Indian Journal of Science and Technology, Vol 9(44), DOI: 10.17485/ijst/2016/v9i44/102234, November 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

An Improved Visual Speech Recognition of Isolated Words using Combined Pixel and Geometric Features

N. Radha[1]*, A. Shahina[1] and A. Nayeemulla Khan[2]

[1]Department of Information Technology, SSN College of Engineering, Chennai, India; radhan, shahinaa@ssn.edu.in
[2]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India; nayeemulla.khan@vit.ac.in

Abstract

Objectives: This paper proposes a method to improve the performance of a Visual Speech Recognition (VSR) system by combining the pixel-based and geometry-based features, so as to augment the performance of audio based Automatic Speech Recognition (ASR) systems in adverse conditions. **Methods/Statistical Analysis:** A video database comprising of 11000 utterances of isolated words, collected from 20 speakers, is used in this study. Pixel based features (DCT and DWT) and geometric features (Active Shape Model or ASM) are fused at two levels, one at the feature level and the other at the decision level. A simple Gaussian mixture HMM word model is built for feature level fusion, while a two stream HMM model is built for decision level fusion. **Findings:** The VSR system built using the combined features shows a significant improvement in performance when compared to individual VSR systems built using pixel and geometric based features. The accuracy of the individual system is 76% for geometric features, 64% for DCT and 72% for DWT pixel-based features. The performance improves for combined features with an accuracy of 80% for ASM+DCT and 84.7% for DWT+ASM. A weighted decision level fusion result in further improvement, with an accuracy of 84% for ASM+DCT and 92% for ASM+DWT. **Application/Improvements:** The combined VSR could be preferred over individual pixel/geometric feature based systems to augment the performance of audio based Automatic Speech Recognition (ASR) systems in adverse conditions. Further studies on improving the VSR system, which could be used in lieu of audio-based ASR systems in adverse situations, are being carried out.

Keywords: HMM, Pixel and Geometric Features, Visual Speech Recognition

1. Introduction

Visual lip reading or VSR involves conversion of visual information, derived from a sequence of images of the lip movements, into text. The basic unit of visual speech is called a viseme. VSR is one of the approaches used to enhance the robustness of speech recognition systems. This is because one of the major drawbacks of ASR systems is their sensitivity to environmental noises. The acoustic features of some phonemes are not clearly distinguishable with respect to their place of articulation (eg., /m/ and /n/). However, in the case of visual speech, these visemes are easily distinguished. In contrast some of the phonemes are visemically very similar, but the acoustic characteristics are distinguished (eg., /t/ and /n/). Lip movement based visual speech provides complementary information to acoustic features which could be exploited to develop robust speech recognition systems. Visual lip reading is relevant for human-computer interaction (HCI). The performance of VSR depends on appropriate features extracted from a sequence of lip movements. Recent studies on different methods for feature extraction are discussed as follows.

In[1] geometric feature, namely height and width of the oral cavity, derived from multiple images templates were used. An improvement in recognition was reported over

*Author for correspondence

single template approach for a 22-consonant (/a/-/c/-/a/) set. In[2] geometric features such as width, height and four distances of upper and lower curves were extracted from the lip contour which was modeled using cubic curves. The extracted feature size was reduced using D-LDA and classified using a HMM classifier[3] had proposed pixel based features for VSR of Chinese syllables. DCT and block-based DCT coefficients with PCA for dimensionality reduction were used. Both DCT and block-based DCT systems were shown to have comparable performance[4] involved a comparative study of four different ROI, derived using gray scale normalization, Fisher transformation, Sobel edge enhancement and binarization techniques. DCT coefficients were used as feature for VSR. Gray scale normalization for DCT features were found to achieve better results[5] had shown that linear discriminant analysis performed on DCT coefficients derived from the ROI located based on the lip contour achieved higher recognition accuracy that using PCA on those features. Both geometric features and pixel-intensity based features were compared for the visual speech recognition task in[6]. Active appearance model (AAM) and ASM were used to derive the geometric features, while multiscale spectral analysis (MSA) was used for deriving pixel-based features. AAM were shown to achieve better recognition rates than ASM, while MSA showed comparable results.

We employ three benchmark VSR systems and two proposed VSR systems (refer Table 1), which are defined based on their input streams. The first VSR system, γ^{c} is based on DCT coefficients as features, the second system, γ^{w} is based on DWT wavelet features, and the third system, γ^{m}, is based on ASM features. The fourth and fifth systems are the proposed combined systems: γ^{cm}, using DCT and ASM features, and, γ^{wm} using DWT and ASM as features. Each VSR system consists of the following sequence of steps: face detection, lip tracking, visual feature extraction, modeling and recognition.

Figure.1 shows a block diagram of proposed combined VSR system. Face detection and lip tracking are performed using the Viola-Jones Algorithm[7]. Once the lip region is identified, the visual features are extracted. Pixel based features are extracted using the DCT and DWT image transformation techniques. Since the pixel based are not robust under variation in illumination conditions, additionally geometric features are extracted from the lip region. But the geometric features do not capture the inner variations in the sequence of lip movements. So, combined pixel and geometric based features are proposed in this paper. The extracted features are concatenated and the corresponding recognition accuracy is tested. Given time-asynchronous pixel and geometric observation feature vectors $o_p(f_c, f_w)$ and $o_g(f_g)$, respectively, concatenative feature fusion is defined as:

$$o_{pg,t} = [o_{p,t}, o_{g,t}] \in R^{f_{pg}} \text{ Where } f_{pg} = f_p + f_g \quad (1)$$

$$p(o_t(pg,t) \mid s) = \prod_{i=1}^{c} p(o_t(p,t) \mid s)^{\gamma_i(p,s,t)} \prod_{i=1}^{c} p(o_t(g,t) \mid s)^{\gamma_i(g,s,t)} \quad (2)$$

Modeling plays an important role due to significant variation in visual speech information during viseme production. The decision fusion also performed a two stream HMM model is used to perform decision fusion in this work. For this model, feature vector is given in (2). For all HMM states $s \in S$, α denotes weights for stream component and uttered viseme frame at t.

This paper is organized as follows. Face detection and ROI tracking are discussed detail in Section 2. Section 3 analysis the experimental results of the visual speech recognition study. Section 4 summarizes the work.

Table 1. Benchmark and proposed systems

Symbols	Bench Mark systems	Symbols	Proposed system
γ^{c}	DCT base	γ^{cm}	DCT-ASM base (feature level)
γ^{w}	DWT base	γ^{wm}	DWT-ASM base (feature level)
γ^{m}	ASM base	γ^{cm}, γ^{wm}	DCT/DWT-ASM (Decision level)

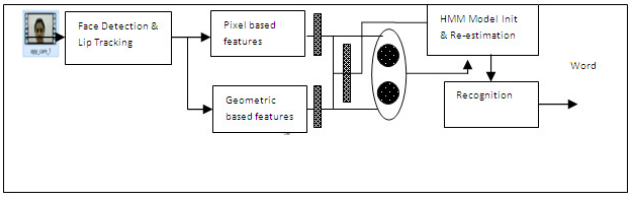


Figure 1. Block diagram of a proposed system.

2. Visual Speech Recognition Study

This section deals with the identifying the face, tracking the ROI, extracting pixel and geometric based features and combining them for building the VSR systems.

2.1 Database for this study

The database used in this study consists of a video corpus which is collected from 20 (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 utterances are recorded under the same lighting and under 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 the speaker's position to the camera is about 32 cm and camera is at a height of 63 cm from the ground. A video of a sound unit consists of a large number image frames. Hence each image frame $(I(x, y))$ is classified further as speaking or non-speaking frame, and only the speaking frames are considered for the study. Face detection which is the first process in VSR system is discussed next.

2.1 Face and Lip Region Detection

The objective of face detection is to determine whether a face is present in the video image. To perform this first videos are converted into sequence of frames. Reliability of the visual speech systems depends upon the accurate detection of the face. Face detection to be more robust to changes in illumination effects and thus exhibits a reliable detection performance. Face detection techniques can be broadly categorized based on knowledge, appearance, template matching, feature invariance and color[8]. Knowledge based techniques encode the human knowledge of a typical face using rules to define the relationship between facial features. Difficulty in precise translation of human knowledge into rules sometimes results in failure to detect faces. In template matching, multiple patterns are used to describe a face to capture the variations in pose and lighting conditions. Enumerating templates for different poses is difficult. Different models are learned from a set of training images in appearance-based methods. Color based schemes are insensitive to variation in expression and rotation, but are sensitive to environment and lighting changes.

In this study, the Viola-Jones algorithm, a feature invariant technique is used. The Viola-Jones algorithm unlike other techniques, 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. This 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. Viola-Jones 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 is normalized into a 64x40 frame which represents the visual speech information. 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. From the extracted ROI, visual features are extracted as discussed in the next subsection.

2.3 Pixel based Visual Feature Extraction

Visual feature extraction for VSR requires the following considerations: A robust method to track the speaker's lips through a sequence of images, variability considerations in properties such as skin color, lip contour, lip width, amount of lip movement during speech, bearded speakers and environmental variability such as lighting conditions, and the appearance of teeth and tongue in the visual signal. Commonly the visual feature extraction methods are classified into a model based, image/pixel based, motion based and color based (hue and saturation). Model-based methods such as deformable templates[1], active snake model[9,10], ASM[11,12] and image transform based methods such as PCA, DCT, DWT and LDA are used for feature extraction. In addition to these, visual motion analysis and hue and saturation thresholding methods[13] are used to extract the visual features. In our work, model and image based methods are used. A combined DCT/DWT-ASM feature is explored. Visual feature extraction using DCT and DWT and ASM is discussed next.

From the given input image database PCA is applied for reducing the dimensionality of the image. Block (8x8) based DCT is then applied and the 64 DCT coefficients are extracted from the feature vector F^c given by:

$$F^c = [f_1^c, f_2^c, f_3^c \ldots\ldots f_{64}^c] where f_1^c = F_{xy} \quad (3)$$

$$F_{xy} = \frac{\alpha_x \alpha_y \sum_{i=0}^{N-1} \frac{\sum_{j=0}^{N-1} I_{i,j} \cos(2i+1)y\pi}{2N} \cos(2j+1)x\pi}{2N} \; 0 \le k \le N-1 \quad (4)$$

where

$$\alpha(0) = \sqrt{\frac{1}{N}}, \alpha(k) = \sqrt{\frac{2}{N}} \; 1 \le k \le N-1 \quad (5)$$

The dynamic lip movement is well captured by the DCT coefficients. DWT was used in this work because of their inherent multi-resolution nature. DWT iteratively sub-band decomposition of the visual speech signal into approximation and detailed coefficients. For this study, level-4 sub-band decomposition is used. 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 256 DWT coefficients per frame are obtained as:

$$f(x) = \frac{1}{\sqrt{M}} \sum_k W_\phi(j_0,k)\phi_{j_0,k}(x) + \frac{1}{\sqrt{M}} \sum_{j=j_0}^{\infty} \sum_k W_\psi(j,k)\psi_{j,k}(x)$$

where j_0 is an arbitrary starting scale

$$W_\phi(j_0,k) = \frac{1}{\sqrt{M}} \sum_{x=0}^{M-1} f(x)\tilde{\phi}_{j_0,k}(x)$$

called the approximation or scaling coefficients

$$W_\psi(j,k) = \frac{1}{\sqrt{M}} \sum_{x=0}^{M-1} f(x)\tilde{\psi}_{j,k}(x)$$

called the detailed or wavelet coefficients

2.4 Geometric based Visual Feature Extraction

Geometric features are extracted using shape based models[14]. Commonly, used shape models are deformable template[1], snake model or ACM, ASM and AAM[1] had presented multiple deformable templates which are specified with a finite number of parametric curves, a set of geometric and a set of rules for fitting the curves to features in an image and require prior knowledge about the shape of a lip image. Deformable template method suffers from changes in lighting conditions. ASM extracts model-based features. This technique obtains information about shape of the lip by fitting statistical shape modes of the lip to the video frames. AAM is an extension of ASM. AAM, which is more robust than ASM, combines the shape model with a statistical model of the grey levels corresponding to the mouth region.

In ACM, the lip contour is approximated by a geometric curve called *snake* by means of energy minimization. The *snake* is guided to fit the contour by image forces which push it towards for features such as edges and contours, while the internal forces ensure smoothness and external forces guide it to the desired local minima[9]. The appearance of teeth and tongue generates a large intensity gradient and causes the snake to diverge from the outer lip contours. Learned models of lips are used to constraint the snake. In this study ASM is used. ASM is an iterative algorithm that is shape-constrained since it is obtained from the statistics of hand labeled training data. During training phase, the lip models are built using annotated visual lip images. The landmark point on a training image shows the connectivity and direction indicating the shape of the model. The training models are aligned using a minimizing function, which are then used to find the mean shape x_s. The principal components about the mean shape are determined. Any valid shape is approximated as:

$$x_s = \bar{x}_s + P_s b_s$$

Where $\{P_s\}$ are the eigen vectors and b_s the corresponding weights. The shape parameters are obtained as:

$$b_s = p_s^T (x_s - \bar{x}_s)$$

The model is iteratively fit to an example image. A cost function such as a statistical model of the grey level profiles of a shape model is used. This iterative process converges when there are no significant changes in the shape parameters. The extracted geometric features and the pixel based features were modeled by HMM separately. HMM models were also built for fusion of the pixel and geometric features carried out at the feature and decision levels.

2.5 Feature fusion based system

In this work, features from the different modalities (DCT/DWT with ASM) are concatenated and modeled together by single stream left-to-right Gaussian HMM as shown in Figure 2. Each state in the HMM is a Gaussian mixture. The probability of the feature vector o being in any of I viseme models denoted by λ, is shown as:

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

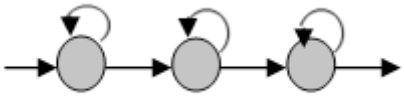
where w_i are the mixture weights and $\sum_{i=1}^{I} w_i = 1$. For each viseme a HMM model is represented by GMM

mean, covariance and a weight parameter given by, $\lambda = \{w_i, \mu_i, \Sigma_i\}$.

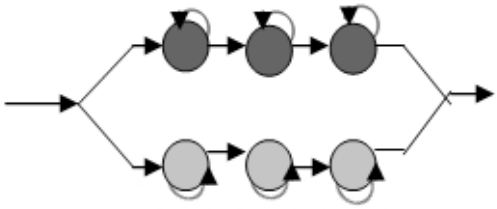
The decision level fusion is modeled by a state synchronous two stream Gaussian L-to-R HMM (product HMM) that combines a stream of log-likelihoods at an intermediate level for visemes[15]. It consists of composite states $s \in S$ with emission score values $s = \{s_n, n \in N\}$. as shown in (6). An example of such a model is depicted in Figure 2.

$$p(o_{p,g,t} \vee s) = \prod_{n \in N} p(o_{n,t} \vee s_n)^{\alpha_n} \quad (6)$$

Each model is represented by 5 states, out of which 2 states are the starting and ending states and the other 3 states represent the features. The products HMM have the same number of mixture weight, mean, and variance parameters. Each state has two stream components and can be controlled by weighing that stream.



a) Left-to-right HMM



b) Two stream L-to-R HMM

Figure 2. HMM models for feature level fusion, and decision level fusion.

3. Performance Analysis

The viseme level HMM models, which have L-to-R states with varying number of Gaussian mixtures (M), are evaluated with DCT, DWT, ACM, combined DCT-ASM and DWT-ASM feature sets denoted as F^c, F^w, F^m, F^{cm}, and F^{wm}, respectively. The corresponding VSR systems built are $\gamma^c(F^c)$, $\gamma^w(F^w)$, $\gamma^m(F^m)$, $\gamma^{cm}(F^{cm})$, $and \gamma^{wm}(F^{wm}.)$ Viseme recognition rates for varying number of Gaussian mixtures with different states are plotted in Figure 3. The pixel based recognition system, $\gamma^c(F^c)$, has highest recognition accuracy for state s=9 and M=12.

The performance analysis of pixel and geometric based feature's based system recognition is shown in Table 2.The DWT based recognition system, $\gamma^w(F^w)$, has the highest recognition rate of 72% for s=7 and M =7. The recognition system, $\gamma^m(F^m)$, has a recognition 76% for state s=9 and M=12. DCT-ASM fusion based system, $\gamma^{cm}(F^{cm})$, and DWT-ASM based systems $\gamma^{wm}(F^{wm})$ have shown improved performance (80.2% and 84.7%, respectively) over the individual feature based systems as seen in Figure 4.

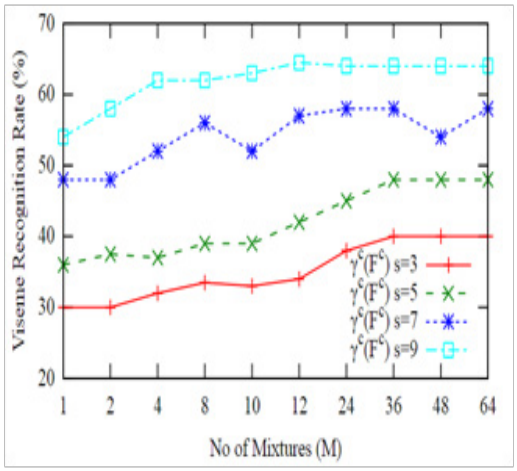
The performance of the system based on decision-level fusion ($\gamma_d^{wm}(F^{wm})$) is the highest (92%) among all the systems. The decision fusion is obtained using a weighted sum of the output probabilities of the DWT-based stream and ASM-based stream given by

$$p = \alpha_w\, p_{\gamma w} + \alpha_m\, p_{\gamma m} p = \alpha_w\, p_{\gamma w} + \alpha_m\, p_{\gamma m} \quad (11)$$

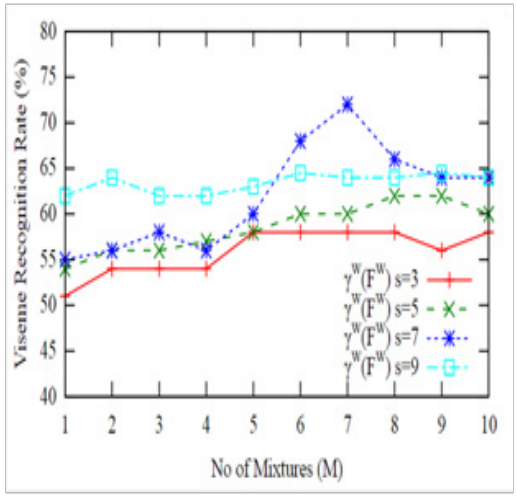
Where α_w and α_m are the weighing factors. The performance for different values of α_w and α_m are shown in Figure. 4(b) $(\alpha_m > \alpha_w, 0.5 \le \alpha_m < 1, and\ 0.1 \le \alpha_w < 0.5)$ and Figure 4(c) $(\alpha_w > \alpha_m, \quad 0.5 \le \alpha_w < 1, and\ 0.1 \le \alpha_m < 0.5)$ the highest performance is obtained when more weightage is given to geometric features than intensity based features, showing that, though there is complementary information in the geometric features and intensity features, the geometric cues play a more relevant role in discriminating visemes. Thus the performance of systems based on both the decision level fusion (92%) and feature level fusion (84.7%) is significantly better than individual systems (64%, 72%, 76% for system based DCT, DWT and ASM, respectively).

Table 2. Performance (%) of Benchmark and proposed systems

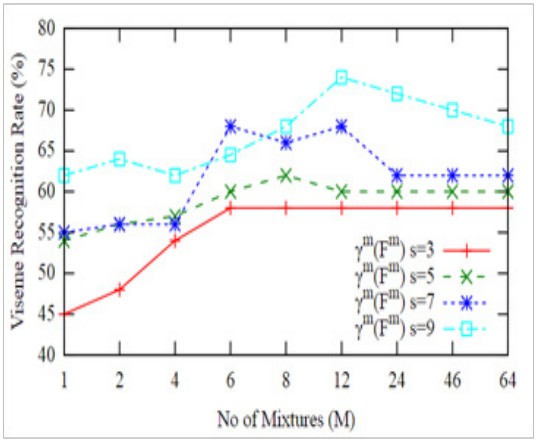
No of states	Conventional(%)			Proposed (%)	
	γ^c	γ^w	γ^m	γ^{cm}	γ^{wm}
1	36.8	50	52	66.5	78.9
3	40	58.7	57.3	69.2	80.5
5	48	62.4	60.1	72	**84.7**
7	58	**72.5**	68.3	**80.2**	83.5
9	**64.5**	65	**76**	78.5	83



a) $\gamma^{c}(F^{c})$ benchmark VSR system

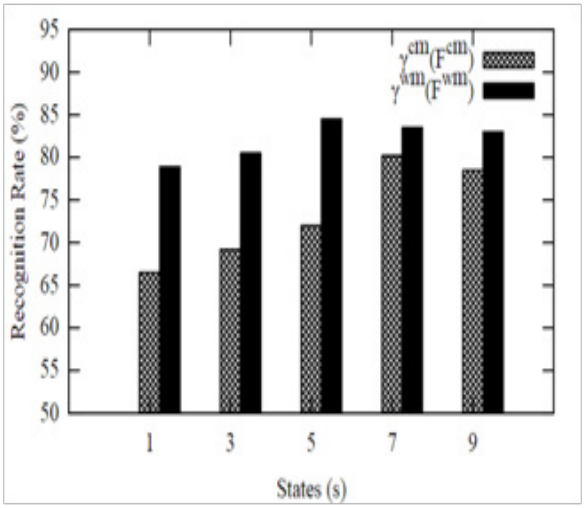


b) $\gamma^{w}(F^{w})$ benchmark VSR system

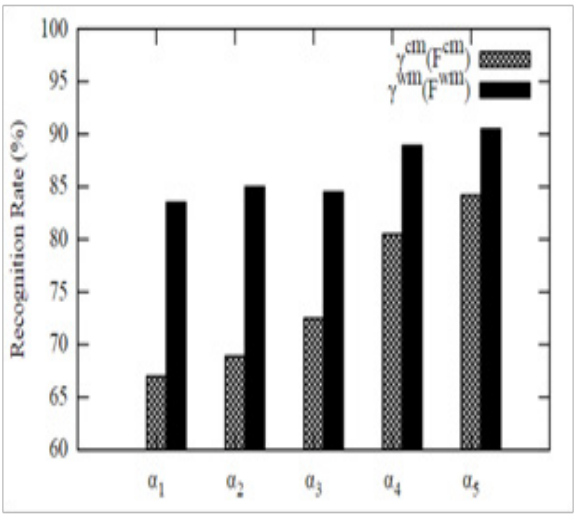


c) $\gamma^{m}(F^{m})$ benchmark VSR system

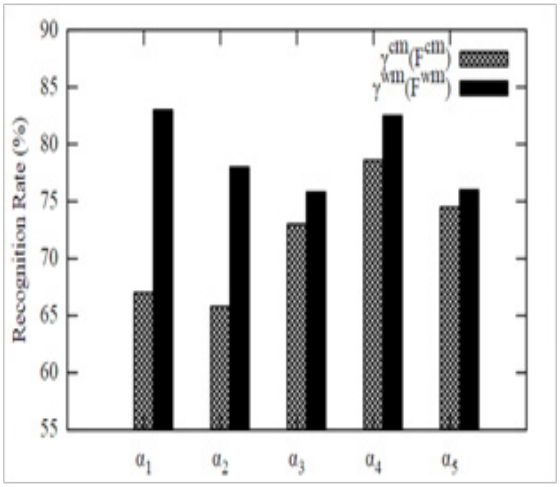
Figure 3. Viseme recognition rates for varying number of Gaussian mixtures with $\gamma^{c}(F^{c})$, $\gamma^{w}(F^{w})$ and $\gamma^{m}(F^{m})$ benchmark systems.



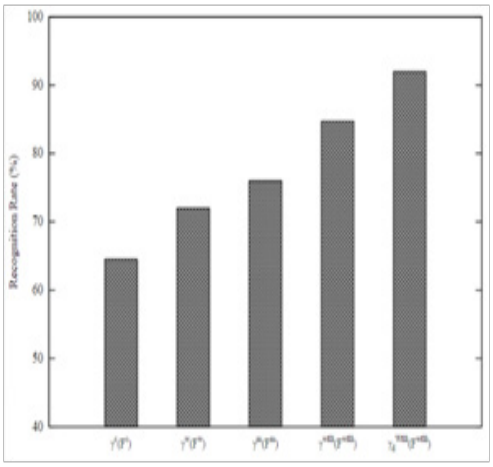
a) Feature level fusion of combined system



b) Weighted decision level fusion of combined system



c) Weighted decision level fusion of combined system



d) Best scores of benchmark and combined system

Figure 4. Viseme recognition rates for combined $\gamma^{cm}(F^{cm})$ and $\gamma^{wm}(F^{wm})$ systems.

4. Conclusion

Since visual cues provide for lesser discriminatory information (than the acoustic cues obtained from audio speech) on the sound units, a small vocabulary of digits is chosen for this study to allow better visual discrimination of sounds. A two-level combined feature framework of pixel and geometric features is proposed to improve visual speech recognition. The viseme model is built using a L-to-R Gaussian HMM model for feature level fusion and a two stream L-to-R Gaussian HMM is used to build a model for decision level fusion. The experimental analysis shows that combined pixel and geometric features significantly improve the performance of the VSR system. The VSR system has the least recognition performance of 64% for DCT features. The performance of the VSR system is 72% and 76% for DWT and ACM features, respectively. VSR system performs significantly better while using the feature-level fusion (84.7%) and decision level fusion (92%). This improvement in performance due to the fusion shows the presence of complementary cues in the pixel based and geometric based features, and also that geometric cues provide better discrimination of visemes.

5. References

1. Chandramohan D, Silsbee PL.A multiple deformable template approach for visual speech recognition, Fourth International Conference on Spoken Language, ICSLP96. Proceedings, USA. 1996; 1. p. 50–3.
2. Alizadeh S, Boostani R, Asadpour V. Lip feature extraction and reduction for HMM-based visual speech recognition systems. 9th International Conference on Signal Processing, ICSP Iran. 2008. p. 561–64.
3. Hong X, Yao H, Wan Y, Chen R. A PCA based visual DCT feature extraction method for lip-reading. International Conference on Intelligent Information Hiding and Multimedia Signal Processing, China. 2006. p. 321–26.
4. Wang X, Hao Y, Fu D, Yuan Y, Chunwei C. ROI processing for visual features extraction in lip-reading. International Conference on Neural Networks and Signal Processing, china. 2008. p. 178–81.
5. Jun H, Hua Z. Research on visual speech feature extraction. International Conference onComputer Engineering and Technology, ICCET Nanchang. 2009. p. 499–502.
6. Matthews I, Cootes TF, Bangham JA, Cox S, Harvey R. Extraction of visual features for lipreading. IEEE Transactions on Pattern Analysis And Machine Intelligence. 2002; 24(9):198–213.
7. Viola P, Jones M. Robust Real Time Object Detection. International Journal of Computer Vision. 2001; 1–30.
8. Brunelli R, Poggio T. Face recognition: features versus templates. IEEE Transactions Pattern Analysis and Machine Intelligence. 1993; 15(10):1042–052.
9. Kass M, Witkin A, Terzopoulos D. Snakes: active contour models, International Journal Computer Vision. 1988; 1:321–31.
10. Bakos B, Marian M. Active contours and their utilization at image segmentation, 5th Slovakian-Hungarian Joint symposium on applied machine intelligence and informatics, Poprad, Slovakia. 2007; 1–5.
11. Cootes TF, Taylor CJ, Cooper DH, Graham J. Active shape models—their training and application. Computer Vision and Image Understanding. 1995; 61(1):38–59.
12. Cootes C, Tim ER, Baldock B, Graham J. An introduction to active shape models image processing and analysis. 2008; 223–4.
13. Yau Y, Chee W, Kumar DK, Weghorn H. Visual speech recognition using motion features and hidden Markov models. International Conference on Computer Analysis of Images and Patterns. Springer Berlin Heidelberg. 2007; 4673:832–39.
14. Brahme A, Bhadade U. Lip detection and lip geometric feature extraction using constrained local model for spoken language identification using visual speech recognition. Indian Journal of Science and Technology (Indjst). 2016; 9(32):1–7.
15. Potamianos P, Gerasimos G. Recent advances in the automatic recognition of audiovisual speech. Proceedings of the IEEE. 2003; 91(9):1306–26.