Audio Visual Isolated Oriya Digit Recognition Using HMM and DWT

Astik Biswas
Department of Electrical Engineering, NIT Rourkela,Orrisa
India
email:-astikbiswas@live.com

P.K. Sahu
Department of Electrical Engineering, NIT Rourkela,Orrisa
India
email:- pksahu@nitrkl.ac.in

Anirban Bhowmick
Department of ECE, IMSEC,Ghaziabad,UP
India
email:- anirban.bhowmick15@rediffmail.com

Mahesh Chandra
Department of ECE, BIT Mesra,Ranchi,Jharkhand
India
email:- shrotriya@bitmesra.ac.in

Abstract

Abstract. Automatic Speech Recognition (ASR) system performs well under restricted conditions but the performance degrades under noisy environment. Audio-visual features play an important role in ASR systems in presence of noise. In this paper Oriya isolated digit recognition system is designed using audio visual features. The visual features of the lip region integrated with audio features to get better recognition performance under noisy environments. Color Intensity and Pseudo Hue methods have been used for lip localization approach with Hidden Markov Model (HMM) as a classifier. For image compression principal component analysis technique has been utilized.

Keywords: AVSR, MFCC, DWT, DCT, HMM.

1. Introduction

The variety of applications of automatic speech recognition (ASR) systems for human computer interfaces, telephony, and robotics has driven the research of a large scientific community in recent decades. However, the success of the currently available ASR systems is restricted to relatively controlled environments and well-defined applications such as dictation or small to medium vocabulary voice-based control commands (e.g., hand-free dialing). In recent years, together with the investigation of several acoustic noise reduction techniques, the study of visual features has emerged as attractive solution to speech recognition under less constrained environments. The use of visual features in audio-visual speech recognition (AVSR) is motivated by the speech formation mechanism and the natural ability of humans to reduce audio ambiguity using visual cues [1]. Traditional audio-based ASR systems perform reasonably well in controlled lab environments. In many environments, however, such as offices or outdoors, the recognition performance decreases drastically due to background noise. Hence, to facilitate the better interpretation of the spoken words,

researchers needed to resort to consider some other distinguishable element of speech.

The importance of visual features for speech recognition, especially under noisy environments, has been demonstrated by the success of recent AVSR systems [2]. One way to increase robustness with respect to acoustic signal distortion is to consider the visual speech modality jointly with the auditory modality. It has been studied that in both ASR and human speech perception, the audio and visual sensory modalities have different strengths and weaknesses, and in fact to a large extent they complement each other. For those distinguishable feature sometime visible speech is usually most informative than only audio. On the other hand, voicing information which is difficult to see visually is relatively easy to resolve via sound. Thus visible speech is not redundant with auditory speech to a large degree.

Hence in this paper information contents of both audio and video are used to train the system. In this paper, we describe a set of methods for visual feature selection and focus is on hidden Markov models for isolated Oriya digit audio-visual speech recognition. The experiments are conducted on clean as well as on noisy data for Oriya speech recognition. Matlab was used to extract wavelet based features and MFCC features. HMM [3] have been used to classify the digits to their respective classes. For extracting visual features wavelets transform (DWT) [4] and discrete cosine transform (DCT) [1] have been used. For image compression principal component analysis (PCA) (Shlens:http://www.snl.salk.edu/~shlens/pub/notes/pca.pdf) method has been used.

A speech recognition system has three major components, database preparation, feature extraction and classification. The recognition performance depends on the quality of database, performance of the feature extraction and classification techniques. Thus choice of features and its extraction from the speech signal should be such that it gives high recognition performance with reasonable amount of computation. The experimental setup used for performing this experiment is shown in Figure 1.

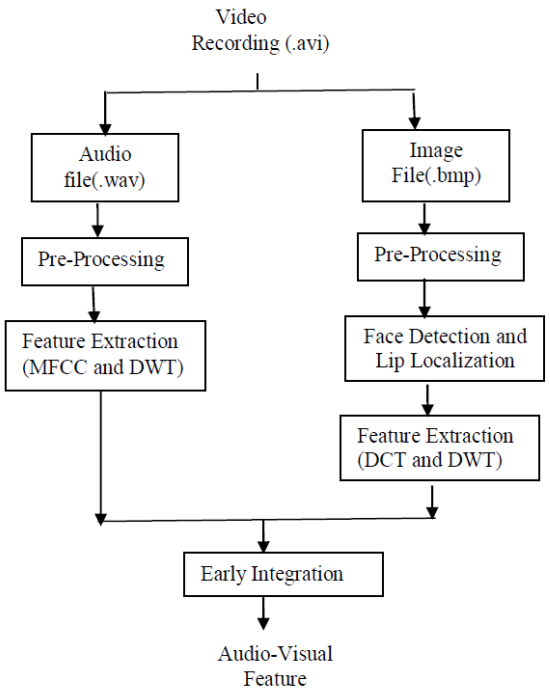


Fig. 1. AVSR System

2. Database Preparation

For this work the database has been prepared by twenty different speakers, belonging to the age group of 20 to 32 years. There are 10 male and 5 female speakers with considerably different speaking-speeds. The background conditions are approximately same for all the speakers. Every speaker has spoken ten Oriya digits ('Shoonya', 'Ek', 'Dui', 'Tini', 'Chaari', 'Paanch', 'Chaw' 'Saath', 'Aatha' and 'Nou') for four times. So there is total of 60 samples of each Oriya digit. The video recording was performed by a digital camera (SONY DSC-H100) having a frame rate of 30 frames per second. The video files are broken into audio only and video only files.

The audio file is stored in uncompressed PCM, Microsoft Wave file format (.WAV) and the video in uncompressed Audio/Video Interleaved file format (.AVI). The two said formats are common and also the readable file formats in MATLAB, the major tool used for further processing of the signals. Out of 60 samples of each digit we took 50 samples for training and rest 10 for testing.

All the audio .wav files of different speakers are resampled to a sampling frequency of 16 kHz. 16 kHz sampling frequency is considered to be a standard for speech processing operations Moreover, the stereo channels are redundant for our desired processing of the signals, one channel is therefore rejected and the resulting audio file with only one channel is stored (mono). Each video instance of each uttered digit was broken into 10 or 15 frames. Since, in a video file, this information is in the form of frames, we converted it into separate image files (.BMP) by writing another program. The standard resolution (1024 * 768) is maintained throughout the experiment.

3. Lip Tracking Method

Since the video recordings contained only the frontal faces of the speakers. But the clutter around the face, including hair, neck, ears, beard and clothes causes

difficulty in adopting a common lip-detection algorithm for all speakers. Hence we used a robust multiple-face-detection program to first determine a rough estimate of the location of the face before going ahead with the actual detection of the lip region. The technique when applied on an image returns a square face boundary box. This algorithm works on gray-scale images alone. So, we first converted the input color image which was in RGB format, with color intensity values ranging from 0 to 255, into its gray-scale intensity equivalent and then applied the procedure.

Fig 2 shows the detected face region and lip region of the speaker. The algorithm being designed for multiple faces, sometimes detected non-face objects as well. Hence, we altered the algorithm to serve our purpose by selecting that detected object as the desired face, which had the largest dimensions.

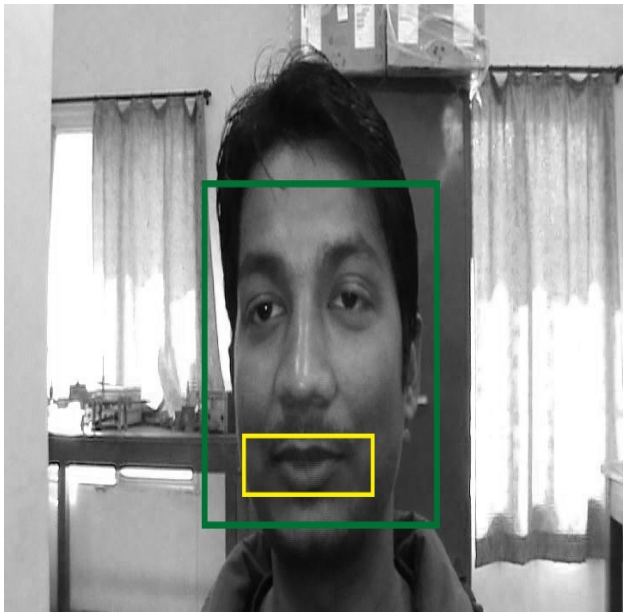


Fig. 2. Detected face and lip region

3.1. *Lip-Localization Method*

For this experiment our ROI is lip region. So to discriminate lip region from other skin region of face lip-localization [2] is necessary. After lip localization binarization process takes place followed by counting of lip-color pixels and plotting of spatial histogram to detect the exact periphery of the lips in order to facilitate in the feature-extraction process. For this, we use a slightly modified procedure as that used prior to face detection. Here, we confined our ROI to the lower one-third of the face due to the absence of any universal pattern among the different vertical histograms. Two lip localization approaches have been applied here like Color-Intensity Mapping and Pseudo-Hue Method.

Color-Intensity Mapping: This method [5] is based on a new color mapping of the lips by integrating color and intensity information. A linear transformation of RGB components has been performed by them. PCA has been employed to estimate the optimum coefficients of transformation. From a set of training images, N pixels of lip and non-lip have been sampled and its distribution is shown. Each pixel is regarded as a 3 dimensional vector $x_i = (R_i, G_i, B_i)$. The covariance matrix is obtained from the three dimensional vector and the associated eigenvectors and Eigen values are determined from the covariance matrix. v = (v1, v2, v3) is an eigenvector corresponding to the third smallest Eigen value where lip and non-lip pixels are the least overlapped.

Experimentally, they have obtained v1=0.2, v2=−0.6, v3=0.3. Thus, a new color space, C is defined as

$$C = 0.2\times R - 0.6\times G + 0.3\times B$$

The new color space C is normalized as

$$C_{norm} = (C - C_{min}) / (C_{max} - C_{min})$$

After normalization, the lip region shows higher value than the non-lip region. By squaring the norm C, the dissimilarity has been further increased between these two clusters as shown. After the color transformation, the C squared image may still show low contrast in the upper lip region. This problem can be resolved by using the intensity information I. The upper lip region typically consists of lower intensity values. So by combining the C squared image (which is well separable in the lower lip) and intensity image (which has a stronger boundary in the upper lip), an enhanced version of the lip color map C can be obtained as follows:

$$C_{map} = \alpha\, C_{squared} + \gamma\, (1/I) \qquad \text{where } \alpha + \gamma = 1$$

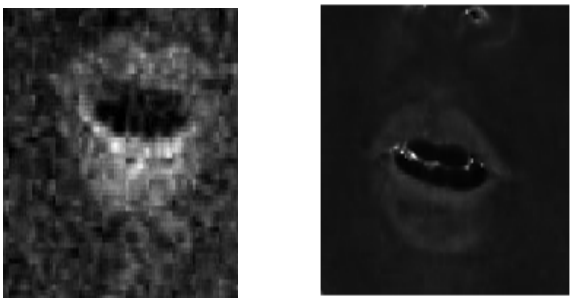
Empirically, $\alpha = 0.75$, $\gamma = 0.25$ are derived. Higher weight is given to the C squared image since it captures most of the lip shape except the upper part and corners of lips.

***Pseudo-Hue Approach*:** Although a number of researchers [6, 7] have suggested that the skin hue is fairly consistent across different people, the colors of the lip and skin region usually overlap considerably. This approach takes into account the R/(R+G) ratio, which is known as the Pseudo Hue value, of the lips.

$$\dot{H} = \frac{R}{R+G}$$

The concept of Pseudo Hue tends to enhance the fact that the difference between R and G for the lip pixels is

greater than that for the skin pixels. In [6] they proposed to use the pseudo hue plane to separate the lips from the skin. Indeed, H has higher values for the lips than for the skin and is robust for lips and skin pixels discrimination even when dealing with different subjects. However, the H of beard and shadow can be very similar as that of the lip. The reason lies in that the pseudo hue value may be very high when all components of RGB are low. Figure 3 shows examples of lip localization method.



(a) (b)

Figure 3 (a) Color Intensity Mapping (b) Pseudo Hue Method.

3.2 Binarization

This process is done to further enhance the separability between the lip and non-lip regions. For this, an optimum threshold is to be first calculated such that the location of lips can be resolved as accurately as possible. The RGB true color image is, in the most basic sense, a three dimensional matrix of 8-bit values, ranging from 0 to 255. Since it is not possible to calculate a threshold for the 3-dimensional image, we need to first convert it into the corresponding 1-dimensional, 8-bit intensity image. This is done using the 'rgb2gray' command of MATLAB. We then developed a few methods to experimentally compute the most appropriate value of threshold for this gray-scale image.

Here pure trial-and-error [1] method has been adopted. We manually tried out different values of thresholds for the images obtained from the two lip-localization approaches. We ran a program that gradually incremented the threshold value from 0 to 255. That value was chosen as the most favourable threshold that fulfilled our goal most; that is, which provided maximum discrimination of the lips from the remaining clutter. We took different threshold value for different speaker to binarize the lip images. Some of the received images are shown in Fig 4.

Fig 4 Binarized Image

4. Experimental Setup & Result

The process of extraction of audio visual for Oriya speech is shown in Figure 1. Audio features of all 15 speakers four all ten digits have been extracted using MFCC [7] and DWT techniques whereas the visual features for all 15 speakers have been extracted using DCT and DWT. The thirteen audio features and fifteen visual features have been integrated to make a composite feature vector of size twenty eight as shown in figure 1. In this experiment there are 10 HMM [8, 9] models trained to recognize all 10 Oriya digits. Each model had five states and three multivariate Gaussian mixtures with a full covariance matrix. Only self-transitions and transitions to the next state are allowed. In the training phase, every training utterance was segmented into equal lengths and then initial model parameters are estimated. Next, the Viterbi decoding algorithm was used to determine an optimum state sequence of each training token and based on this every token was re-segmented. Model parameters are re-estimated repeatedly until the estimates are unchanged or the maximum number of iterations was reached. As mentioned earlier out of 60 samples of each digit, 50 samples used for training and rest 10 used for testing. The experiments are performed on both clean as well as noisy database. Noisy database was prepared using babble, car and F16 noise at 5dB and 10dB SNR levels . Specifically lower SNR levels are taken to see influence of noise on audio visual feature. HMM have been used as classifier for all the recognition experiment. For video features extraction different lip localization methods have been used. Table 1 shows recognition performance of audio only and audio visual features using color intensity mapping (lip localization) and HMM. Table 2 shows the recognition performance of audio only and audio visual features using pseudo hue method (lip localization) with HMM.

It is also observed that pseudo hue approach is providing better result than color intensity mapping. This happens because pseudo hue approach is separating lip with skin colors on the basis difference between red and green color. It is also observed that audio visual features are giving better recognition

efficiency compared to audio features specially in case of increase background noise.

Table 1. Recognition performance of audio visual features using Color Intensity Mapping

Noise Level		Audio only		Audio + Video			
		MFCC	DWT	MFCC+DCT	MFCC+DWT	DWT+DCT	DWT+DWT
Clean Data(60 dB)		92	98	75	74	79	83
Babble	5dB	48	49	54	56	62	69
	10dB	56	58	64	67	66	74
F16	5dB	38	43	41	36	41	46
	10dB	43	49	45	47	52	52
Car	5dB	33	37	42	47	46	57
	10dB	44	46	51	53	55	60

Table 2. Recognition performance of audio visual features using Color Intensity Mapping

Noise Level		Audio only		Audio + Video			
		MFCC	DWT	MFCC+DCT	MFCC+DWT	DWT+DCT	DWT+DWT
Clean Data(60 dB)		95	97	76	779	76	83
Babble	5dB	59	62	67	72	70	74
	10dB	63	67	73	75	76	79
F16	5dB	54	59	63	68	70	72
	10dB	67	69	73	75	77	76
Car	5dB	61	59	68	70	69	75
	10dB	68	73	70	74	73	77

5. Conclusion

In this paper, the recognition performance of isolated Oriya digits has been evaluated using DWT features in clean and noisy environments and compared with existing MFCC feature. Recognition efficiency for clean data as well as noisy data at different SNRs has been tested with different features. DWT shows better recognition performance compared to MFCC because it is more invariant to fixed spectral distortion and channel noise. The discrete wavelet transform has been utilized to produce wavelet coefficients which are used for classification. By using wavelets we are able to extract more temporal information of the speech signal, especially for stops or plosives. Sometimes stops or plosives may be of 5-10 ms duration, whose temporal information are extracted in better way with wavelet based features as compared to other features Two methods for lip localization have been applied for our case and necessary modifications are done. Audio-only and the audio-video-integrated recognizers are tested for increased background noise in the audio. As expected, the audio-only recognizer's performance goes down with the decreasing value of SNR while the video only recognizer's performance remains unaltered. Hence the audio video integrated recognizer proves better for speech recognition applications in presence of noise.

6. References

1. Mumtaz, S.M. and Khanam, R,: *Hindi Viseme Recognition from Continuous Speech,* Project Report, AMU, Aligarh, India. (2010).
2. Estellers, V. and Thiran, J.P, : *Multi-pose lipreading and audio-visual speech recognition, EURASIP Journal on Advances in Signal Processing,*Vol.2012 ,Issue.51. pp.1-23 (2012).
3. Tian, Y., Zhou, J.L., Lin, H., and Jiang, H, : *Tree-Based Covariance Modeling of Hidden Markov Models, IEEE transactions on Audio, Speech and Language Processing*, Vol. 14,Issue. 6, pp. 2134-2147. (2006)
4. Siafarikas, M., Mporas, I., Ganchev, T. and Fakotakis, N,: *Speech Recognition using Wavelet Packet Features' ,Journal of Wavelet Theory and Applications*, Vol .**2** ,pp. 41–59.(2008)
5. Ooi, W.C., Jeon, C., Kim, K., Ko, H. and Han, D.K.,: Effective Lip Localization and Tracking for Achieving Multimodal Speech Recognition, *Multi-sensor Fusion and Integration for Intelligent Systems*, Vol. 35, pp.33-43. (2009)
6. Eveno, N., Caplier , A. and Coulon, P.Y,: *Accurate and Quasi-Automatic Lip Tracking, IEEE Transactions on Circuit and Systems for Video Technology*, Vol.14,Issue.5, pp.706 –715.(2004)
7. Davis, S. B., and Mermelstein, P. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech and Signal Processing, (ASSP), vol. 28, no. 4, pp. 357–366,(1980).
8. Rabiner, L. R. and Juang, B. H.: *Fundamental of Speech Recognition,* Prentice Hall, USA. (1993)
9. Young, S.J. and Woodland, P.C.: The use of state tying in continuous speech recognition., *In: 3rd European Conference on Speech Communication and Technology (EUROSPEECH 93)*, Berlin, Germany pp. 2203-2206. (1993)

[a] Due to page limit constraint, we skipped the feature extraction procedure MFCC and DWT