Lip Reading by Using 3-D Discrete Wavelet Transform with Dmey Wavelet

Sunil S. Morade *ssm.eltx@gmail.com*
PhD Student,
Electronics Engineering Dept,
SVNIT, Surat.

Suprava Patnaik *suprava_patnaik@yahoo.com*
Professor, Department of E and TC Engineering,
Xavier Institute of Engineering, Mumbai, India.

Abstract

Lip movement is an useful way to communicate with machines and it is extremely helpful in noisy environments. However, the recognition of lip motion is a difficult task since the region of interest (ROI) is nonlinear and noisy. In the proposed lip reading method we have used two stage feature extraction mechanism which is précised, discriminative and computation efficient. The first stage is to convert video frame data into 3 dimension space and the second stage trims down the raw information space by using 3 Dimension Discrete Wavelet Transform (DWT). These features are smaller in size to give rise a novel lip reading system. In addition to the novel feature extraction technique, we have also compared the performance of Back Propagation Neural Network (BPNN) and Support Vector Machine(SVM) classifier. CUAVE database and Tulips database are used for experimentation. Experimental results show that 3-D DWT feature mining is better than 2-D DWT. 3-D DWT with Dmey wavelet results are better than 3-D DWT Db4. Results of experimentation show that 3-D DWT-Dmey along with BNNN classifier outperforms SVM.

Keywords: 2-D DWT, 3-D DWT, Dmey Wavelet, BPNN, SVM, Lip Reading.

1. INTRODUCTION

Lip reading is a technique by which seeing the lip movement one can recognize the speech and is helpful for hearing impaired person. Potential uses of lip reading are communication during disaster like earthquake, noisy factory areas and IVR system. Two fundamental steps involed in lip reading system are: 1) feature extraction and 2) feature classification. Lip features are extracted either by a geometrical model or by an image transform model. Lip geometrical model depends on extraction of lip contour. Inaccuracy in extraction of lip contour affects the different geometrical parameters such as width, height and area. Because of the associated risk of inaccuracy 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 the focus is on image transform model which is also known as appearance 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 with2D-DCT or 2D-DWT as the foremost step of appearance model. Important constraints of the image transform techniques is the feature vector size.

State of art literatures on appearance model are many, out of which few noteworthy literatures are cited here for basic understanding of challenges in lip reading paradigm. E. Petajan [1] experimented on lip-reading to enhance speech recognition by using visual information. The speech reading system proposed by Bregler et al. [3] used Eigen lips as feature vectors. Potamianos et al. [4] compared three linear image transforms namely PCA, DWT and DCT transform techniques. R. Seymour et al. [5] used comparison of image transform features in visual speech recognition of clean and corrupted videos. They evaluated Fast Discrete Curvalet Transform (FDCT), DCT, PCA and Linear Discriminant Analysis (LDA) methods. Wang et al. [6] used different region of interest (ROI) as a visual features in lip reading process and discussed about impact of different ROI processing methods recognition accuracy. N. Puviarasan et al. [7] used DCT and DWT methods for visual feature extraction. They generated data base of hearing impaired persons and observed that DWT with HMM gives better result. A.Shaikh et al. [8] used optical flow information as a feature vector for lip reading. The vocabulary used in their experiment was visemes. Visemes are the basic visual movements associated with phonemes. They tested the result of lip reading using SVM classifier with Gaussian Radial Basis kernel function. The classification performance parameters such as specificity, sensitivity and accuracy are used to test classifiers. Meyor et al. [9] 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. [10] 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. However to our knowledge in none of the publications attempt has been made towards discriminative feature mining from volume information of video, by using 3D transforms.

Selecting predefined number of coefficients from sequence of frames gives feature vectors of defined size for all classes however can't guarantee for efficient feature. Efficient feature is a set of coefficients with interclass variation as maximum as and with the class variation as minimum as possible. While a digit is uttered by *M* people and by each one for N times, the feature vector is required to be more or less similar however for different digits the expectation is to deal with the feature vectors as different as conceivable. 2D transforms would work well for frames with uniform activity and variations. In lip reading framework depending on dynamism or speed of utterance many times the trailing frames seems to have silence and hence non-informative. This can be ruled out by use of 3D transforms.

2. PROPOSED LIP READING FRAMEWORK

A typical lip reading system consists of four major stages: video frame normalization, Face and lip detection, feature extraction, and the finally the 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 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. digits, names etc. Present experimentation is limited to digit utterance.

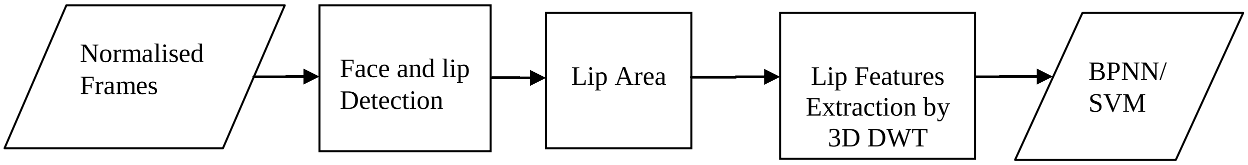


FIGURE 1: Lip reading process.

2.1 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 σ_i, which is defined in (1), is computed for all the frames. These are 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 vector~~s~~ size remains same for each utterance.

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

where, $I_i(x, y)$ stands for the (x, y) spatial location of i^{th} video frame and each frame is of size M*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. 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 [12]. Using Adaboost algorithm for face and mouth detection result is shown in Fig. 2 (a and b).

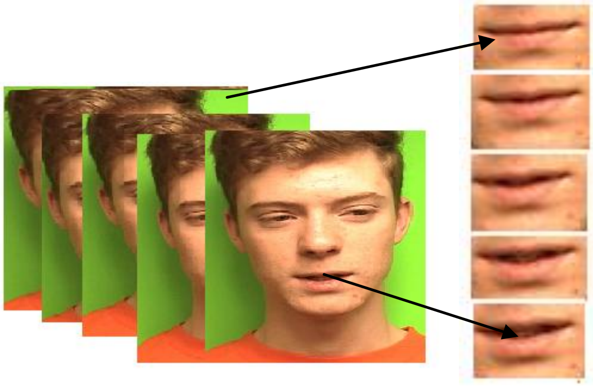


FIGURE 2: (a) Detection of face and lip area for CUAVE database s02m (b) Lip portion.

2.2 3-D Discrete Wavelet Transform (3-D DWT)

1-D transform is used in speech and music, while 2-D is used in image processing. According to Lee, et al.[12], the 3-D DCT is more efficient than the 2-D DCT for image compression application. The DWT has many advantages over the DCT. First, DCT difference coding is computationally expensive. Second, wavelets do not cause blocking artifacts. The 3-D DWT is separable, which means that the 3-D transform is able by applying 1-D DWT in each dimension.DWT considers correlation of images, which translates to better feature vector. Wang and Huang [13] showed the 3-D DWT outperforming the 2-D DWT by 40-90% in image compression. Likewise, one would expect the 3-D DWT outperform the 2-D DWT for lip reading application. Initially for lip reading application we used two dimensional data i.e. Z-axis along the frame was not considered. This axis gives the variation in lip movement so it is important to use 3-D DWT. The results get improved by using 3-D DWT as compared to 2-D DWT.

While 2-D DWT is used for computing the feature vector. Goal is to select only those coefficients which play the dominant role in the representation of lip motion. In standard 2-D 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.The 3-D 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 3-D DWT, the 1D analysis filter bank is applied in turn to each of the three dimensions [2].This is shown in Fig. 3.

DWT computations, the input are multiplied by the shifts (translation in time) and scales (dilations or contractions) of the wavelet. Below are variables commonly used in wavelet architecture. The outputs of low-pass and high-pass filters are given by equation (2) and (3) respectively.

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

$$W_h(n,j) = \sum_{m=0}^{2n} w(m,j-1) * g(2n-m) \qquad (3)$$

where $\mathrm{W(n,j)}$ 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 $\mathrm{w(n,j-1)}$ is the input signal. V. Long and L. Gang [14] proposed a new method for choosing the best wavelet base for speech signals. They have compared Haar, Daubechies, Dmey, Biorthogonal, Coiflets, and Symlet and concluded that Dmey wavelets outperforms for speech signal synthesis. The results from [14] motivated us to select Dmey wavelet, as in lip reading application also the speech information is extracted from visual information.

Fig. 4(a) shows the number of frames of lip in each direction. Fig. 4 (b) shows that first the process transforms the data in the x-direction. Next, the low and high pass outputs both feed to other filter pairs, which transform the data in the y-direction. These four output streams go to four more filter pairs, performing the final transform in the z-direction. The process results in 8 data streams. In our experiment approximation component (LLL) is important so only low pass filter outputs are shown in Fig. 4(b).

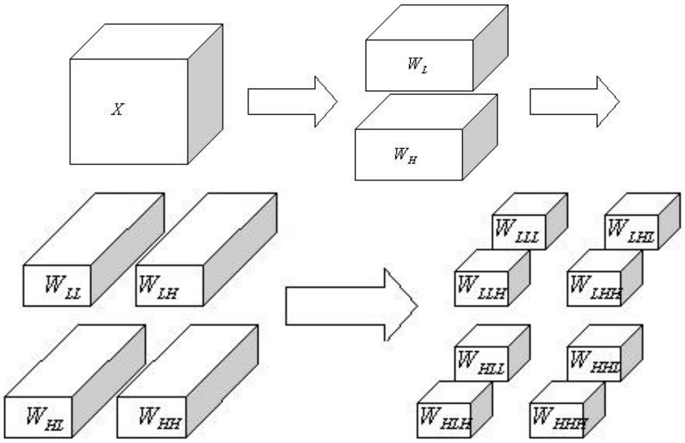
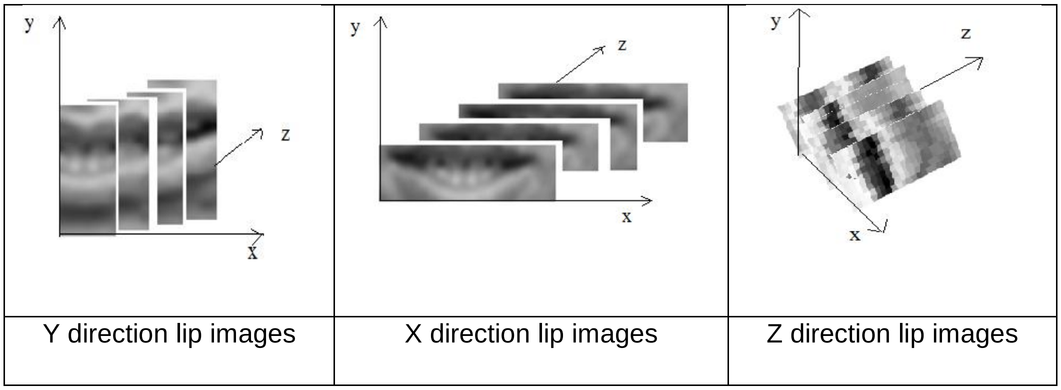
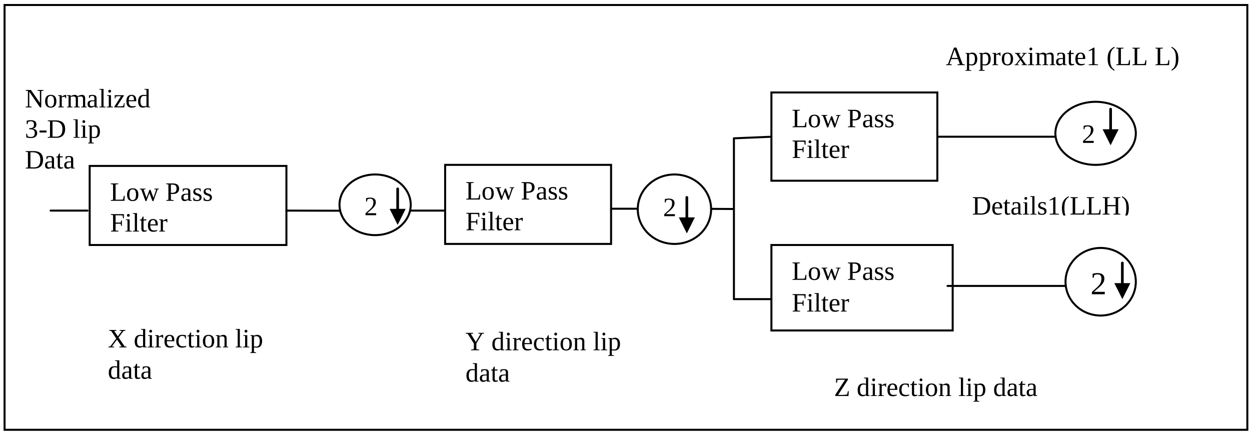


FIGURE 3: The resolution of a 3-D signal is reduced in each dimension.



.

4(a)



4(b)

FIGURE 4(a): Lip images for a digit in three directions (X, Y, Z) **(b)** Single level decomposition for 3-D DWT only for LLL.

3. CLASSIFICATION

The last processing stage of lip reading is feature classification. Researchers have used Naïve Bayes, KNN, ANN, SVM, and HMM as classifier for lip reading. Out of this classifier HMM is mostly used classifier but it is complex and required large training data. Also after doing experimentation it is observed that the performance of Naïve Bayes and KNN is poor as compared to SVM and ANN for lip reading application. A.K. Jain et al. [15] in their review paper of pattern recognition compared different classifier and found that ANN and SVM are most suitable for lip reading application. So in this paper experimentation results for SVM and ANN are compared.

3.1 BPNN

Artificial Neural Network (ANN) models are useful to solve pattern recognition problems due to their low dependency on domain-specific knowledge and due to the availability of efficient learning algorithms. It is important to note that neural networks themselves can lead to in any different classifiers depending on how they are trained. Layers in multilayer perceptron network architecture allow to identify nonlinear decision boundaries.

ANN is mathematical model consists of number of highly interconnected processing elements organized into layers, geometry and functionality of which have been resembled to that human brain. Bregler and Y. Konig [16] in their experiment used a TDNN for speech and visual speech data. The ANN may be regarded as possessing learning capabilities in as much as it has natural. The first layer consists of input element in accordance with feature vectors. The most frequently used training algorithm in classification problem is the back propagation algorithm. The neural network has been trained to adjust the connection weights and biases in order to produce desired mapping.

We have used a BPNN classifier. The nodes of this network are sigmoid. Number of hidden layers tried varies in the range 10 to 30. Best performance of the neural network for feature vector of DWT is obtained with 20 hidden layers. Learning rate and moment coefficients are set at 0.3 and 0.001 respectively. Input vector X is a feature vector of size n which is the length of feature vector. The ten neurons in the output layer were used to represent the digit.

3.2 Support Vector Machines (SVM)

One of the most interesting and recent developments in classifier design paradigm is the introduction of support vector machine classifier by Vipnik [17]. It is a two class classifier. The optimization criteria is the width of the margin between the classes i.e. empty area around the decision boundary defined by the distance to the nearest training patterns .These patterns are called support vectors and finally used to define the classification function.

The important advantage of the support vector classifiers is that it offers possibility to train generalizable, nonlinear classifiers in high dimensional spaces using small training set. The support vectors replace the prototypes with the main difference being that they characterize the classes by decision boundary. Moreover this decision boundary is not defined by minimum distance function but by a more nonlinear combination of these distances [15]**.**

Data separation is completely possible by using nonlinear separation but it is not using linear separation. 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 are expressed in terms of Φ by (9). 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.

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

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

$$k(\boldsymbol{x_i}, \boldsymbol{l_j}) = \boldsymbol{\Phi^T}(x_i)\boldsymbol{\Phi}(l_j) \qquad (4)$$

Kechman in his literature discussed the mapping and different kernel function [18]. SVM maximizes the distance of separating plane from the closest training data point. Linear kernel is defined by (5a) while polynomial kernel is defined by (5b) where d is the degree of polynomial.

$$k(\boldsymbol{x}, \boldsymbol{l}_i) = (\boldsymbol{x^T}\boldsymbol{l}_i) \qquad (5a)$$

$$k(\boldsymbol{x}, \boldsymbol{l}_i) = (\boldsymbol{x^T}\boldsymbol{l}_i + 1)^d \qquad (5b)$$

For classification, the decision function for a two class problem derived by a support vector can be written by (6) using a kernel function $k(x, l_i)$ of a new pattern x and a training pattern l_i.

$$f(x) = \sum_{i=1}^{N} y_i \alpha_i \, k(x, l_i) + b \quad (6)$$

Where k kernel function, b scalar bias, α langrage's multiplier and $y_i = \pm 1$ is the label of object x_i and l_i support vector obtained from training data. In equation (7) we need to find suitable Lagrange multipliers α to get the following function reach its maximum value.

$$L_d(\alpha) = \sum_1^l \alpha_i - \frac{1}{2}\sum_1^l y_i \alpha_i \alpha_j {\Phi_i}^T \Phi_j \qquad (7)$$

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

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 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 [19]. E. Osuna et al. has suggested two modifications in Platt's SMO algorithm so that the SMO algorithm speeds up to train SVM in many situations [20].

Large SVM training data can fit inside of the memory of an ordinary personal computer. 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 approximately thousand times faster for linear SVMs and ten times faster for non-linear SVMs [20]. Because of its ease of use and better scaling with training set size, SMO is the standard SVM training algorithm. In this experiment SMO is used for training SVM with 2nd degree polynomial kernel function.

4. PROPOSED LIP READING METHODOLOGY

A Flow chart of the steps involved in our simulation technique is shown in Fig. 5. Three major execution steps of the algorithm are: 1) pre-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 applying 3-D DWT, low frequency components (LLL) of the image are taken as a feature vector for classification. 3-D DWT or 2-D DWT attempts to transform image pixels of significant lip frames into a new space which separates redundant information and provides better discrimination. Then the final feature vectors of all the train lip frames are stored in the training database along with class level.

5. CORPUS AND RESULT

5.1 CUAVE Database

CUAVE [21] (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 booth at a resolution of 720 x 480 with the NTSC standard of 29.97 fps using 1 Megapixel-CCD camera. This database is a speaker-independent database consisting of connected and continuous digits spoken in different situations. The database consists of two major sections: one of speaker pairs and the other one of individuals.

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. Disruptive mistakes are removed, but occasional vocalized pauses and mistakes in speech are kept for realistic test purposes. The data was then compressed into individual MPEG-2 files for each individual speaker and group of two speakers. It has been shown that this does not significantly affect the collection of visual features for lip reading. The object of the video captured for the presence of two speakers speaking simultaneously does not affect significant features for lip reading.

Each individual speaker was asked to move side-to-side, back-and-forth, or tilt the head while speaking 30 isolated digits. In addition to this isolated section, there is also a connected-digit section with movement as well. So far, much research has been limited to low resolution, pre-segmented video of only the lip region.

5.2 TULIPS Database

Tulips1.0 is a small Audiovisual database of 12 subjects saying the first 4 digits in English. Subjects are undergraduate students from the Cognitive Science Program at UCSD. The database was compiled at R. Movellan's laboratory at the Department of Cognitive Science, UCSD.

Figure 6 shows the 6 frames for utterance of digit 0 using TULIPS database. The "Raw Data" directory contains two directories: Tulips1.A and Tulips1.V. The "Preprocessed Data" directory contains representations used by different researchers on this database. Tulips1.V contains the video files in 100 x 75 pixel 8 bit gray level, .pgm format. Each frame corresponds to 1/30 of a second. R. Movellan presents work on speaker independent visual speech recognition system and used simple HMM as a classifier used Tulips database of 1 to 4 digits for testing result [22].

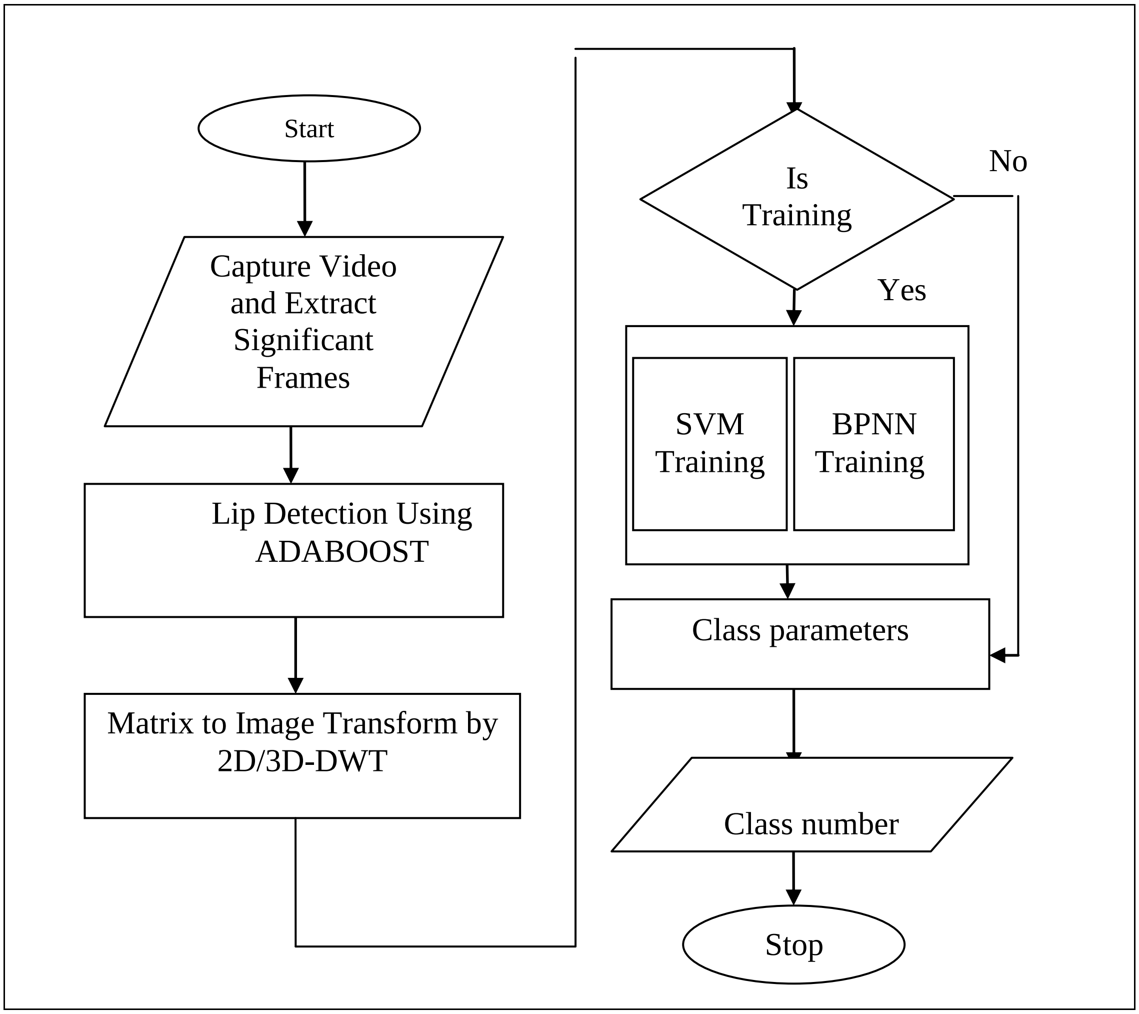

Start
Capture Video and Extract Significant Frames
Lip Detection Using ADABOOST
Matrix to Image Transform by 2D/3D-DWT
Is Training
No
Yes
SVM Training
BPNN Training
Class parameters
Class number
Stop

FIGURE 5: Flowchart for lip reading system.

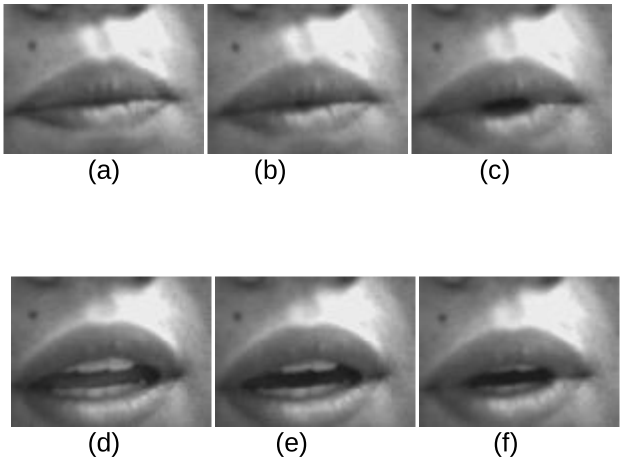


FIGURE 6 (a): to (f) Number of frames of zero utterance using TULIPS database.

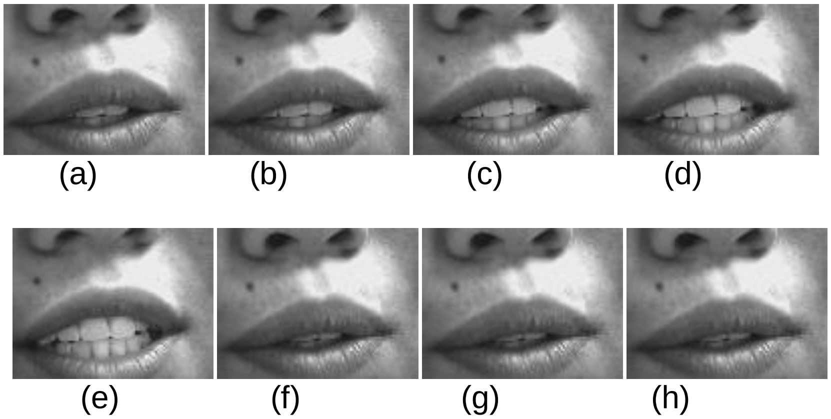


FIGURE 7: (a) to (h) Number of frames for utterance of three using TULIPS database.

6. 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 22 x 33 and passed through an LPF to remove high frequency noise. In proposed experimentation 3-D DWT with three decomposition levels are applied to lip area. DWT with Db2, Db4 and Dmey wavelets are used and respective coefficients are generated per frame. DWT Coefficients are calculated for 10 normalized frames in CUAVE database. This results in a feature vector of size 162 x 1. Seven users and each one uttering each digit 5 times produces 350 x 162 dimensional training dataset Feature vectors are levelled with 10 different classes each class corresponding to a digit. For 10 frames 162 coefficients are generated after 3D DWT as compared to 300 required for 2-D DWT.

7. EXPERIMENTAL RESULTS

7.1 Results for CUAVE Database

This section deals with the results. Table 1 shows recognition rate of lip reading for individual candidate with DWT and BPNN classifier. It shows that as the Recognition Rate (R.R.) of persons is more they provide more visual speech information.Table 2 indicate that Confusion Matrix for 2-D DWT with Dmey wavelet with BPNN classifier with M=0.001. Table 3 shows that Confusion Matrix for 3-D DWT with Dmey wavelet with BPNN classifier with M=0.001.Table 4 shows that recognition results of 3-D DWT are better than 2-D DWT for Db2 and Dmey wavelets. DWT with Db4 wavelet gives better result than Db2. DWT with Dmey wavelet shows highest recognition rate with BPNN classifier. BPNN classifier outperforms SVM classifier.Table 5 shows that the performance improvement of each digit for 3-D DWT compared to 2-D DWT with BPNN and SVM classifier. Average % improvement is more in BPNN as compared to SVM. For 3-D DWT 0 digit has greater performance improvement.

TABLE 1: Lip reading results using 2-D DWT and BPNN classifier for individual candidate.

Candidate Number	Reco . Rate(%)
1	94
2	81
3	81
4	72
5	84
6	76
7	72

TABLE 2: Confusion Matrix for 2-D DWT with Dmey wavelet with BPNN classifier with M=0.001.

Digits	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**% R.R.(BPNN)**
0	18	0	3	0	1	1	7	3	1	1	51.42
1	0	25	2	2	2	1	2	0	1	0	71.42
2	2	0	32	0	0	0	0	1	0	0	91.42
3	0	1	0	28	1	1	4	0	0	0	80
4	0	2	0	0	32	1	0	0	0	0	91.42
5	0	2	0	3	0	26	0	1	0	3	74.32
6	1	0	2	2	0	0	26	2	1	1	74.32
7	2	0	0	0	0	0	7	19	1		54.28
8	2	1	0	0	0	0	1	0	26	5	74.28
9	0	0	0	0	0	0	2	2	6	25	71.42
Average Result											**73.43**

TABLE 3: Confusion Matrix for 3-D DWT with Dmey wavelet with BPNN classifier with M=0.001.

Digits	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**% R.R.(BPNN)**
0	24	0	3	1	1	0	3	2	1	0	68
1	1	26	0	2	2	1	1	1	1	0	74.28
2	1	0	34	0	0	0	0	0	0	0	97.1
3	0	0	1	26	1	2	4	1	0	0	74.28
4	0	0	1	0	33	1	0	0	0	0	94.28
5	0	1	0	0	1	30	0	2	0	1	85.71
6	2	2	0	2	0	0	29	0	0	0	82.85
7	1	0	0	1	0	0	3	27	0	3	77.14
8	0	0	0	0	0	1	0	0	31	3	88.57
9	0	0	0	0	0	0	0	1	5	29	82.85
Average Result											82.50

TABLE 4: Reco. Result (R.R.) for 3-D DWT using BPNN and SVM.

Type of Transformation	SVM R.R.%	BPNN R.R, %
2-D DWT (Db2)	67.14	71.14
2-D DWT (Db4)	55.14	59.71
2-D DWT (Dmey)	70.85	73.71
3-D DWT (Db2)	73	78
3-D DWT (Db4)	75.23	80
3-D DWT (Dmey)	78.56	82.50

TABLE 5: % improvement in R.R. for feature vector using 2-D DWT and 3-D DWT for BPNN and SVM.

	BPNN	SVM
Digits	**%R.R. Improvement**	**%R.R. Improvement**
0	16.58	17.15
1	2.86	2.85
2	5.68	0
3	-5.72	8.58
4	2.86	8.58
5	11.39	20
6	8.53	0
7	22.86	8.57
8	14.29	8.54
9	11.43	2.86
Avg	9.07	7.7

7.2 Results for TULIPS Database

SVM and BPNN are trained for feature classification. Table 6 shows the confusion matrix for feature vector using 2-D DWT with BPNN classifier with TULIPS database. From confusion matrix 1 and 4 are found to be most recognized digits and 3 is less recognized. From Table 7, 3-D DWT with Dmey wavelet performance found to be better with, as compare to 2-D DWT. Tulips database results for digit 3 are less as compare to CUAVE database because orientation of lip is not proper and nose portion also appear in lip image as shown in Fig. 7.

TABLE 6: Confusion matrix for 4 digits using 2-D DWT Dmey wavelet feature vectors with BPNN.

		Digit Presented				**Recognition Rate in %**
		1	**2**	**3**	**4**	
Subject Response	1	21	0	2	1	87.5
	2	0	18	4	2	75
	3	3	2	16	3	66.7
	4	1	1	1	21	87.5
Average Result						79.2

TABLE 7: Confusion matrix for 4 digits using 3-D DWT Dmey with BPNN.

		Digit Presented				Recognition Rate in %
		1	2	3	4	
Subject Response	1	22	0	2	1	91.67
	2	0	18	4	2	75
	3	3	2	16	3	66.7
	4	1	1	1	22	91.67
Average Result						81.25

8. CONCLUSION

In this paper, we have compared 2-D DWT with 3-D DWT features for lip reading. Naturally 3-D DWT performance is better as compare to 2-D DWT. SVM and BPNN are trained for feature classification. BPNN (classifier), performance found to be better with 3-D DWT with Dmey wavelet, as compare to the feature vector from other transform techniques. So BPNN is the most appropriate classifier with 3-D DWT. Among the digits, '4' is found as most discriminative and has been always acknowledged. '0' has less recognition rate as compared to other numbers. As using 3-D DWT length of feature vector is small, to build the training model SVM and BPNN required less computation time. Performance of lip reading system using both 2-D and 3-D DWT is less for Tulips database because of lip orientation. Further experimentation may reduce the 3-D DWT feature vector size by using proper discrimination technique and the performance of lip reading can be improve for real time application.

9. REFERENCES

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

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

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

[4] Potamianos, H. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lip reading," Int. Conf. on Image Processing, 173–177, 1998.

[5] 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.

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

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

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

[9] 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.

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

[11] P. Viola, M. Jones, “Rapid Object Detection using a Boosted Cascade of Simple features”, IEEE Int. Conf., 511-517, 2001.

[12] H. Lee, Y. Kim, A. Rowberg, and E. Riskin, "Statistical Distributions of DCT Coefficients and their Application to an Inter frame Compression Algorithm for 3-D Medical Images," IEEE Transactions of Medical Imaging, Vol. 12, 478-485, 1993.

[13] J. Wang and H. Huang, "Three-dimensional Medical Image Compression using a Wavelet Transform with Parallel Computing," SPIE Imaging Physics Vol. 2431, 16-26,1995,

[14] V. Long and L. Gang “Selection of the best wavelet base for speech signal” IEEE. Intelligent multimedia, video and speech processing, 2004.

[15] A. K. Jain, R. P. Duin, and J. Mao, “Statistical Pattern Recognition: A Review” IEEE Transactions On Pattern Analysis And Machine Intelligence, 22, 1, 2000.

[16] C. Bregler and Y. Konig, “Eigenlips” For Robust Speech Recognition”, IEEE conf. Acoustics, Speech, and Signal Processing, 1-4, 1994.

[17] V.N. Vapnik, “stastical learning theory” New York John Wiley & Suns, 1998.

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

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

[20] 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

[21] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, “CUAVE: a new audio-visual database for multimodal human computer- interface research”, Proceedings of IEEE Int. conf. on Acoustics, speech and Signal Processing, 2017-2020, 2002.

[22] J. R. Movellan “Visual Speech Recognition with Stochastic Networks”, Advances in Neural Information Processing Systems, MIT Pess, Cambridge, 1995.