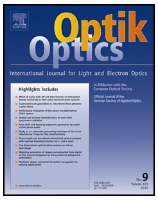
Optik 126 (2015) 5753–5761

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo



Comparison of classifiers for lip reading with CUAVE and TULIPS database

Sunil S. Morade[a,∗,1], Suprava Patnaik[b,2]

[a] *Department of Electronics Engineering, SVNIT, Surat, India*
[b] *Department of E and TC Engineering, Xavier Institute of Engineering, Mumbai, India*

ARTICLE INFO

Article history:
Received 29 August 2014
Accepted 25 August 2015

Keywords:
Back propagation Neural Network (BPNN)
Discrete Cosine Transform (DCT)
Discrete Wavelet Transform (DWT)
Lip reading
Support Vector Machine (SVM)

ABSTRACT

Automatic lip reading is a technique of understanding the uttered speech by visually interpreting the lip movement of the speaker. The two major parts, which play crucial role in lip reading system, are feature extraction followed by the classifier. For automatic lip reading, there are many competing methods published by researchers for feature extraction and classifiers. In this paper, we compare some of these leading methods. We have compared Support Vector Machine (SVM), Back Propagation Neural Network (BPNN), K-Nearest Neighborhood (KNN), Random Forest Method (RFM) and Naive Bayes (NB) classifiers, on the basis of recognition performance and training time. Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) are studied to extract feature vectors. The CUAVE and Tulips database are used for experimentation and comparison. It is observed that SVM outperforms the rest for CUAVE database. Training time of SVM is also less than others.

1. Introduction

Automatic lip reading is an active research area. It acts as a supplement to acoustic speech recognition technology for noisy environment. It has many standalone applications like multimedia phones for hearing impaired, person identification in video surveillance systems, remote machine control in industrial environments, etc. Unlike to acoustic speech recognition, the major challenge for Visual Speech Recognition (VSR) is the lack of standards. Lip movements are neither learned through grammar of phonemes or pronunciation dictionary nor have well defined statistical models like mel-frequency, ceptral coefficient, etc. People do not concern about producing sufficient visual signal while talking, but actually visual speech provides information which may not be present in the audio signal. Acoustic automatic speech recognition systems tend to perform poorly in noisy environment. Of course, it is important that the visual modality provides information which is more stable and robust, and so there have been numerous studies going on to assess and improve the performance of visual speech recognition.

Visual features can broadly be classified into two categories. First category is shape based features, which uses geometrical parameters such as height, width and area of lip contour. Geometry of mouth is captured efficiently however fails to capture many significant features such as cavity information. A major drawback of geometrical features is these are sensible to the accuracy of extracted lip contour. Second category includes image transformation based feature extraction methods such as DCT, DWT and Principal Component Analysis (PCA). Image transform methods capture cavity information along with the geometrical parameters, hence are efficient than geometrical feature vectors.

Discrete Cosine Transform (DCT) transforms image from the spatial domain to the frequency domain. For lip reading major amount of information remain in DC coefficient and lower frequency AC coefficient, therefore selected DCT coefficients are used as feature vector. Each level of DWT decomposes and splits the input signal into a low and high frequency sub-bands. In image applications one level of DWT results four sub-band images each one vector, one fourth size of the original image. Multiple level of decomposition followed by feature vector extraction significantly reduces the feature vector size. For PCA, pixel information is converted into eigenspace and significant eigenvectors are selected, as feature vectors.

Efficient classification of feature vectors is useful as final result after feature extraction depends on classifier. For lip reading KNN, Neural network, SVM and HMM are used as classifiers. Many researchers have used HMM classifier; however, it involves more

[∗] Corresponding author. Tel.: +91 253 2531692.
E-mail addresses: ssm.eltx@gmail.com, ssmorade@kkwagh.edu.in (S.S. Morade), suprava_patnaik@yahoo.com (S. Patnaik).
[1] Department of E and Tc Engineering, K.K. Wagh Institute of Engineering and Research, Nashik, India
[2] (Ex-Professor, Department of Electronics Engineering, SVNIT, Surat, India)

http://dx.doi.org/10.1016/j.ijleo.2015.08.192
0030-4026/© 2015 Elsevier GmbH. All rights reserved.

computation and large training data, so we proposed the comparison among other classifiers.

2. Related work

To our knowledge of literature survey, first geometrical based feature system was developed by Petajan et al. [1]. Also they used image transform based approach, which obtains a compressed representation of the image pixel values that contain speaker's lip area. Authors have experimented features on both approach and the results are compared for visual recognition. It is shown that the image transform based approach results in superior lip reading performance.

The speech reading system proposed by Bregler et al. [2] used eigenlips as feature vectors. They proposed Time delay Neural Network (TDNN) as classifier, which consists of an input layer, a hidden layer of 256 neurons, phone state layer and an output layer of 63 neurons. Back-propagation learning was used to train the neural network. Authors have described another connectionist approach for combining acoustic and visual information by using a hybrid MLP–HMM speech recognition system.

Potamianos et al. [3] have compared Principal Component Analysis (PCA), DWT and DCT transform techniques for digit recognition. They found that PCA requires intensive training phase and is not amenable to fast implementation; the use of DWT or DCT is more recommended. Authors concluded that image transform features are robust to video degradation.

Matthews et al. [4] compared different transform techniques for Large Vocabulary Continuous Speech Recognition (LVSCR) and found that word error rate is more for LVSCR. They also compared Active Appearance Model (AAM) with image transform model. Image transform model is better compared to AAM.

Heckman et al. [5] investigated on selection of the DCT coefficients and the influence on the recognition scores. They used a hybrid HMM audio-visual speech recognition system. DCT is applied to a rectangular block of sub image bounding the lip area. This approach does not require lip contour extraction. Due to simplicity and stability of DCT features are superior to geometrical features. They observed that energy of DCT coefficient has direct relation with the performance. They concluded that 30 DCT coefficients are sufficient for recognition of digits. This is also getting support by the fact that DCT coefficients preserve cues related to appearance of teeth, tongue, and shape of muscles around the lips. Matthews et al. [6] used three methods for parameterization of lip. Two methods used inner and outer lip contours for lip modeling and derived lip reading features from a PCA and third method used a non-linear scale space analysis to form features from the pixel intensity.

Meyor et al. [7] 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. After DCT transform, an increase in the DC coefficient between consecutive frames correspond to increase in mean energy and is likely due to appearance of teeth. On other side a DC coefficient represent large mouth opening leading to smaller mean value. They concluded that best set of features having 8 coefficients in vertical direction and 8 coefficients in horizontal direction of low frequency and dynamic DCT coefficients for each direction.

Rothkrantz et al. [8] 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 the LGE method competitive with some strong points on its favor. Wang et al. [9] used active shape model using HMM and RDA classifier for digit recognition.

Seymour et al. [10] 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. They used 10 states HMM classifier for experiment. Wang et al. [11] used different regions of interest (ROI) as a visual feature in lip reading process. Authors discussed about different ROI processing methods expressed its impact on recognition accuracy. Four ROIs are obtained by using gray scale normalization, different enhancements, edge enhancement and image segmentation. The experimental results show that DCT based features with normalized gray scale image can achieve the best recognition performance among the other processed ROIs.

Puviarasan et al. [12] used DCT and DWT for visual feature extraction. They used data base of hearing impaired person and observed that DWT with HMM gives better result. Shaikh et al. [13] 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. The classification performance parameters such as specificity, sensitivity and accuracy are used to test classifiers.

Zhao et al. [14] calculated spatiotemporal local texture features directly from image sequences and trained SVMs upon the features for classification. In this paper SVM classifier was selected since it is well founded in statistical learning theory and has been successfully applied to various object detection tasks in computer vision. The author used the second degree polynomial kernel Function, which provide the best results after the comparison of linear, polynomial, and RBF kernels in experiments. Pei et al. [15] used an efficient lip reading approach by using the unsupervised random forest manifold alignment (RFMA). The density random forest is employed to estimate affinity of patch trajectories in speaking special videos.

Zhou et al. [16] proposed a generative latent variable model to provide a compact representation of visual speech data. The model uses latent variables to separately represent the interspeaker variations of visual appearances and those caused by uttering within images. They used a simple classification model than HMM or Dynamic time warping.

3. Lip feature extraction

3.1. Lip reading process

The basic steps of lip reading process can be described by the block diagram given in Fig. 1. Always the first step is to separate out important frames from the captured video for the utterance of digits, second is face and lip detection and third is lip area separation.

3.2. Video data separation

Pratt software is used for audio analysis of captured video. From audio analysis, the time duration of each digit is calculated, which is used to separate important video frames of a digit. Frames are captured before 0.2 s from starting of each digit. Fig. 2 indicates audio waveform for digit 0 – 9 used for separation of video frames.

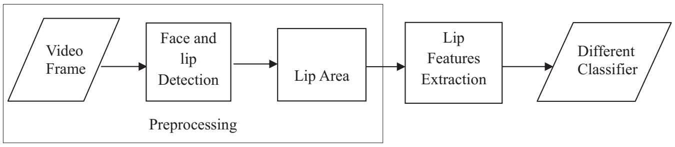


Fig. 1. Lip reading using different classifiers.

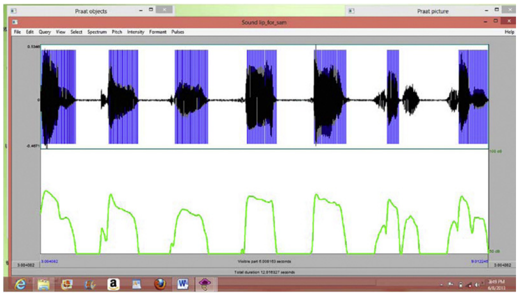

Praat objects
Praat picture
Sound lip_for_sam
File Edit Query View Select Spectrum Pitch Intensity Formant Pulses
Help
0.5346
0
-0.4671
100 dB
50 dB
3.004082
9.012245
Visible part 6.008163 seconds
Total duration 12.016327 seconds
3:49 PM
6/8/2013

Fig. 2. Audio waveform for digit 0 – 9 for separation video frames.

In our experimentation, we have used gray images. There are large inter and intra subject variations due to the variation in speed of utterance and this results difference in the number of frames for each utterances. Approximately 16 – 25 frames are required to utter a digit. We have used Mean squared difference (σ_i) between consecutive frames to get active and prominent frames by using Eq. (1). Ten frames with the higher values of σ_i are selected. The number of frames for each utterance is made same in order to extract feature vectors of same size. Complexity of computation is reduced by using significant frames.

$$\sigma_i = \frac{1}{M \times N} \sum_{x=0}^{x=M-1} \sum_{y=0}^{y=N-1} [I_i(x, y) - I_{i+2}(x, y)]^2. \tag{1}$$

where, I_i is ith frame, I_{i+2} is $i+2$nd frame of size $M \times N$ pixels. σ_i is the mean squared difference between the two frames. i varies from 1 to number of frames required to utter a digit.

3.3. Face and lip detection

Lip detection or segmentation is not trivial due to the low contrast 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 speaker's clothing can cause segmentation failures.

For lip area detection and tracking there are many algorithms available which have their own strengths and weaknesses. Some of them use contours, some use templates and others used filters. These algorithms are computationally expensive. Vio and Jones invented an algorithm based on Adaboost classifier to rapidly detect any object including human face [17]. They presented a face detector which uses a holistic approach and is much faster than any contemporaries. Adaboost classifier which cascades the Haar like features and not pixels features, results in rapid detection of human faces. Therefore, we have selected Viola and Jones algorithm for face and lip detection. Result of face and mouth detection using Adaboost algorithm is shown in Fig. 3. Lip portion is identified after cropping the image accordingly.

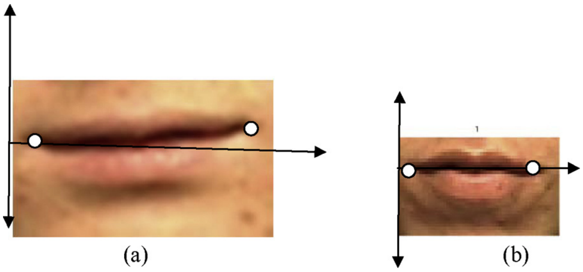


Fig. 4. Cropped lip images (a) before and (b) after normalization.

Visual features will usually be extracted from the video frames. A major problem in generating visual features is the enormous quantity of data in video sequences. Each video frame contains more than thousand pixels, from which feature vector of size between 10 and 100 elements must be selected. Ideally, these features vector should be robust to variables as different talkers, head poses and light conditions. One way of approach which we have followed is to obtain features directly from image and another approach consists of parameter values encapsulated into a geometric model. Image transform method attempts to transform image pixels of video frames into new compact representation space which removes redundant information and provides better class of discrimination. Fig. 4 shows cropped lip portion of size 44 × 65 pixels. Before performing the transformation lip image is pre-processed for orientation, intensity and size normalization for down sampled to size 22 × 32 pixels. Normalization is done all the way through pixel intensities divided by maximum value of pixels for 10 frames.

3.4. Image transform based features mining

This section aims to compare the strength of transform domain feature to geometric features. The area of lip is transformed by using either DCT or DWT method. DCT and DWT are very useful for dimensionality reduction—usually a small number of low frequency coefficients can approximate well most time series images.

3.4.1. DCT of lip area

The two dimensional DCT of matrix A of size $M \times N$ is defined by Eq. (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 × 32 DCT coefficients. To select DCT coefficients, upper triangle mask is preferred over rectangular mask because it gives lower frequency component information.

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

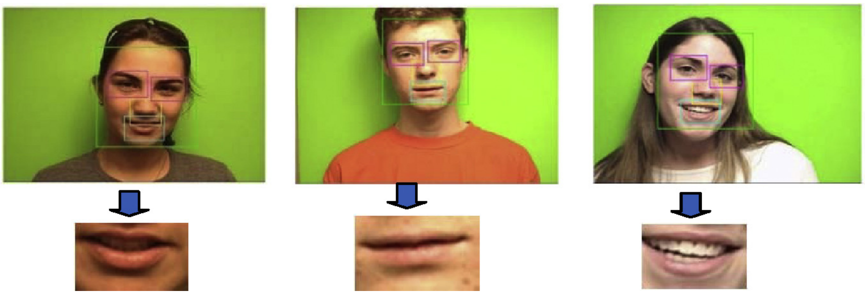


Fig. 3. Result of detection of face and lip bounding box for samples from CUAVE database.

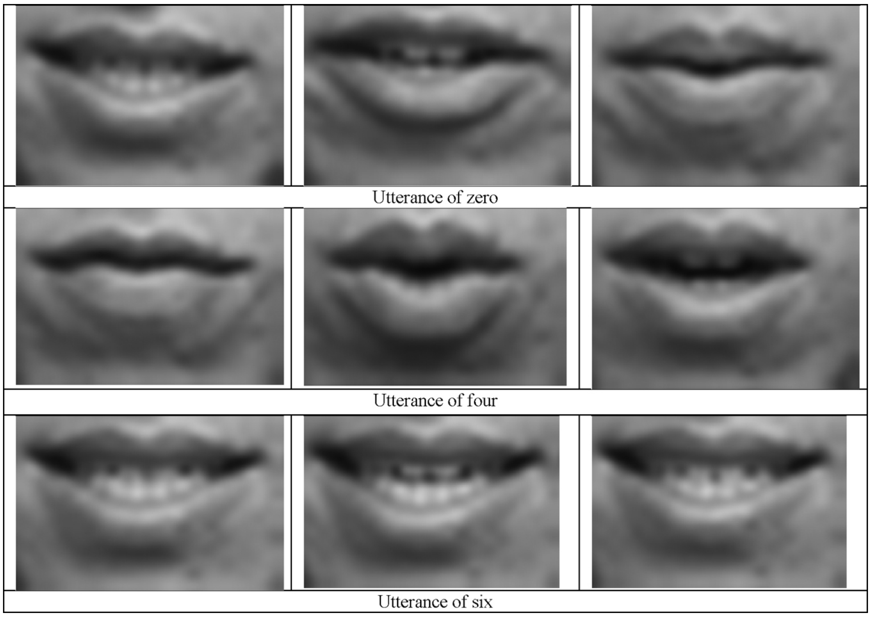


Fig. 5. Frames 1, 5 and 10 for utterance of digits 0, 4 and 6 using CUAVE database.

$$\alpha_p = \left\{ \begin{array}{cc} \dfrac{1}{\sqrt{M}}, & p = 0 \\ \sqrt{2/m}, & 1 \le p \le M - 1 \end{array} \right\}.$$

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

This section aims to compare the strength of transform domain feature. 0th, 5th and 10th frame obtained after the normalization for utterance of digit 0, 4 and 6 are shown in Fig. 5. Comparison of DCT features with DWT features is given in Table 1. All the parameters are expressed in normalized form. Normalization is done using the parameter is divided by maximum value of parameter for 10 frames. The first coefficient of the DCT is F_{00} which is DC (zero frequency) component. There is a conflict among researchers about inclusion of DC coefficient. We strongly recommended the DC component, as DC coefficient provides information related to percentage visibility of teeth and tongue. Comparing the utterance of zero to nine it is trivial to feel that lip variation is significant for digits 0 and 4 another side of variation it is minimum for digit 6. Frequency parameters are more suitable for digit recognition as they exhibit more variations and hence suitable for discrimination as shown in Table 1.

3.4.2. Discrete Wavelet Transforms (DWT) of lip portion

Temporal data have a unique structure—high dimensionality and high feature correlation. DWT maps the signal into a joint time—frequency domain. DWT hierarchically decomposes the signal using windows of different sizes to offer multi resolution analysis. Good time resolution and poor frequency resolution at high frequencies and good frequency resolution and poor time resolution at low frequencies. DCT and DWT bear comparable in energy preservation, but DWT is faster to calculate and offers a multi-resolution decomposition or time-frequency localization. DCT measures global frequencies and the signal is assumed to be periodic. The latter assumption may cause poor approximation at the border of a time series. The basis of the DWT consists of infinitely many scaled and shifted versions of a mother wavelet function, often with compact support. In discrete implementation of DWT, working with a dyadic grid of resolution, the wavelet basic functions are obtained as

$$\Psi_{j,k}\Psi(n) = 2^{\frac{j}{2}}\Psi\left(2^j n - k\right). \tag{3}$$

where, Ψ is the mother wavelet function. For any square inferable real function $x(n)$ one stage of decomposition results in

$$\mathrm{DWT}_{(x(n))} = \left\{ \begin{array}{l} d_{j,k} = \left\langle \Psi_{j,k}(n), x(n) \right\rangle \\ a_{j,k} = \left\langle \Phi_{j,k}(n), x(n) \right\rangle \end{array} \right\}. \tag{4}$$

where, $\Phi(x)$ is known as scaling function and in multiresolution analysis as an example of Haar function, the scaling and wavelet functions are related as given in Eq. (5).

$$\Phi(x) = \Phi(2x) + \Phi(2x - 1). \tag{5}$$

$$\Psi(x) = \Phi(2x) - \Phi(2x - 1).$$

The coefficients $d_{j,k}$ refer to the detail components in signal $x(n)$ and correspond to the wavelet function, whereas, $a_{j,k}$ coefficients refer to the approximation component of the input signal. More detail about wavelet is not within the scope of this paper and is available in many standard literatures. In our lip reading experimentation we have used DAUB4 (db4) filters. These filter coefficients extend in a

Table 1
Variance of DCT and DWT coefficients for digits 0 – 9.

Digits	0	1	2	3	4	5	6	7	8	9	$\Delta\sigma_{\text{avg}}$	$\Delta\sigma_{\text{min}}$
DCT σ	0.8	0.77	0.86	0.92	0.88	0.57	0.54	0.71	1	0.83	0.79	0.03
DWT σ	0.72	0.96	0.82	0.63	1	0.68	0.44	0.56	0.37	0.5	0.67	0.04

way to generate up to $N = 4$ moments to be equal to zero and hence better approximation. The two well established feature extraction techniques using DWT for time series classification applications are: (1) Saving only the first k coefficients as they preserve a rough sketch of the time series and (2) To use the k largest normalized coefficients as they preserve the optimal amount of energy. The later choice is based on the fact that better energy preservation guarantee that a prototype reconstruction, using the extracted features, will be more similar to the original input pattern with respect to the Euidean distance. Thus, it also will preserve the Euclidean distance between the input time sees intuitively. Keeping the largest instead of approximation coefficients in general causes memory problem when dealing with several time series or videos. It involves higher storage for location information and distance computations and not suitable in particular for lip reading applications which involves time series indexing. Further in feature mining extracting knowledge that is interpretably a domain expert is the ultimate goal of knowledge discovery. Trivial clustering algorithms rely on a meaningful distance function to group data vectors that are closed to each other and distinguish themselves from others that are far away. But lip reading being samples from enormous high dimensional space, the contrast between the nearest and the farthest neighbor gets increasingly smaller, making it impossible to find meaningful groups. Therefore in the present experiment we have considered only the approximation sub-band coefficients obtained after 3-level of decomposition using db4 wavelet filters.

In geometric or shape based approach rules are based on only a few time points that decide the height, width, area, etc. Subsequently, the generated rules are highly questionable, particularly when applied on semantically uncontrolled applications. More so, discrimination, based on only Euclidean distance or nearest search approach, for a large number of patterns increases the challenge exponentially. Based on our experimentation, Table 1 illustrates a straightforward discrimination ability analysis of feature attributes for two different approaches: DCT and DWT. Corpus considered for this are digits, 0 – 9; each digit uttered five times and by seven different speakers. Inter class average variance are compared to judge the discrimination ability. In a trivial notion difference between any observed variance and the calculated average class variance could indicate the class index. In order to get a rough analytical estimate of the capacity, we compared the variance of a random pattern. It is observed that DWT parameters are more suitable for digit recognition as they exhibit more dynamism in variance σ. Measure of variation is nothing but the average of class variance σ_i.

$$\sigma_i = \frac{1}{N}\sum_{n=1}^{N}\sigma_{ni}. \tag{6}$$

where, N is the number of utterances and σ_{ni} is the variance of nth utterance for the ith digit. Sample frames 0th, 5th and 10th obtained after the normalization for digits 0, 4 and 6 are shown in Fig. 5 for CUAVE database.

Comparing the utterance of 0 – 9, it is observed that lip variation is significant and well discriminating for digits 0 and 4, and in contrast variation is most indiscriminating for digit 6. DCT and DWT coefficients are more suitable for digit recognition as they exhibit higher σ_i. For digit 6, indicate less variation so there is a possibility of decrease in recognition performance. Two other metrics are minimum between the class variance $\Delta\sigma_{\min}$ and average all class variance $\Delta\sigma_{\text{avg}}$. The smaller $\Delta\sigma_{\text{avg}}$ exhibits better energy representation and large $\Delta\sigma_{\min}$ corresponds to the better identification.

$$\Delta\sigma_{\min} = \min_{i \neq j}\left(\sigma_i - \sigma_j\right) \quad \text{for} \quad i, j = 1, 2, \ldots 10 \tag{7}$$

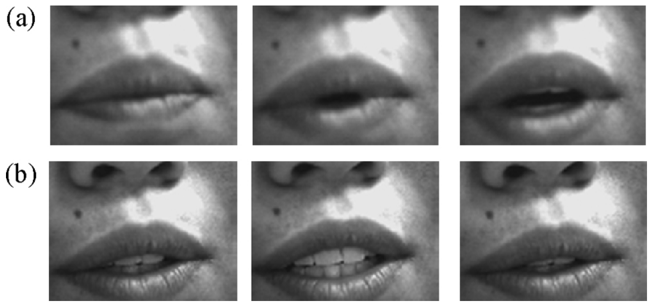


Fig. 6. Frames 1, 3 and 6 for utterance of (a) digit 1 and (b) digit 3 using Tulips database.

Sample frames 1st, 3rd and 6th are obtained after the normalization for digits 1 and 3 are shown in Fig. 6 for Tulips database. Original lip area is shown in Fig. 7(a). Fig. 7(b) is the normalized coefficient of three levels. Small rectangle in Fig. 7(b) indicates the selected coefficients. Fig. 7(c) is the reconstruction of the lip image, containing LL3 band along with 6 significant coefficient LH3 band. These coefficients are used for the classifier input. Recognition performance improves 7 – 10% with addition of some LH3 coefficients compared to only LL3 sub-band coefficients.

3.5. Feature classifiers

Classifier relies on basic assumption that each observation pattern belongs to particular categories. Input to the classifiers is DCT or DWT coefficients as a feature vectors. Feature vectors are classified using different classifiers.

3.5.1. Naïve Bayes

Naïve Bayes is a probabilistic model, which works on Bayes rule with strong assumption that features are independent of given class. Naïve Bayes classifiers can be trained very efficiently in a supervised learning. The Naïve Bayes classifier greatly simplifies learning by assuming that features are independent given class. Although, in practice, independence of features is a poor assumption, Naïve Bayes often competes well with more sophisticated classifiers. In supervised learning problem, probability of Y with respect to X is given by $P(Y|X)$, where, Y is a Boolean-valued random variable, and X is a vector containing n Boolean attributes.

$$X = X_1, X_2, \ldots, X_n.$$

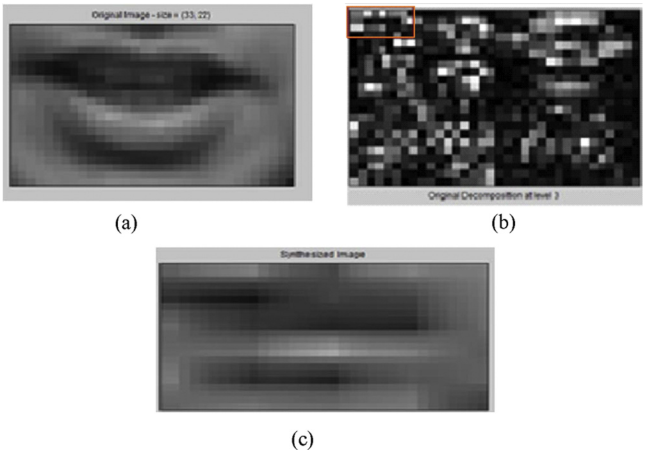

Original Image - size = (33, 22)
Original Decomposition at level 3
(a)
(b)
Synthesized Image
(c)

Fig. 7. (a) Original Lip image (b) Display of three levels coefficient (c) Synthesized image with 36 coefficients (LL-30 and LH-6).

where, X_i is the Boolean random variable denoting ith attribute of class. Applying Bayes rule, $P(Y=y_i|X)$ can be represented as

$$P(Y=y_i|X=x_k)=\frac{P(X=x_k|Y=y_i)P(Y=y_i)}{\sum_j P\left(X=x_k|Y=y_j\right)P\left(Y=y_j\right)}. \tag{8}$$

where, y_i denotes the ith possible value for Y, x_k denotes the kth possible vector value for X. 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.

By making a conditional independence assumption, the Naïve Bayes classifier reduces the number of parameters to be estimated when modeling, $P(X|Y)$, from our original $2(2^n-1)$ to just $2n$. More generally, when X contains n attributes, which are conditionally independent of one another given Y, we have

$$P(X_1\ldots X_n|Y)=\prod_{i=1}^{n}P(X_i|Y). \tag{9}$$

Let us now derive the Naïve Bayes algorithm, assuming in general that Y is any discrete-valued variable, and the attributes $X_1\ldots X_n$ are any discrete or real valued attributes. Our goal is to train a classifier that will output the probability distribution over possible values of Y, for each new instance X that we ask it to classify. Now, assuming the X_i are conditionally independent given Y, we can use Eq. (9) and the expression for the probability that Y will take on its kth possible value, according to Bayes rule is given by,

$$P(Y=y_k|X_1\ldots X_n)=\frac{P(Y=y_k)\prod_i P(X_i|Y=y_k)}{\sum_j P\left(Y=y_j\right)\prod_i P\left(X_i|Y=y_j\right)}. \tag{10}$$

where, the sum is taken over all possible values y_j of Y. Eq. (10) is the fundamental equation for the Naïve Bayes classifier. Given a new instance $X^{\text{new}}=X_1\ldots X_n$, this equation shows how to calculate the probability that Y will take on any given value, given the observed attribute values of X^{new} and given the distributions $P(Y)$ and $P(X_i|Y)$ estimated from the training data. If we are interested only in the most probable value of Y, then we have the Naïve Bayes classification rule:

$$Y\leftarrow \underset{y_k}{arg\max}\frac{P(Y=y_k)\prod_i P(X_i|Y=y_k)}{\sum_j P\left(Y=y_j\right)\prod_i P\left(X_i|Y=y_j\right)}. \tag{11}$$

which, simplifies to the following (because the denominator does not depend on y_k).

$$Y\leftarrow \underset{y_k}{arg\max}\,P(Y=y_k)\prod_i P(X_i|Y=y_k) \tag{12}$$

A discrete probabilistic model classifier is given by Eq. (13).

$$P(y|X_1,\ldots,X_n)=\frac{1}{z}P(y)\prod_{i=1}^{n}p(X_i|y) \tag{13}$$

where, y is a dependent class variable with small number of outcomes. X_1, ..., X_n are feature variables. z is a scaling factor dependent on X_1,...., X_n. $P(y)$ is a class prior and independent of probability distributions. Naïve Bayes classifiers efficiently trained in a supervised learning. An advantage of this classifier is that it requires small amount of training data. In lip reading problem, Bayesian classifiers assign the most likely class of digit described by its feature vector. For lip reading of a digit, variables are given by Eq. (14). Y is the class variable for 10 digits. Total 270 input samples are used. Each sample contains 300 attributes.

$$Y=[y_0,y_1,\ldots,y_9] \tag{14a}$$

$$X=[X_0,X_1,\ldots,X_{270}] \tag{14b}$$

$$X_i=[x_0,x_1,\ldots,x_{300}]. \tag{14c}$$

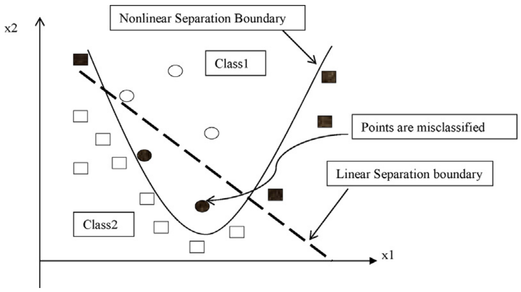


Fig. 8. Two class classifier with SVM separation plane.

3.5.2. *Back Propagation Neural Network (BPNN)*

ANN is the network model consisting of a large number of highly interconnected processing elements organized into layers, of which, geometry and functionality have been similar to the human brain. The ANN may be regarded as possessing learning capabilities in as much as it has natural.

The neural network employed as a classifier for present experimentation is a 10-layered structure. The first layer consists of input element in accordance with feature vectors. The 10 neurons in the output layer are used to represent the digit. The neural network has been trained to adjust the connection weights and biases in order to produce desired mapping by use of back propagation training. The nodes of this network are sigmoid. The network can also be monitored and modified during training time. Learning rate α is 0.3 and moment is 0.5.

3.5.3. *Support Vector Machine (SVM)*

SVM maximizes the distance of separating plane from the closest training data point. Linear and nonlinear data separation by line and hyperbola is shown in Fig. 8. Data separation is completely possible by using nonlinear separation, but it is not using linear separation. The separation can be implemented by using polynomial kernel of SVM [18]. For classification, class decision is based on $f(x)$ value given by Eq. (14).

$$f(x)=\sum_{i=1}^{N}y_i\alpha_i K(X,X_i)+b \tag{14}$$

where, b is a scalar bias, α is the langrage's multiplier, y is an output function and x_i is the support vector obtained from training data. Applying kernels, we do not even have to know that what the actual mapping is. A kernel is a function K, such that

$$K(\boldsymbol{X},\boldsymbol{X}_i)=\Phi^T(\boldsymbol{X})\Phi(\boldsymbol{X}_i). \tag{15}$$

where, $\Phi^T(\boldsymbol{X})\Phi(\boldsymbol{X}_i)$ is scalar product of mapping.

The simplest linear kernel is given by Eq. (15a). Second order Polynomial is used as a kernel of is given by Eq. (15b). John Platt's sequential minimal optimization algorithm (SMO) is used for training a Support Vector Machine. SMO is fast training algorithm.

$$K\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right)=\boldsymbol{X}_{\boldsymbol{i}}^{\boldsymbol{T}}X_j \tag{15a}$$

$$K\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right)=\left[\boldsymbol{X}_i^T\boldsymbol{X}_j+1\right]^2 \tag{15b}$$

3.5.4. *K-Nearest Neighbor (KNN)*

N is non-parametric method of classifying objects based on closest training data. KNN classification is the similar observation belongs to similar classes. Thus, one simply has to look for the nearest neighbors in all the classes, assign the class with maximum number to the nearest neighbors to the unknown. In practice given an instance x, KNN finds the k neighbors nearest to the unlabeled data from the training space based on selected distance measure. In our case, the Eucliden distance is used. Now, let the k neighbors nearest to x be $N_k(x) = p$; $c(z)$, the class label of z and n, the number of classes.

$$j \in \{1, \ldots, n\}$$

$$N_k^j(x) = \{z \in N_k(x) : c(z) = j\} .$$

Finally, the classification result $j^* \in \{1, \ldots, n\}$ defines majority of vote,

$$j^* = arg \left| \max_j \left(N_k^j(x) \right) \right| . \quad (16)$$

KNN is also called as Instance-Based learner (IBk) with fixed neighborhood. K sets number of neighbors to be used. IB1 is equivalent to IBk and the value of k is equal to 1. Recognition rate is more for $k = 1$ as compared to parameters $k = 3$ and $k = 5$.

3.5.5. *Random Forest Tree (RFT)*

A random forest is a classifier, consisting of a collection of tree structured classifiers $\{h(x, \theta_k)\}$, $k = 1, \ldots, n$, where, the θ_k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x [19]. Random forests area combination of tree predictors, such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest.

Forest of tree classifiers error depends on the strength of the individual trees in the forest and the correlation among them. Using a random selection of features to split each node yields error rates that compare favorably to Adaboost classifier, but are more robust with respect to noise. In this experiment, 100 trees and 10 feature vectors are used.

3.5.6. *Performance parameters of classifier*

Performance of the classifiers is evaluated using parameters precision, recall and ROC area. In Eqs. (17 and 18), TP indicates true positive correctly selected number and FP false positive, falsely selected number. Precision is related to TP and FP. FNs false negative. F-score measure that combines precision and recall is the harmonic mean of precision and recall.

In simple terms, high recall means that an algorithm returned most of the relevant results, while high precision means that an algorithm returned substantially more relevant results than irrelevant. Precision is also referred as Positive Predictive Value (PPV). True Positive Rate (TPR) is given by Eq. (17) and false positive rate is given by Eq. (19). Receiver operating characteristics (ROC) analysis is plot of TPR verses FPR. All possible combinations of TPR and FPR compose ROC space.

$$\text{TPR/precision} = (\text{TP})/(\text{TP} + \text{FP}) \quad (17)$$

$$\text{Recall/sensitivity} = (\text{TP})/(\text{TP} + \text{FN}) \quad (18)$$

$$\text{FPR} = (\text{FP})/(\text{FP} + \text{TN}) \quad (19)$$

$$F = 2 \times \text{precision} \times \text{recall}/(\text{precision} + \text{recall}) \quad (20)$$

4. Experimentation

4.1. *CUAVE database*

CUAVE (Clemson University Audio Visual Experiments) [20] 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 × 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 the speaker pairs and the other one of the individuals.

Table 2
Recognition Rate (R.R.) of lip reading using DCT and SVM classifier for individual person.

Person	R.R. (%)
1	96
2	82
3	82
4	74
5	86
6	78
7	74

It contains a mixture of speaker with white and black skin. Database digits are continuous and with pause. Data are recorded with sequential and random manner. Some videos are taken from side view. Total 36 videos are in database, 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 were then compressed into individual MPEG-2 files for each individual speaker and group of two speakers. It has been shown that this does not affect significantly to 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, many researches have been limited to low resolution, pre-segmented video of only the lip region.

4.2. *Tulips database*

A Tulips 1.0 is a small audio-visual 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 Movellan's laboratory at the Department of Cognitive Science, UCSD.

A Tulips 1.V contains the video files in 100 × 75 pixel 8 bit gray level, .pgm format. Each frame corresponds to 1/30 of a second. Movellan presented work on speaker independent visual speech recognition system using simple HMM as a classifier, where TULIPS database of 1 – 4 digits is used for testing [21].

4.3. *Result analysis using different classifier*

The classifier is evaluated by 10-fold Cross-Validation (CV). Cross-validation is a standard evaluation technique in pattern classification, in which dataset is split into n parts (folds) of equal size. $n - 1$ folds are used to train the classifier. By using percentage split (splitting 66% training and 33% testing), results are slightly less than CV (90% training and 10% testing) result. Table 2 indicates individual candidate recognition result of CUAVE database and also shows individual candidate recognition rate (R.R.) for 1 – 7 candidates. Candidate 1 is having highest R.R., while candidates 4 and 7 are having lowest R.R.

Table 3
Lip features using different transformation and % average R.R.

Lip features using different transform	Average R.R. (%)
DCT	75.9
DWT (LL3 sub-band)	78.5
DWT (LL3 + LH3 sub-band)	81.11

Table 4
Average training time for classifiers for CUAVE database.

Type of transform	Naïve bayes (s)	BPNN (s)	SVM (s)	KNN ($n=1$) (s)	RFT (s)
DCT	0.06	20.03	0.7	0.1	1.07
DWT (db2)	0.16	25.63	0.56	0.1	1.11

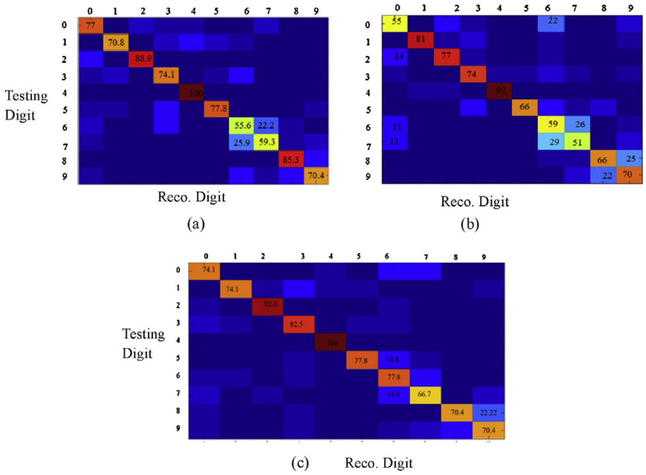


Fig. 9. Confusion matrix for samples of 270 digit utterance for a CUAVE database using feature vectors as (a) DWT with only LL3 coefficients (b) DWT with LL3 and some of LH3 coefficients (c) DCT with upper triangle significant coefficients.

The performance of SVM classifier testing with 10 folds cross validation is shown in Fig. 9(a)(b) and (c). Total 270 data instances are used for training and testing. These results are tested on feature vectors for DWT-dB2 with LL3, DWT-dB2 with LL3 + LH3 and DCT. These figures are confusion matrix for feature vectors with DWT and DCT. The diagonal squares indicate recognized digit and other squares indicate the confusion for input digit. Fig. 9 shows the percentage recognized digit and more confused digit.

Result of confusion matrix shows that six and seven are least recognized digit for DCT and DWT respectively. Four and two are the most recognized digits for both DCT and DWT features. Most of the digits have small confusion with next digit. For DCT and DWT features, zero has more confusion with seven, seven has more confusion with six and six has more confusion with seven. These results show that overall recognition rate for DWT feature vectors is more as compared to DCT. Table 3 shows DWT-dB2 along with LL2 band and some coefficients of LH2 band are used, recognition rate increases by 5 – 7%.

Performance of Naïve Bayes, BPNN, KNN, RFT and SVM classifiers for each digit is evaluated in Fig. 10 with DWT-db2 and DCT features. Average recognition rate for all classifiers with two types of feature vectors indicates that 2 and 4 are the most recognized digits. From Fig. 11, it is found that SVM and BPNN classifiers with DWT–db2 wavelet outperform the rest. Table 4 shows computation time required to train the classifiers. Naïve Bayes and KNN requires less time to train the model, while BPNN is complex network, which requires more time to train among the other classifiers.

Table 5 shows precision, recall, *F*-measure and ROC area with DWT–db2 as feature vectors. These parameters are significant for SVM and BPNN classifiers for CUAVE database. Table 6 indicates similar parameter result for DWT–db2 coefficients for Tulips . For both database Bayes classifier, false positive rate is more compared to other method indicating selection of more irrelevant digits. The

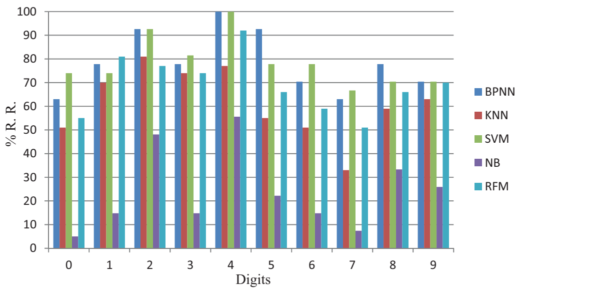


Fig. 10. Classification Rate for different digit utterance for DWT-DB2 (LL3 sub-band) for CUAVE database.

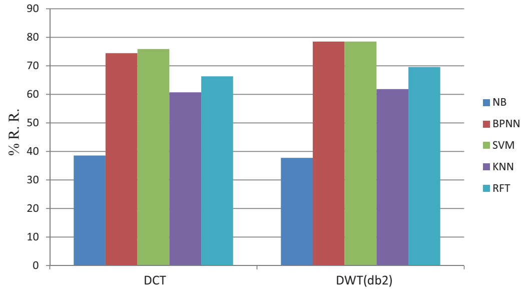


Fig. 11. R.R. (%) of different classifiers result with DCT and DWT coefficients for CUAVE database.

Table 5
Performance parameter of classifier for DWT-db2 using CUAVE database.

Classifiers	FP rate	Precision	Recall	*F*-measure	ROC area
BPNN	0.024	0.792	0.782	0.786	0.963
KNN	0.042	0.659	0.619	0.628	0.788
Bayes	0.085	0.244	0.237	0.226	0.684
SVM	0.024	0.796	0.785	0.787	0.943
RFT	0.034	0.696	0.722	0.704	0.921

Table 6
Performance parameter of classifier for DWT-db2 using Tulip database.

Classifiers	FP rate	Precision	Recall	*F*-measure	ROC area
BPNN	0.049	0.86	0.854	0.853	0.953
KNN	0.139	0.624	0.583	0.593	0.731
Bayes	0.201	0.394	0.396	0.382	0.608
SVM	0.083	0.754	0.75	0.751	0.881
RFT	0.142	0.575	0.573	0.573	0.794

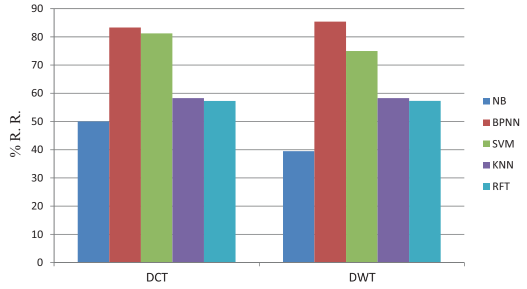


Fig. 12. R.R. (%) of different classifiers results with DCT and DWT coefficients with TULIPS database.

value of recall (sensitivity) is less as compared to precision. Result of ROC area indicates that with DWT-db2 or DCT features, SVM and BPNN classifiers are excellent. For the same features KNN is better and Naïve Bayes classifier cannot be effectively used.

Performance of Naïve Bayes, BPNN, KNN, RFT and SVM classifiers for each digit is evaluated in Fig. 12 with DWT-dB2 and DCT features with Tulips database. For this database BPNN classifiers with DWT-DB2 wavelet outperform the rest.

5. Conclusion

Summarizing the literature review, it has been found that performance of DCT and DWT are at par. DCT with DC component plays an important role as it carries the information related to cavity. In DCT, DC along with low frequency AC coefficients are used as feature vector. In DWT, after applying three levels of decomposition, all LL coefficients and 6 significant LH coefficients are used as feature vector. Classifier namely SVM, BPNN, RFT, KNN and Naive Bayes are compared. Using image transform features and different classification techniques, it is found that digit six is least recognized and four is most discriminative digit for CUAVE database. Performance of SVM is found best, after classification studies done on recognition rate, training time, precision, recall, *F*-measure and ROC area for CUAVE database, while performance of BPNN is found to be the best in Tulips database.

References

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

[2] C. Bergler, Y. Konig. Eigenlips for robust speech recognition. Proceedings of the IEEE International Conference on Acustics, Speech and signal processing, 1994.

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

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

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

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

[7] G.F. Meyor, J.B. Mulligan, S.M. Wuerger, Continuous audio-visual using N test decision fusion, Elsevier J. Inf. Fusion 5 (2004) 91–100.

[8] L. Rothkrantz, J. Wojdel, P. Wiggers, Comparison between different feature extraction techniques in lip reading applications, SPECOM 2006 (2006) 25–29.

[9] S.L. Wang, A. Liew, W.H. Lau, S.H. Leung, An automatic lip reading system for spoken digits with limited training data, IEEE Trans. Circuit Syst. Video Technol. 18 (12) (2008) 1760–1764.

[10] R. Seymour, D. Stewart, Ji Ming, Comparison of image transform-based features for visual speech recognition in clean and corrupted videos, EURASIP J. Video Process. 2008 (2008) 1–9.

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

[12] N. Puviarasan, S. Palanivel, Lip reading of hearing impaired persons using HMM, Elsevier J. Expert Syst. Appl. (2010) 1–5.

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

[14] G. Zhao, M. Barnard, M. Pietikainen, Lip reading with local spatiotemporal descriptors, IEEE Trans. Multimed. 11 (7) (2009) 1254–1265.

[15] Y. Pei, T. Kim, H. Zha, Unsupervised Random Forest manifold alignment for lip reading, ICCV 2013 (2013) 129–136.

[16] Z. Zhou, X. Hong, G. Zhao, M. Pietikainen, A compact representation of visual speech data using latent variables, IEEE Trans. Pattern Anal. Mach. Intell. 36 (1) (2014) 181–187.

[17] P. Viola, M. Jones. Rapid Object Detection using a Boosted Cascade of simple features. IEEE International Conference, 2001, pp. 511–517.

[18] V. Kechman, Learning and Soft Computing: Support Vector Machines, Neural Networks and Fuzzy Logic Models, MIT Press, Cambridge, 2001.

[19] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[20] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy. CUAVE: a new audio-visual database for multimodal human computer-interface research. Proceedings of the IEEE International Conference on Acoustics, speech and Signal Processing, 2002, vol. 2, pp. 2017–2020.

[21] J.R. Movellan, Visual speech recognition with stochastic networks, in: Advances in Neural Information Processing Systems, vol. 7, Cambridge, MIT Press, 1995.