Visual Speech Recognition using MV Features

N. Radha, S. Swetha, Siva

SSN college of Engineering, Chennai, India

radhan@ssn.edu.in, svkumar141@gmail.com

Abstract:- **This paper proposes a visual speech recognition (VSR) system built using combined pixel based features and mean-variance (MV) feature. DCT and DWT are the two pixel transform approaches which are used to extract the features from the visual speech (visemes). This work focuses on the MV features for lip reading that are extracted from each visemes and have been combined with pixel based transform features. VSR system is built separately using DCT coefficients, wavelet coefficients, combined DCT-MV and DWT-MV. The extracted features are modeled by L-to-R Gaussian HMM. The English digit datasets were used for the experimental purpose. The recognition performance of VSR system using DCT features gives 65% accuracy and DWT based features gives 72.3%. The proposed combined system gives 73.5% recognition accuracy for DCT-MV and 86.6% for DWT-MV features. Experiments prove that the combined system improves the recognition accuracy.**

Keywords: *Hidden Markov model, Region of interest, Discrete cosine transform, Discrete Wavelet Transform*

1. INTRODUCTION

Visual speech recognition recognizes the spoken words based on visual information of lip movements which are not affected by any acoustic noise. VSR is an area that has great potential in solving problems in speech processing. Difficulties in the audio based speech recognition system can be significantly reduced by additional information provided by the visual features. It is well known that visual speech information through lip-reading is very useful for human speech perceptions. The various applications of visual lip reading includes speech recognition, speaker's identification and recognition [9].

In [1] the Audio-Visual speech recognition for the embedded devices was proposed. Three functional blocks such as audio processing, lip localization system followed by the feature extraction and audio-visual integration. The MFCC features were extracted from audio signal for automatic speech recognition. Lip finding and lip tracking are the two primary processing steps used in visual processing. To find out no lip position available in the lip frame, geometric model of that lip is used and followed by lip tracking. Visual features are extracted using ASM and DCT techniques and then the feature dimensions are reduced by LDA. DCT feature based recognition is higher than ASM. CUAVE database with continuous digit datasets were used for recognition.

A new robust approach to improve lip localization and tracking is proposed in [2]. Fast lip region detection method using openCV technique is also proposed. The combination of haar features and adaboost classifier are used to detect face from the video. For lip tracking, a component of lab color space value is used and it is computed using mean value of a component specified as the threshold. Experiments conducted for different lip shapes with different lighting conditions and different head poses. Normalized lip images database used

for recognition. In [3] have presented approach for automatic localization for lip feature points. ASM is used to locate different feature points on lip frames. Database for French vowels uttered by multiple speaker were used and the recognition was tested against the database.

Visual feature extraction using DCT based nonlinear predictive coding (NPC) is proposed in [4]. DCT coefficients are extracted and trained using NPC structure in which a feed forward multilayer perceptron is used to extract the features. This process limits the DCT coefficient to a proper size without any significant loss in recognition accuracy. Visual speech database with uttered 10 sentences out of 2 are same sentences and others are different sentences. In [5] lip reading system using DCT-PCA based method was proposed. ROI is tracked from video and has been normalized. Two different ways of DCT coefficients are extracted from ROI such as block based and entire manner. In block based 8x8 and for entire manner 32x16 frame size is used. PCA based features also extracted from ROI. A database HIT Bi-CAVDB consists of syllables of Chinese language was used for experimental purpose. Best recognition is achieved for block based DCT method. DCT has been demonstrated to be superior to PCA for visual feature extraction.

DWT transforms a discrete visual speech signal into wavelet representation [6].Visual feature extraction method using DWT is an effective multi resolution procedure and generates visual feature (wavelet coefficients) can be used as an efficient feature vector for VSR [7]. The advantage of DWT over DCT is transforms the whole visual image into inherent scaling concept. PCA is used to reduce the dimension of DCT, DFT and DWT based features and these features used to improve the recognition accuracy of VSR system [4,5]. The most frequently used dimension reduction techniques for VSR such as LDA, PCA and maximum likelihood linear transform. Visual speech system using pixel based method contributes more towards to extract robust features for that system. Pixel methods are sensitive to lighting conditions and variant to basic transformation such as translation, rotation and scaling. The pixel based visual speech system performance of degrades under various circumstances. We proposed new MV features for VSR.

Figure 1. shows the block diagram of the proposed system. Face detection, Lip localization and lip detection are performed using Violo-Jones algorithm. One of the temporal segmentation technique called pixel pair wise comparison is used to reduce the speaking image frame and non speaking one[7,9]. Three types of features namely DCT coefficients, wavlet coefficients and MV features are extracted from the lip ROI. DCT and DWT features (Image based) were modeled by HMM individually. Then the DCT, DWT based visual speech system performance improved by combining MV (μ, σ^2) features with pixel based features. The recognition performance tested for both individual and the combined one. Given time-asynchronous pixel based and MV feature vectors $o_{\mathrm{P}}(f_{c,}f_{w})$ and $o_{mv}(f_{\mathrm{mv}})$, respectively, a simple concatenative feature fusion is used for combining and thus defined as:

$$o_{pmv,t} = \left[o_{p,t}\left(o_{c,t}, o_{w,t}\right), o_{mv,t}\right] \in R^{f_{pmv}} \quad \text{where } f_{pmv} = f_p + f_{mv} \qquad (1)$$

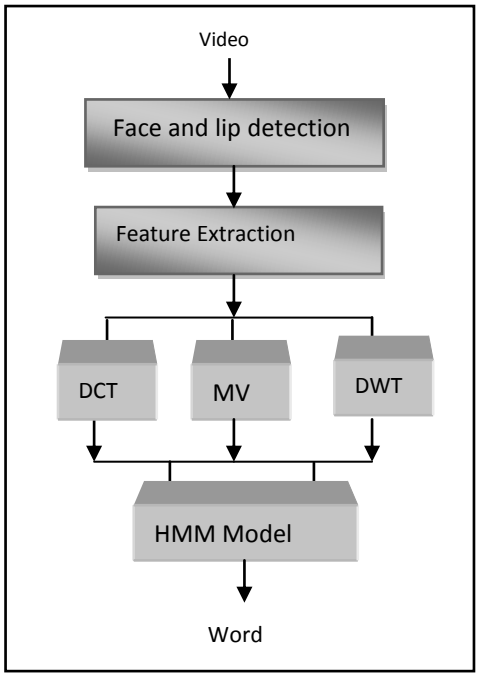


Fig 1. Proposed VSR system

Outline of paper is as follows. We introduce the MV based analysis for visual speech in Section 2. In Section 3 we present a various feature extraction methods and modeling used in this work. The performance analysis is discussed in Section 4. Section 5 concludes our paper.

2. VISUAL SPEECH ANALYSIS

The visemes are analyzed by mean absolute difference. The two cost functions namely mean absolute difference ($C1$) and mean squared error are used to match the closest one in the current block of a frame. In this work $C1$ is used for analysis. Cost function is given as.

$$C1 = \frac{1}{k^2}\sum_{m=0}^{k-1}\sum_{n=0}^{k-1}|c_{mn} - p_{mn}|$$

The mean differences computation of visual phonemes of digits for different utterances analysis given in figure 2. The digits of zero and four corresponding viseme mean absolute differences are given in figure 2.a,b. For digit zero and four, frame number from 9 to19 and 5-35 respectively shows the speaking viseme differences. Different utterances uttered for the same word ‘zero’ and ‘four’ have similar mean differences. The next section discuss about the database for study, different feature extraction methods and modelling used in visual lip reading.

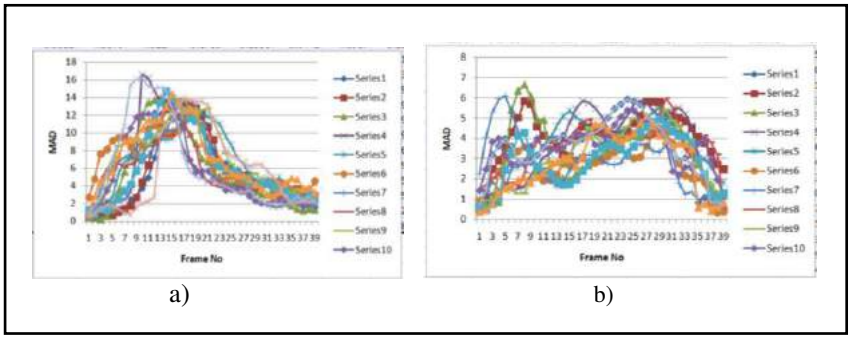


Fig. 2. Mean absolute difference of consecutive frames for the different utterances of the same word "zero" and "four" ((a) and (b)).

3. FEATURE EXTRATION

This section deals with the database for study, face identification, ROI detection, extracting motion based features and MV features for building the VSR systems.

3.1. *Database for this study*

The database used in this study consists of a video corpus which is collected from 10 (12 male and 8 female) speakers. Simultaneous recording of audio and visual speech is carried out using a microphone and a camera, respectively. The SONY Handycam HDR-PJ660/B Camcorder is used for the video recording. Each speaker records simultaneously the audio and video of 50 utterances of each of the10 digits, 0 to 9, out of which 35 utterances are used for training and 15 utterances are used for testing. A total of 7000 utterances are used for training and 3000 utterances are used for testing for all the ten digits.

All the visual utterances are recorded under the same lighting and normal environmental conditions. The video consists of 50 frames per second with a frame width of 640 and a frame height of 480. The horizontal distance from d1 the speaker's face position to the camera is about 32 cm and camera is at a height d2 of 63 cm from the ground. A video of a sound unit consists of a large number image frames. Hence image frame $\left(I(x,y)\right)$ for each viseme is reduced further using temporal segmentation technique. Face detection from video is the first process in VSR system is discussed next.

3.2. Temporal segmentation

In this study, a feature invariant technique the Viola-Jones algorithm is used. This algorithm detects a face in an image by scanning sub windows of the image multiple times with a re-scalable detector. The scale invariant detector is constructed using an integral image and Haar-like features. The Viola-Jones algorithm uses a 24x24 window as the base window size to evaluate the features. Since a large number of rectangular Haar like features have to be evaluated, to reduce computation, to find the best features and eliminate redundancy, Adaboost machine learning algorithm is used. This classifier constructs a strong classifier as a weighted combination of weak classifiers. This algorithm is invariant to pose and orientation changes. Once the face is detected the lip region is extracted next using the same algorithm.

The ROI extraction is a pre-processing step for extraction of visual features. It's simply defined as a rectangle containing the intensity of the speaker's mouth region. The ROI is normalized into a 64x40 frame which represents the visual speech information. Temporal

segmentation (TS) is performed over an image frame (ROI). The aim of TS is automatically find the start and end frames from an image sequence. Pair wise pixel comparison is one of the temporal segmentation concept used to remove unused frames in visemes. A small shot changes are detected for lip frames using a simple global inter-frame difference measure [7], defined as

$$\left(\sum_{x=1}^{X}\sum_{y=1}^{Y} P(I_t(x,y)) - \sum_{x=1}^{X}\sum_{y=1}^{Y} P(I_{t-1}(x,y))\right) > T$$

Lip frames are discarded if the differences of current and the previous lip frame comparison is greater than the threshold. Table 1 shows the temporal segmentation calculated using frame difference measure for digits.

Table 1: Temporal segmentation using frame difference measure

Digits	Total no frames (Before detection)	Total no frames (After detection)
One	54	47
Two	58	49

Total frames out of 7 and 11 frames are discarded for digit one and two respectively. Similar operation carried out for all other digits. After temporal segmentation, visual feature extraction using MV and pixel based from the extracted ROI is discussed in the next sub section.

3.3. Image transforms

From the given input visual lip image databases DCT coefficients are calculated from every viseme and PCA is used to reduce the dimensionality of that features. The total of 64 DCT coefficients per digit is extracted and thus the feature vector F^{mc} is given as

$$F^{mc} = [f_1^{mc}, f_2^{mc}, f_3^{mc} \ldots\ldots f_{64}^{mc}] \quad where f_1^{mc} = F_{xy} = \mathrm{DCT(I(x,y))} \qquad (2)$$

$$F_{xy} = \alpha_x \alpha_y \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} I_{i,j}\, cos\frac{(2i+1)y\pi}{2N} cos\frac{(2j+1)x\pi}{2N} 0 \le k \le N-1 \qquad (3)$$

$$\text{where } \alpha(0) = \sqrt{\frac{1}{N}}, \alpha(k) = \sqrt{\frac{2}{N}} 1 \le k \le N-1 \qquad (4)$$

The dynamic lip movements are well captured by the DCT coefficients. The visual feature extraction using DCT is sensitive to lighting conditions. Hence DWT is used and extracts wavelet coefficients from the lip frames. DWT can be applied as a convolution of a selected wavelet function with an original image or it can be seen as a set of two matrices of filters, row and column one shown in figure 3. Using separability property of DWT, the first part of decomposition consists of an application of row filters to the original image. The column filters are used for further processing of an image. This image decomposition can br mathematically described by,

$$w = I(x,y) * w_m$$

Where w is the matrix wavelet coefficients, w_m represents wavlet function. In the first level of decomposition of 2D DWT, the image is separated into four parts. Each of them has a quarter size of the original image. They are called approximation coefficients (LowLow or LL), horizontal (LowHigh or LH), vertical (HighLow or HL) and detail coefficients (HighHigh or HH). Approximation coefficients obtained in the first level can be used for the next decomposition level.

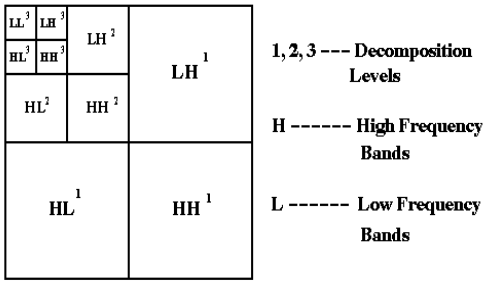


Fig 3. DWT decomposition into two levels

DWT have inherent multi-resolution nature property. DWT is an iterative sub-band process, decomposes the visual signal into approximation and detailed coefficients. Level-4 sub-band decomposition is used in this work [8]. The low pass filtered approximation coefficients contain significant amount of information about lip as compared to the other coefficients. Hence only approximation coefficients are considered as a feature in this work. The total of 256 DWT coefficients per frame is extracted. The visual features are combined with MV features is discussed in the next section.

3.4 *MV based visual feature extraction*

A simple absolute mean difference concept is used to detect MV feature from the lip frames. From each current frame and the previous frame difference of absolute mean μ and variance σ^2 is calculated and defined as

$$\mu = \left(\sum_{x=1}^{X} \sum_{y=1}^{Y} P(I_t(x,y)) - \sum_{x=1}^{X} \sum_{y=1}^{Y} P(I_{t-1}(x,y)) \right) / N$$

$$\sigma^2 = \sum_{x=1}^{X} \sum_{y=1}^{Y} (P(I_t(x,y) - \mu)^2 / N$$

The dimension of MV feature (μ, σ^2) is 39. The MV features are combined with DCT and DWT features. The extracted visual features and the combined features were modelled using HMM is discussed in next section.

3.5. Viseme Modelling

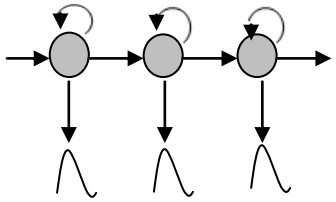
Each Visemes are modelled by a single stream of left-to-right Gaussian HMM is shown in figure 4. Each state in HMM is a Gaussian Mixture Model (GMM). GMM is a parametric model characterized by a mean, and the variance. The common method for computing visemes likelihoods is a GMM probability density function (pdf) of the observed feature vector, o is given by

$$b_i(o) = \frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma_i{}^{-1}|^{\frac{1}{2}}}\, e^{-\frac{1}{2}(o-\mu_i)^{\mathrm{T}}\Sigma_i{}^{-1}(o-\mu_i)}$$

Where μ_i is the state mean vector, Σ_i is the covariance matrix and N is the dimension of feature vector o. The probability of the feature vector o being in any of I Viseme models denoted by , is shown as.

$$p(o|\lambda) = \sum_{i=1}^{I} w_i \; b_i(o)$$

Where w_i are the mixture weights and $\sum_{i=1}^{I} w_i = 1$. For each Viseme a GMM model is represented by GMM mean, covariance and a weight parameter given by, $\lambda = \{w_i, \mu_i, \Sigma_i\}$



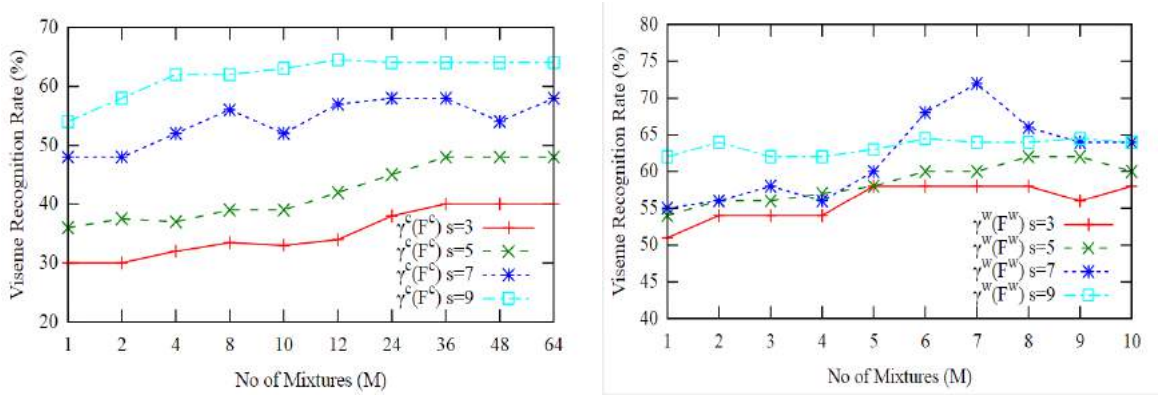
a) Left-to-right HMM

Fig 4. Viseme-state synchronous single stream HMM

HTK toolkit is used for model building. The result of this study is discussed in Section 4.

4. EXPERIMENTAL RESULTS

The viseme level HMM models, which are having L-to-R states with varying number of Gaussian mixture(M), are evaluated with DCT, DWT, combined DCT-MV and DWT-MV feature sets F^c, F^w, F^{cmv}, and F^{wmv}, respectively. The corresponding visual speech systems built are, $\gamma^c(F^c)$, $\gamma^w(F^c)$, $\gamma^{cmv}(F^{cmv})$, $\gamma^{wmv}(F^{wmv})$, Viseme recognition rates for varying number of Gaussian mixtures with 5 states are plotted in Figure 5. The pixel based recognition system, $\gamma^c(F^c)$, has state s=9 and M=12 which maintains highest recognition 65% accuracy. This system yields poor performance recognition rates at which state of s=1/s=2, and M=2[8].



a) b)

Fig 5: Recognition rate of different visemes of $\gamma^c(F^c)$ and $\gamma^w(F^w)$ systems[8].

The recognition system, $\gamma^w(F^w)$, sets a baseline recognition rate at 72% with state branch s=7 and the Gaussian mixture per state is M =7. The proposed system, $\gamma^{cmv}(F^{cmv})$, has been evaluated at the fusion of feature level and improves the performance of $\gamma^c(F^c)$ the system. Similarly, $\gamma^{wmv}(F^{wmv})$, has also been experimented at feature level fusion mode and thus default improves the performance $\gamma^w(F^w)$ the system. Viseme recognition rates of the γ^{cmv}, γ^{wmv}, system of with estimated corresponding feature vectors F^{cmv}, F^{wm} for varying different states are plotted in figure 6. Noted that the system γ^{cmv}, γ^{wmv}, has achieves high recognition accuracy 72.3%, 86.6% at the state s=9, M=24 and s=9 M=7 respectively .

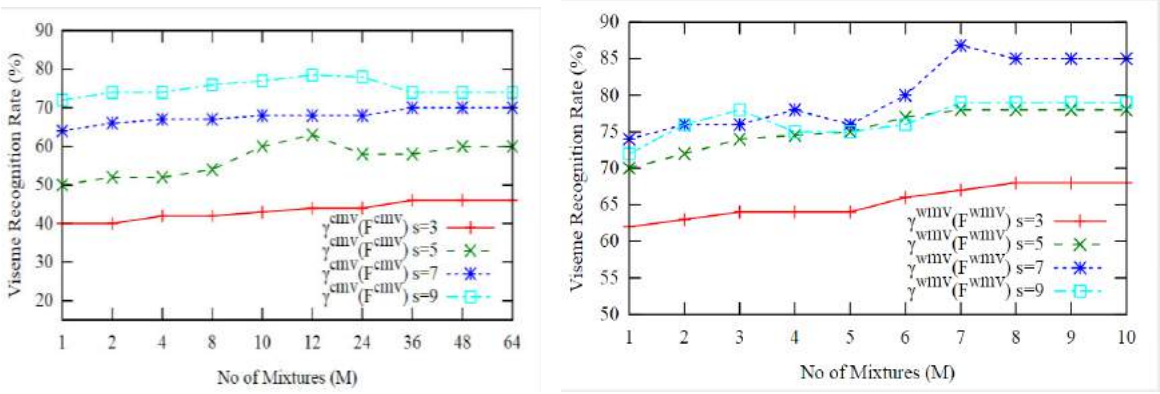


Fig 6: Recognition rate of different visemes of $\gamma^{wcv}(F^{wcv})$ and $\gamma^{wmv}(F^{wmv})$ systems.

5. CONCLUSION

Visemes provide lesser discriminatory information than compared to acoustic signal on the sound units, a vocabulary of digits dataset is chosen for this work and allows better discrimination in sounds. We have presented a pixel based visual feature extraction method that extract DCT and wavelet coefficients. A combined feature based framework for pixel based features and MV features are proposed, used to improve the conventional pixel based VSR system. A simple plain feature concatenation technique is used for combining features. The four viseme models of two models built individually and remaining two for combined features is done using L-to-R Gaussian HMM and the recognition accuracy is tested.

The experimental analysis shown that combined features significantly improve the performance of the pixel based VSR system. The VSR system has the least recognition performance rate of 65% for DCT features and 72.3% DWT features[9]. VSR system performs significantly better while using the DCT+MV (74.5%) and DWT+MV (86.6%). This improvement in performance due to the fusion shows the presence of complementary cues in the pixel based and MV based features, and also that MV features provide better discrimination of visemes. The overall relative performance of the proposed work of VSR system is encouraging.

References

[1] Pérez, Jesús F. Guitarte, et al. "Lip reading for robust speech recognition on embedded devices." *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*. Vol. 1. IEEE, 2005.

[2] WenJuan, Yao, Liang YaLing, and Du MingHui. "A real-time lip localization and tacking for lip reading." *Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on*. Vol. 6. IEEE, 2010.

[3] Werda, Salah, Walid Mahdi, and Abdelmajid Ben Hamadou. "Lip localization and viseme classification for visual speech recognition." *arXiv preprint arXiv:1301.4558* (2013).

[4] Azar, Mahmood Yousefi, and Farbod Razzazi. "A DCT based nonlinear predictive coding for feature extraction in speech recognition systems." *Computational Intelligence for Measurement Systems and Applications, 2008. CIMSA 2008. 2008 IEEE International Conference on*. IEEE, 2008.

[5] Hong, Xiaopeng, et al. "A PCA based visual DCT feature extraction method for lip-reading." *Intelligent Information Hiding and Multimedia Signal Processing, 2006. IIH-MSP'06. International Conference on*. IEEE, 2006.

[6] Nehe, Navnath S., and Raghunath S. Holambe. "DWT and LPC based feature extraction methods for isolated word recognition." *EURASIP Journal on Audio, Speech, and Music Processing* 2012.1 (2012): 7.

[7] Lefèvre, Sébastien, Jérôme Holler, and Nicole Vincent. "A review of real-time segmentation of uncompressed video sequences for content-based search and retrieval." Real-Time Imaging 9.1 (2003): 73-98.

[8] Radha, N., A. Shahina, and A. Nayeemulla Khan. "An Improved Visual Speech Recognition of Isolated Words using Combined Pixel and Geometric Features." *Indian Journal of Science and Technology* 9.44 (2016).

[9] Radha, N., A. Shahina, and A. Nayeemulla Khan. "A person identification system combining recognition of face and lip-read passwords." *International Conference on Computing and Network Communications (CoCoNet), 2015*.

[10] Koprinska, Irena, and Sergio Carrato. "Temporal video segmentation: A survey." *Signal processing: Image comunication* 16.5 (2001): 477-500.