face and lip area and separate the lip portion.
5. Extract the features using 2D and 3D transform.
6. Apply PCA / LDA / LSDA.
7. Using training set, train SVM and Naive Bayes. Find out the class parameter.
8. Using testing set, SVM and Naive Bayes is tested to find the class number.

7. CORPUS AND RESULT

7.1 CUAVE database

CUAVE [25] (Clemson University Audio Visual Experiments) was recorded by E. K. Pattererson of Department of Electrical and Computer Engineering, Clemson University, US. The database was recorded in an isolated sound room at a resolution of 720 x 480 with frame rate of 29.97 fps. This database is a speaker-independent database consisting of connected and continuous digits spoken in different situations. It contains mixture of speaker with white and black skin. Database digits are continuous and with pause. 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 for male speaker and 17 are for female speaker.

7.2 Use of 2D-DWT and 3D-DWT and discrimination technique for feature extraction

Transformation is applied on lip ROI which aliened 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 DCT are applied to lip area. In DCT 30 Coefficients are taken from upper triangle of total DCT Coefficients. Total coefficients are calculated for 10 normalized frames. This results in a feature vector of size 300 x 1. Seven users and each one uttering each digit 5 times produces 350 x 300 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 discreminating much smaller vectors of size 30 x 1. In proposed experimentation 2D DWT with three decomposition levels are applied to lip area. In DWT-db4 30 (5 x 6) Coefficients are generated for per frame. Total coefficients are calculated for 10 normalized frames. This results in a feature vector of size 300 x 1. Seven users and each one uttering each digit 5 times produces 350 x 300 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 discreminating much smaller vectors of size 30 x 1. Similar way 3D-DWT transform we get feature vector size of 350 x164 and for 3D-DCT feature vector size of 350 x 149.

8. RESULTS

Mean of 35 feature vectors is given by (30). There are m coefficients for each digit feature vector and interclass distance (d) is calculated by Ecludian Distance by (31) for 10 digits. This section deals with the results. Table 2 shows that inter-class distance between different digit is more in LDA and LSDA. Seven candidates uttered each digit five times so total 35 feature vectors for each digit.

$$f_i^{(j)} = \frac{1}{35}\sum_{i=1}^{35} x_i^{(j)} \quad (30)$$

$$f^i = [f^1, f^2 \ \dots\dots\dots f^{10}\,]$$

$$D_1 = \sum_{j=1}^{10}\sum_{i=1}^{m}\sqrt{((f_i^{(1)} - f_i^{(j)})^2)} \quad (31)$$

$$T_{int} = \sum_{i=1}^{10} D_i \quad (32)$$

So its recognition result is better as compare to DCT, DWT and PCA. Table 2 shows different transform methods are used to calculate distance coefficients. Table 1 indicate that four is the most recognized digit and its interclass distance is more compare with other digits. DWT with LSDA and LDA have the maximum inter-class distance.

Fig. 3 shows comparison between two classifier namely: Naive Bayes and SVM for all digits. It shows that Naive bayes performance is better for all digit. Fig. 4 shows confusion matrix with 2D-DWT+LSDA features and SVM classifier. It shows that nine has confusion with six though they acoustically very different. Zero has confusion with six. Digit six is found to be more confusable compared to other digits. Tables 2 and 3 indicates that Naive Bayes performance is best than SVM when 3D-DWT or 3D-DCT used along with LSDA. Tables 2 and 3 indicate 3D-DWT and LSDA with Naive Bayes classifier performance is the best.

Table 1 Inter-Class Distance for Different Transform for Different Digits

TRANS /DIGITS	D_0	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	T_{INT}
DWT	0.69	0.8	0.72	0.69	1.00	0.7	0.74	0.69	0.92	0.71	7.72
DCT	0.66	0.71	0.77	0.7	0.78	0.66	0.68	0.64	1	0.64	7.24
DWT+PCA	0.71	0.81	0.76	0.80	1.00	0.91	0.79	0.70	0.85	0.73	8.04
DWT+LDA	0.86	0.83	0.89	0.77	1.00	0.83	0.74	0.78	0.86	0.85	8.40
DWT+LSDA	0.86	0.93	0.89	0.77	1.00	0.83	0.84	0.78	0.86	0.85	8.60

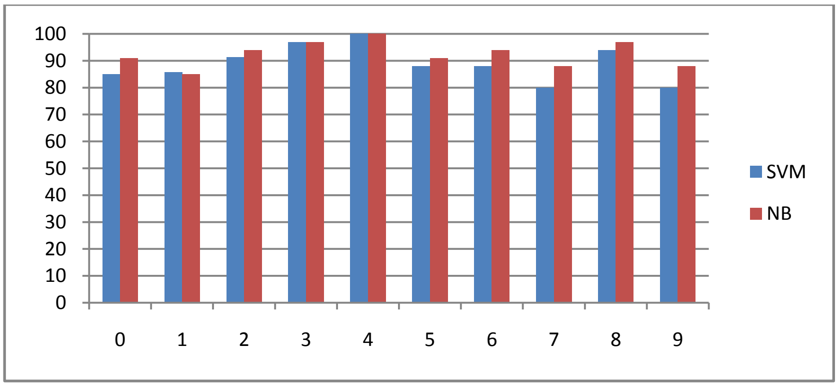

Fig. 3 Classification Rate for different digit for feature vectors DWT+LSDA using SVM and Naive Bayes

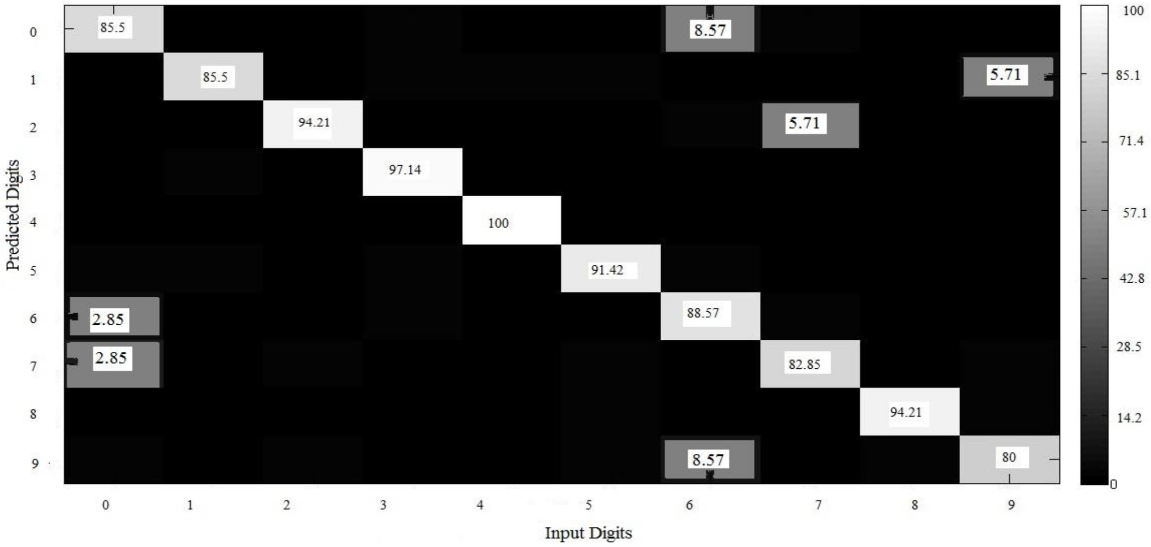

Fig. 4 Confusion Matrix for digit using 2D-DWT with LSDA and SVM classifier

Table 2 RR for with PCA, LDA and LSDA with 2D-DWT and 3D-DWT (CUAVE)

Type of Trans.	2D-DWT +PCA	2D-DWT +LDA	2D-DWT +LSDA	3D-DWT +PCA	3D-DWT +LDA	3D-DWT +LSDA
SVM (%)	70.56	88.33	90.28	74.00	95.71	97.0
NB (%)	60.66	89.14	91.85	60.71	97.16	98.57

Table 3 RR with PCA, LDA and LSDA with 2D-DCT and 3D-DCT (CUAVE)

Type of Trans.	2D-DWT +PCA	2D-DWT +LDA	2D-DWT +LSDA	3D-DCT +PCA	3D-DCT +LDA	3D-DCT +LSDA
SVM(%)	65.66	82.14	84.67	74.00	97.71	98.00
NB (%)	55.5	85.36	87.34	59.00	98.57	98.57

9. CONCLUSION

The performance of LSDA with DCT or DWT is found to outperform other combinations. DWT +LSDA performance is marginally better than DCT +LSDA. The results of LDA are better as compared to PCA with DCT or DWT. Experimental result shows that 2D-DWT+LSDA feature vector size is reduced to 90% of that 2D-DWT feature vector size while resulting in 13% improvement in RR. In case of 3D-DWT+ LSDA improvement in RR is 20% compared to 2D-DWT. The reason behind the improvement in LSDA result is that feature vectors of LSDA are simultaneously better representative of its cluster and discriminative from other clusters while PCA are representative and LDA are discriminative only. Naive Bayes (classifier), performance is found to be better with LSDA and LDA, as compare to the feature vector from other transform techniques. So Naive Bayes is the most appropriate classifier with LSDA. Among the digits, '4' is found as most discriminative and has been always acknowledged. Six is more confused as compared to other numbers. As feature vector length is small, to build the training model SVM and Naive Bayes required less computation time. In future combination of different 3D transforms can be used to get most discriminative feature between them. Further if LSDA technique is used on selected coefficient by using discriminative analysis results may improve.

10. REFERENCES

[1] E. D. Petajan, Automatic lip-reading to enhance speech recognition, Ph.D. Thesis University of Illinois, 1984.

[2] I. Matthews, G. Potamianos, C. Neti and J. Luettin, "A comparison of model and transform-based visual features for audio-visual LVCSR", IEEE International Conference on Multimedia and Expo, 825–828, 2001.

[3] C. Bergler and Y. Konig, ""Eigenlips" For robust speech recognition," in Proc. IEEE Int. Conference on Acustics , Speech and signal processing, 1994.

[4] 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.

[5] G. Potamianos, C. Neti, J. Huang, J. H. Connell, S. Chu, V. Libal, E. Marcheret, N. Haas, J. Jiang, "Towards practical deployment of audio-visual speech recognition", ICASSP-2004.

[6] 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.

[7] X. Wang, Y. Hao, D. Fu, and C. Yuan, "ROI processing for visual features extraction in lip-reading", IEEE Int. Conference Neural Networks & Signal Processing, 178-181, 2008.

[8] N. Puviarasan, S. Palanivel, Lip reading of hearing impaired persons using HMM, Elsevier Journal on Expert Systems with Applications, 1-5, 2010.

[9] A. Shaikh and J. Gubbi, "Lip reading using optical flow and support vector machines", CISP 2010, 327-310 (2010).

[10] G. F. Meyor, J. B. Mulligan and S. M. Wuerger, "Continuous audio-visual using N test decision fusion", Elsevier Journal on Information Fusion, 91-100 (2004).

[11] L. Rothkrantz, J. Wojdel, and P. Wiggers, "Comparison between different feature extraction techniques in lipreading applications", SPECOM- 2006, 25-29 (2006).

[12] M. Heckmann, K. Kroschel, C. Savariaux, and F. Berthommier, "DCT-based video features for audio-visual speech recognition," *7th International Conference on Spoken Language Processing, 1925–1928, 2002.*

[13] P. Viola, M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple features", IEEE Int. Conference, 511-517, 2001.

[14] S. Morade and S. Patnaik, "Lip reading by using 3-D Discrete Wavelet Transform with Dmey wavelet" , IJIP, Vol 8, 385-396, 2014.

[15] M. C. Weeks "Architectures For The 3-D Discrete Wavelet Transform" , Ph.D. Thesis University of Southwestern Louisiana ,1998.

[16] Y. Fan, S. Chen, K. Wu, and J. You, "3D-DCT Chip Design for 3D Multi-view Video Compression", Appl. Math. Inf. Sci. 6 No., 2S, pp. 567S-572S, 2012.

[17] K. Min and M. Fac, "A lip reading method based3D DCT and 3-D HMM" ,IEEE conf. ICIOE, 115-119, 2012.

[18] L.Yaling, Y. Wenjuan, D. Minghui, "Feature Extraction Based on LSDA for lipreading", Proceedings of IEEE International conference, 2010.

[19] H. Jun, Z. Hua, l. Jizhong, "LDA based feature extraction method in DCT domain in lipredaing", computer engineering and application, 45(32), 150-152, 2009.

[20] Deng Cai, X. He, K. Zhou, J.Han, H. Bao, "Loaclity discriminant analysis", International joint conference on artificial Intelligence Hydrabad Morgankauffimann Publishers, 2007.

[21] V. Kechman, "Learning and soft computing, support vector machines, Neural Networks and Fuzzy logic models", MIT Press Cambridge,1-58, 2001

[22] J. C. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines", Microsoft research reports, 1-21,1998.

[23] E. Osuna, R.Freund and F.Girosi, An Improved Training Algorithm for Support Vector Machines, , Neural networks for signal processing , Proc. of IEEE 1997, 276-285, 1997.

[24] T. M. Mitchell," Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression",1-15, 2010.

[25] 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.