Proc. Natl. Acad. Sci., India, Sect. A Phys. Sci. (January–March 2022) 92(1):103–110
https://doi.org/10.1007/s40010-020-00724-7

RESEARCH ARTICLE

Recognition of Isolated Digit Using Random Forest for Audio-Visual Speech Recognition

Prashant Borde[1] · **Sadanand Kulkarni**[1] · **Bharti Gawali**[1] · **Pravin Yannawar**[1]

Received: 17 January 2017 / Revised: 9 May 2019 / Accepted: 29 October 2020 / Published online: 13 November 2020

Abstract The proposed research work clearly investigates the effective use of two modalities (audio and visual inputs) toward designing functional audio-visual speech recognition system. The promising results presented in this piece of work were obtained on vVISWa (visual Vocabulary of Isolated Standard Words) dataset of audio-visual words and CUAVE (Clemson University Audio-Visual Experiments) database, respectively. The discrete cosine transform (DCT), local binary pattern (LBP) features of full frontal visual profile and MFCC features for acoustics signals were fused together for recognition purpose and were classified using random forest classifier.

Keywords Face detection · Lip tracking · Local binary pattern (LBP) · Discrete cosine transform (DCT) · Mel-frequency cepstral coefficients (MFCC) · Linear discriminant analysis (LDA) · Random forest

✉ Prashant Borde
borde.prashantkumar@gmail.com

Sadanand Kulkarni
sankalpsadanand.georai@gmail.com

Bharti Gawali
drbhartirokade@gmail.com

Pravin Yannawar
pravinyannawar@gmail.com

[1] Vision and Intelligent System Laboratory, Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India

1 Introduction

Over a last few decades, automatic speech recognition (ASR) has enhanced human computer interaction with high-level reliability. The performance of many automatic speech recognition (ASR) system was reported low, when the acoustic signal is corrupted with noise [1]. The major challenge faced by ASR research community is to improve robustness of traditional ASR in face of audible noise. As the visual modality is not directly affected by audio noise, it can stand as potential source to make ASR systems more robust and to be transformed into AVSR (audio-visual speech recognition system). Lip reading is the technique to recognize what a person is saying by visually interpreting the movements of the lips, face, and tongue. The hearing-impaired or listeners with normal hearing use visual information of lip movements as a primary source of speech perception [2]. These approaches have been adopted to improve the performance of AVSR system in presence of noise [3, 4].

Gurbuz et al. [5] have described the incorporation of visual lip tracking and lip-reading algorithm that utilizes the affine invariant Fourier descriptors from parametric lip contours to improve the audio-visual speech recognition system. Saenko et al. [6] have discussed the approach for visual speech modeling based on articulatory features under visually challenging conditions. This idea was used to set stage for parallel support vector machine (SVM) classifier to extract different articulatory attributes from the input images and then combine their decisions to obtain higher-level units, such as Visemes or words. They evaluated their approach in preliminary experiments on a small audio-visual database, using several image noise conditions, and compared it to the standard Visemes-based modeling approach. Sagheer et al. [7] have presented a

visual speech feature representation approach that combines hyper column model (HCM) with hidden markov model (HMM) to perform a complete lip reading system. Hong et al. [8] has presented a PCA-based method to reduce the dimensionality of DCT coefficients for visual only lip reading systems. The redundancy in the visual cues in audio-visual speech recognition have been examined by Yannawar et al. [9]. Borde et al. [10] have discussed the contribution of visual features that are computed through Zernike moments in association with MFCC for recognition of isolated words. Varpe et al. [11] have discussed isolation of region of interest (ROI) for multi-pose AVSR and designed mechanism of ROI detection based on skin color and also compared its effect with Viola–Jones algorithm under multi-pose AVSR scenario. Morade and Patnaik [12] have discussed the lip tracking using active contour model and proposed a geometrical feature extraction approach for lip reading; these features were classified using 3 state and 5 state HMM and tested over digit. Noda et al. [13] have proposed a noise robust AV ASR system by utilizing two different models to extract noise robust features from audio and video. They employed a deep denoising auto encoder and a convolutional neural network (CNN) encoder to represent AV features, respectively.

In this paper, we present a method for visual feature fusion in audio-visual speech recognition system, and it is based on the extraction of visual features using DCT, LBP and acoustic features using MFCC method.

2 *'vISWa'* Database

Some robust audio-visual speech recognition system database have been designed and databases includes CUAVE [14], AVICAR [15] and WAPUSK [16]. CUAVE [14] database is available free of cost for the researchers to use it for their study. The database entitled vVISWa (Visual Vocabulary of Independent Standard Words) was developed by Borde et al. [17] to deal with multi-pose audio-visual speech recognition system for three languages that is, Marathi (The Native language of Maharashtra), Hindi (National Language of India) and English (Universal language). The vVISWa dataset is consisting isolated words like digits, months, days, most frequent isolated words used in daily interactions in English, Hindi and Marathi. The set of isolated digits uttered by native speakers in English language were considered for evaluations. This dataset consists of 15 individual speakers out of which 8 subjects were male and 7 were female with 10 random utterances of digits. Each speaker uttered words in close-open-close constraint without the head movement. The database comprised of 1500 utterance ($15 \times 10 \times 10$) of these independent standard words in three channels (full frontal, 45° and side pose) so volume is ($4500 = 15 \times 10 \times 10 \times 3$). The stream for channel 1 (C1-Full frontal) of isolated digits uttered in English language is considered in this paper for evaluation. The framework of data acquisition is shown in Fig. 1.

The visual utterances are recorded at the resolution of 720×576 in (*.avi) format using high-definition digital camera in three angles comprising full frontal using camera C1, 45° face using camera C2 and side pose using camera C3. Lighting is controlled under a dark gray background. This database are utilized for building applications based on multi-pose AVSR in Indian regional language.

3 Proposed AVSR System

The proposed audio-visual speech recognition system takes the data stored in the form of vVISWa database. The full frontal visual profiles are considered, and the visual isolated words are passed to the sampler to generate audio stream as well as visual stream suitable for processing from the composite visual stream.

The sampling rate for visual stream contained in vVISWa dataset was 25 fps (frames per second) and accordingly each visual utterance was converted into discrete frames and passed for ROI processing and feature extraction as shown in Fig. 2.

3.1 Face and ROI Identification

The visual speech recognition part is divided into three parts that is face detection, mouth localization and visual feature extraction. In order to achieve robust face detection, Viola–Jones algorithm has been used and detects the face using detector based on AdaBoost classifier that uses cascades of weak classifiers to boost [18]. At first, the detector detects the face from each frame of the visual stream and subsequently extracts the mouth portion that is ROI from frame. Finally, a region-of-interest (ROI) is resized to

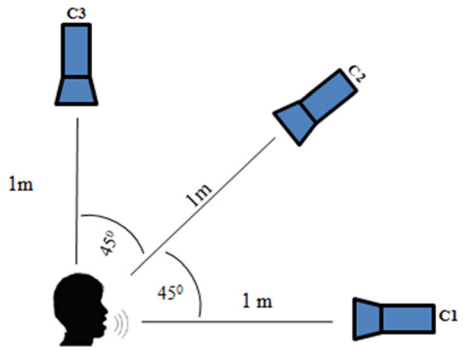


Fig. 1 Acquisition of utterances

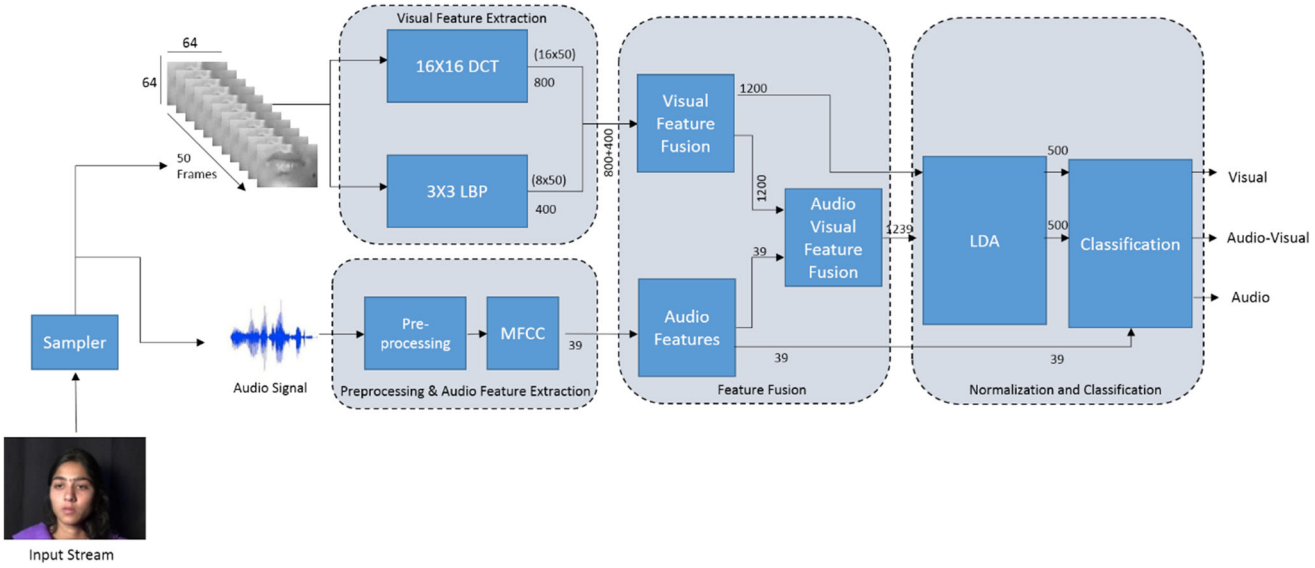


Fig. 2 Architecture of AVSR system

64 × 64 size as shown in Fig. 3. This resized ROI is preprocessed for color space conversion from RGB to gray and passed for visual feature extraction which aims to obtain discriminative visual features of ROI.

3.2 Visual Feature Extraction

The preprocessed visual stream containing ROI is processed for visual feature extraction which is divided in to appearance based and shape based. The appearance-based visual feature extraction treats entire ROI as an information for visual speech recognition and is applied using discrete cosine transform (DCT), local binary patterns (LBP) and linear discriminant analysis (LDA). Shape-based visual features are modeled with a parametric or statistical lip contour model like active contour model (ACM), active shape model (ASM) and active appearance model (AAM). The appearance-based features are computationally efficient, and there is no manual support of hand labeling of data as that of shape based features; therefore, for extracting robust visual features the appearance-based feature have been utilized in this work.

3.2.1 Visual Features Using Local Binary Pattern

LBP is a simple and efficient texture operator which labels the pixels of an image by thresholding neighborhood of each pixel with the value of the center pixel and considers the result as a binary number. It is also invariant to any monotonic gray level changes. Therefore, LBP features are essentially a binary vector that is computed from a neighborhood around the current image pixel. The commonly used neighborhood is 3 × 3 pixels, which are referred as non-interpolated LBP. This is applied for visual feature extraction of each ROI frame such that center of the mask is updated in accordance with neighborhood. The center pixel c contains eight pixels in its neighborhood, $c = p_0, p_1, p_2 \ldots p_7$ and the value of c gets updated in accordance with neighborhood satisfying thresholding criteria. The final LBP feature vector is composed by thresholding the luminance of each p_i against the center pixel c. If the luminance of c is smaller than or equal to the luminance of p_i, the result t_i for p_i is 1, and 0 otherwise. The results t_i are organized as a binary vector t_7, t_6, $t_5 \ldots$ t_0, which is interpreted as an 8-bit unsigned integer value [19] and is calculated for all the samples. Figure 4 shows the process of LBP features extraction.

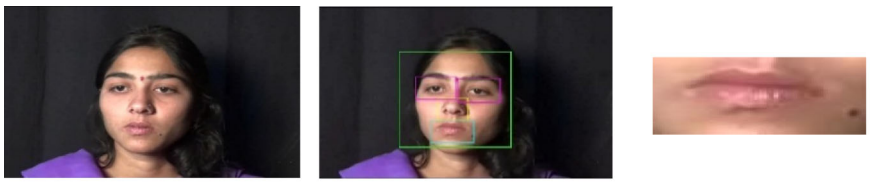


Fig. 3 ROI segmentation using Viola-Jones algorithm. **a** Original frame, **b** facial component-detected frame, **c** mouth localization frame

a Original frame, **b** facial component detected frame **c** mouth localization frame

Fig. 4 Small neighborhood computation using LBP. **a** 3 × 3 neighborhood pixel values and its neighbors values. **b** 3 × 3 neighborhood values greater then center marked as 1 otherwise 0, after thresholding

2	8	9
10	5	3
1	12	4

a 3x3 neighborhood Pixel values and its Neighbors values

0	1	1
1		0
0	1	0

b 3x3 neighborhood values greater then center marked as 1 otherwise 0, After Thresholding

Binary Pattern: 01100101 (101)

The LBP features were stored in the LBP feature vector corresponding to the visual word of vVISWa dataset uttered by the subject. Similarly ROI was also passed for computation of DCT features.

3.2.2 *Visual Features Using Discrete Cosine Transform (DCT)*

DCT has been applied for visual feature extraction of each frame. It describes an image in terms of its frequency components and is used in image reconstruction, filtering and image compression applications. It has three major properties like *decorrelation*, *energy compactness* and *computation effectiveness*. We have used block-based DCT [20]. It divides ROI image of size [64 × 64] into 16 non-overlapping blocks of size [16 × 16], and we have applied DCT to each block to obtain the transform coefficients $T_i^{'}$ corresponding to each block, where i = 1, 2, 3… 16. The energy coefficients of each block $E_{(i)}$ are calculated using

$$E(i) = \sum_{j=1}^{16} \sum_{k=1}^{16} T(j,k) \tag{1}$$

where i = 1, 2, 3, 4…16. Sixteen energy coefficients corresponding to ROI frame are computed and stored in a vector. This has been applied to all ROI frames of visual word sample and target vector representing word energy feature of size 1 × 800 elements as shown in Fig. 5. All such vectors corresponding to the isolated visual words of vVISWa dataset are stored, and DCT energy feature matrix are constituted.

LBP and DCT features are extracted and stored in the feature vector. The features of acoustic signals are also processed.

3.2.3 *Acoustic Feature Extraction Using MFCC*

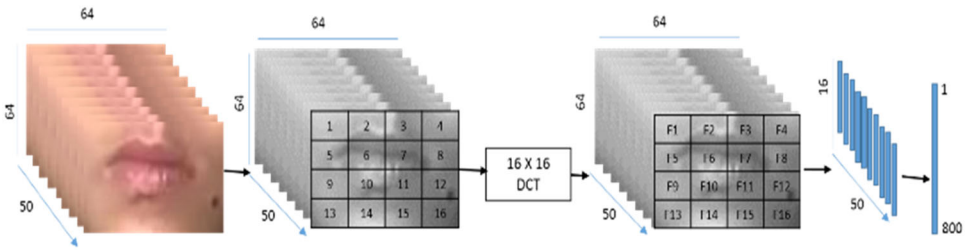
The sampled acoustic signal representing vVISWa data set is preprocessed for acoustic feature extractions. The preprocessing part includes cleaning of signal and removal of silence which exists in the signal before utterance and after utterance. This procedure returns the absolute signal representing only user utterance corresponding to isolated word. The preprocessed acoustic signal is passed for computation of Mel frequency cepstral coefficients (MFCC). Due to its spectral base as parameters for recognition, MFCCs represent audio based on perception of human auditory systems. In MFCC, the frequency bands are positioned logarithmically (i.e. on the Mel scale) which approximates the human auditory system's response more closely than the linearly spaced frequency bands of FFT or DCT [21, 22]. Figure 6 shows the block diagram of MFCC features extraction process.

The sampled speech signal corresponding to the isolated word from *vVISWa* data set is passed for audio feature extraction using MFCC. Thirteen MFCC features are computed along with its first-order and second-order derivative of the MFCC coefficients. Finally, 39 MFCC (base, first order, second order) are extracted for each utterance and stored into feature vector, respectively.

3.3 Fusion of Feature Vector and Normalization

The visual and audio features are extracted and stored corresponding to words of vVISWa dataset which are required to be fused together to generate a robust visual and audio feature to be utilized at the recognition phase. In practice, data fusion is introduced differently at different

Fig. 5 DCT energy feature extraction



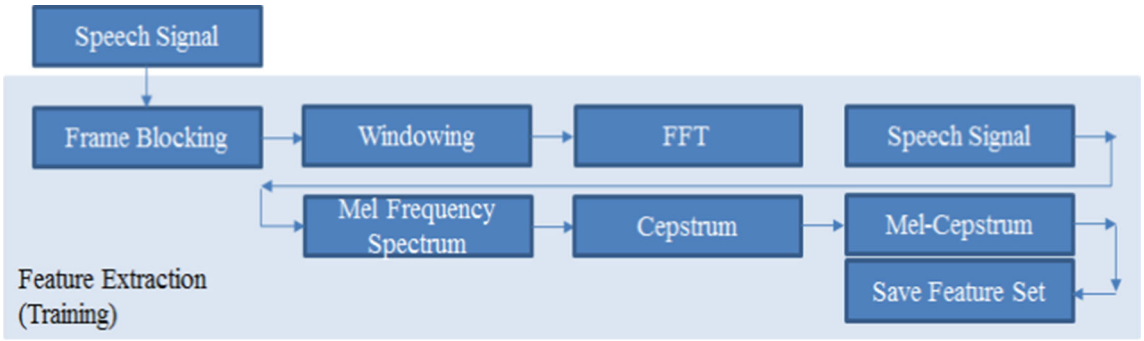


Fig. 6 Block diagram of MFCC feature extraction

level of fusion like sensor level, feature level, matching score level and decision level. The features level fusion is realized by simply concatenating the features obtained from different sources of information [23]. The concatenated features have better discrimination power than the individual feature vectors. We have used the feature level fusion of DCT (800 features) and LBP (400 features), and this results in new fused feature vector of size 1200 features. This feature vector is further normalized as per classes and trained for classification using random forest classifier. Similarly 39 MFCCs are also used for recognition based on audio features only. These 39 audio features and combined visual features (1200 features) are also fused together in audio-visual feature fusion resulting in 1239 features corresponding to one word, and similarly training is made for all words of vVISWa dataset by class-oriented normalization using LDA. The normalized audio-visual features are used for classification.

4 Experiment and Result

The process of visual feature extraction and audio feature extraction was applied on vVISWa and CUAVE dataset, and the performance of the system was studied and compared. The CUAVE database contained continuous and randomly uttered visual sequence of 0–9 words with associated time-stamp information. This timestamp helped in navigation of stream and holds information of start and end time of utterance. CUAVE data set include 37 speaker and was recorded at 720 × 480 resolution of pixels with

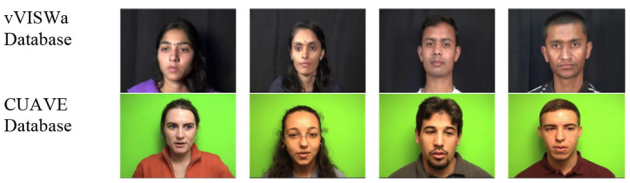


Fig. 7 Samples of *vVISWa* and *CUAVE* database

frame rate 29.97 fps. Figure 7 shows the sample image from CUAVE and vVISWa database. The accuracy of the proposed system was tested on digit utterances from CUAVE and vVISWa dataset.

The utterances from these dataset were processed for feature extraction as discussed above. The DCT visual features corresponding to the word are shown in Table 1; these features are global information representation of frame containing ROI. These global DCT features from all frames corresponding to word utterance are combined into matrix of size 800 (that is 16 DCT features each for 50 frames).

The local information of ROI was computed using LBP as presented in Table 2. LBP features corresponding to all frames contained in words were computed and stored in matrix representing 400 LBP feature corresponding to the word.

Due to the different characteristics of DCT representing global information and LBP representing local information, the features were fused into the fusion matrix. Each word visual feature is represented by 1 × 1200 vector; such n vectors are placed into the fusion matrix. This fusion matrix were normalized as per classes using LDA [24], and total 500 features for training set and test set, respectively, were considered for classification using random forest.

The acoustic feature of utterance was computed using MFCC; 39 MFCC features comprised of 13 primary MFCC, 13 first derivative coefficients and 13 s derivative coefficients, respectively; these coefficients are listed in Table 3.

The DCT + LBP + MFCC features corresponding to word from vVISWa dataset were fused using audio-visual feature fusion and resulting 1 × 1239 feature vector. The final AVF fusion matrix for all training set and test set were normalized using LDA and passed for RF classifier. RF classifier classifies the set based on parameters to adjust the number of trees to grow. In general, the default value of trees offered by the classifier was considered to be a good choice for evaluation of these parameters [25]. For

Table 1 DCT global features of ROI

DCT feature frame	1	2	3	4	5	6	7	…	16
1	458	463	386	337	475	466	344	…	209
2	455	472	386	345	494	478	328	…	207
3	453	478	405	327	473	485	334	…	208
4	469	482	389	322	472	473	333	…	210
5	468	478	387	330	482	461	493	…	210
6	468	469	373	328	473	640	616	…	211
7	473	467	358	318	488	817	590	…	213
…	…	…	…	…	…	…	…	…	…
50	469	449	436	443	446	448	450	…	209

Table 2 LBP local features of ROI

LBP features frame	1	2	3	4	5	6	7	8
1	113	143	638	1734	742	218	198	58
2	91	148	666	1762	711	212	189	65
3	112	110	696	1724	748	212	177	65
4	115	138	670	1684	790	219	171	57
5	111	119	674	1761	716	231	189	43
6	107	132	711	1709	715	236	180	54
7	116	111	696	1792	688	222	173	46
:	:	:	:	:	:	:	:	:
50	122	147	710	1689	696	228	198	54

Table 3 MFCC features for audio samples

MFCC coefficients →	1	2	3	4	5	6	7	…	39
Sample 1	4.136	−1.388	3.226	−0.132	1.437	−0.604	0.458	…	0.202
Sample 2	3.840	−1.607	3.417	−0.314	1.517	−0.630	0.515	…	0.251
Sample 3	4.148	−1.117	3.399	−0.095	1.328	−0.585	0.484	…	0.203
Sample 4	4.134	−0.990	3.383	−0.056	1.271	−0.581	0.491	…	0.215
Sample 5	4.185	−0.961	3.304	−0.031	1.373	−0.555	0.575	…	0.185
Sample 6	4.305	−1.079	3.318	−0.048	1.332	−0.598	0.509	…	0.141
Sample 7	4.223	−0.941	3.361	−0.003	1.316	−0.480	0.527	…	0.169
Sample 8	4.262	−1.071	3.329	−0.010	1.387	−0.501	0.468	…	0.165
Sample 9	4.207	−1.169	3.284	0.042	1.336	−0.527	0.514	…	0.180
Sample 10	3.798	−1.515	3.417	−0.246	1.518	−0.502	0.577	…	0.291
:	:	:	:	:	:	:	:	:	:

vVISWa and CUAVE databases, the classification process were performed using 500 trees. A *random forest* multi-class model was trained, where each class corresponds to a word in the database. The classification recognition rates of the proposed method over aforesaid databases are found to be better than [26]. The experimental results on *vVISWa* and *CUAVE* dataset are shown in Tables 4 and 5.

Table 4 provides information about the performance of recognition without feature normalization. It was seen that when DCT + LBP-fused feature matrix was considered for recognition, the recognition performance of vVISWa dataset on visual features was observed to be degraded as compared to CUAVE database, because number of participating ROI frames involved in the deformation of mouth at the time of utterance were maximum as compare to the CUAVE dataset corresponding to the word. The CUAVE dataset utterance was sampled with 29.97 fps and resulted in 22 frames and utterance duration was <1 s. In

Table 4 Recognition performance on CUAVE and vVISWa dataset without feature normalization

	Method	CUAVE	vVISWa
Visual only	DCT + LBP	77.85	63.92
Audio only	MFCC	87.14	73.21
Audio-visual	DCT + LBP + MFCC	97.14	75.85

Table 5 Recognition performance on CUAVE and vVISWa dataset with feature normalization

	Method	CUAVE	vVISWa
Visual only	DCT + LBP	41.25	83.75
Audio only	MFCC	87.14	73.21
Audio-visual	DCT + LBP + MFCC	76.5	100

case of vVISWa dataset, the sampling rate is 25 fps and resulted in 50 frames and utterance duration was 2 s. As the number of participating frames is maximum, the size of features vector becomes large and thus un-normalized features result in lower classification rate as compared to CUAVE dataset. MFCC features extracted for CUAVE and vVISWa datasets have been computed and classified, and result shows that system performs better on CUAVE database as compared with vVISWa database. Similar results have been seen when the audio-visual features of DCT + LBP + MFCC were fused together and classifier results were in better performance with CUAVE database as compared with vVISWa dataset.

Feature normalization plays crucial role in recognition system. It was seen from the Table 5 that visual recognition of utterance using normalized DCT + LBP features is better in vVISWa dataset as compared to CUAVE dataset as the duration of sample of utterance is 2 s in vVISWa set where as in CUAVE dataset is less than 1 s. The visual deformation of mouth at the time of utterance was clearly represented in ROI if the utterance duration is about 2 s; otherwise, visual deformation is overlapped or improper and will contribute in less recognition rate. In [25], DCT global information features of ROI and LBP that is local information features were computed, and their dimensions were reduced as well as feature was selected using Mutual Information Feature Selector (MIFS) and linear discriminant analysis (LDA). DCT global information of ROI and LBP local information of ROI were fused together into to the feature vector and used for classification by normalization and selection of features using LDA. The result produced by the recognition system on vVISWa dataset outperforms over CUAVE dataset.

The recognition based only on acoustic samples revealed significant improvement in CUAVE dataset as compared with vVISWa dataset. The performance of vVISWa dataset is low due to the presence of silence before and after utterance of the word. The 'CUAVE' dataset contains visual streams of continuous randomly uttered words and was sampled with the help of labeled data (contains three column attribute information like start-time, end-time, word). Labeled data were supportive in selection of acoustic signal so that MFCC features of selected segment were extracted and represented in feature vector. This increases the recognition results of CUAVE acoustic data over vVISWa acoustic data. Table 5 shows that when normalized features of DCT + LBP + MFCC were used for recognition of isolated word from CUAVE and vVISWa dataset, the performance of the system was better in vVISWa dataset, and recognition result was significantly increased as compared with '*CUAVE*' and that of [26].

5 Conclusion

This paper presents the performance of audio-visual speech recognizer using CUAVE and vVISWa database of digits. It is seen from the result that recognizer outperforms over vVISWa as compared with CUAVE dataset when bimodal input was given. The approach of fusion of DCT + LBP features reduces the feature space and minimizes the overheads of managing large feature sets. The combination of DCT + LBP + MFCC features increases the performance of audio-visual recognizer significantly.

Acknowledgements The authors gratefully acknowledge support by the Department of Science and Technology (DST) for providing financial assistance for Major Research Project sanctioned under *Fast Track Scheme for Young Scientist*, vide sanction number SERB/1766/2013/14 and the authorities of Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS) India, for providing the infrastructure for this research work.

References

1. Dupont S, Luettin J (2000) Audio-visual speech modeling for continuous speech recognition. IEEE Trans Multimed 2(3):141–151
2. Železný M, Krňoul Z, Císař P, Matoušek J (2006) Design, implementation and evaluation of the Czech realistic audio-visual speech synthesis. Signal Process 86(12):3657–3673
3. Neti C, Potamianos G, Luettin J, Matthews I, Glotin H, Vergyri D, Sison J, Mashari A (2000) Audio visual speech recognition. No. REP_WORK. IDIAP.
4. Heckmann M, Berthommier F (2002) Kroschel K (2002) Noise adaptive stream weighting in audio-visual speech recognition. EURASIP J Appl Signal Process 1:1260–1273
5. Gurbuz S, Patterson EK, Tufekci Z, Gowdy JN (2001) Lip-reading from parametric lip contours for audio-visual speech recognition. In: Seventh European conference on speech communication and technology, INTERSPEECH, pp 1181–-1184.
6. Saenko K, Darrell T, Glass JR (2004) Articulatory features for robust visual speech recognition. In: Proceedings of the 6th international conference on Multimodal interfaces, ACM, pp 152–158.
7. Sagheer A, Tsuruta N, Taniguchi R-I, Maeda S (2005) Visual speech features representation for automatic lip-reading. In: Proceedings (ICASSP'05). IEEE international conference on acoustics, speech, and signal processing, 2005, vol. 2, p 781.
8. Hong X, Yao H, Wan Y, Chen R (2006) A PCA based visual DCT feature extraction method for lip-reading. In: Proceedings of the 2006 international conference on intelligent information hiding and multimedia, IEEE, pp 321–326.
9. Yannawar PL, Manza GR, Gawali BW, Mehrotra SC (2010) Detection of redundant frame in audio visual speech recognition using low level analysis. In: Proceedings of the 2010 IEEE international conference on computational intelligence and computing research, IEEE, pp 1–5.
10. Borde P, Varpe A, Manza R, Yannawar P (2015) Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition. Int J Speech Technol 18(2):167–175
11. Varpe A, Borde P, Sukale S, Perdeshi P, Yannawar P (2015) Analysis of induced color for automatic detection of ROI in multipose AVSR system. Information systems design and intelligent applications. Springer, New Delhi, pp 525–538
12. Morade SS, Patnaik S (2014) A novel lip reading algorithm by using localized ACM and HMM: tested for digit recognition. optik 125(18): 5181–5186.
13. Noda K, Yamaguchi Y, Nakadai K, Okuno HG, Ogata T (2015) Audio-visual speech recognition using deep learning. Appl Intell 42(4):722–737
14. Patterson EK, Gurbuz S, Tufekci Z (2002) Gowdy JN (2002) Moving-talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus. EURASIP J Appl Signal Process 1:1189–1201
15. Lee B, Hasegawa-Johnson M, Goudeseune C, Kamdar S, Borys S, Liu M, Huang T (2004) AVICAR: audio-visual speech corpus in a car environment. In: Proceedings of the eighth international conference on spoken language processing INTERSPEECH-2004, pp 2489–2492.
16. Vorwerk A, Wang X, Kolossa D, Zeiler S, Orglmeister R (2010) WAPUSK20-a database for robust audiovisual speech recognition. In: LREC, pp 3016–3019.
17. Borde P, Manza R, Gawali B, Yannawar P (2016) vVISWa'-A multilingual multi-pose audio visual database for robust human computer interaction. Int J Comput Appl 137(4):25–31
18. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition, 2001. CVPR 2001, Vol. 1, pp. I-I.
19. Bianconi F, Fernández A (2011) On the occurrence probability of local binary patterns: a theoretical study. J Math Imaging Vis 40(3):259–268
20. Estellers V (2012) Thiran JP (2012) Multi-pose lipreading and audio-visual speech recognition. EURASIP J Adv Signal Process 1:51
21. Goh C, Leon K (2009) Robust computer voice recognition using improved MFCC algorithm. In: Proceedings of the 2009 international conference on new trends in information and service science, IEEE, pp. 835–840.
22. Gold B, Morgan N, Ellis D (2011) Speech and audio signal processing: processing and perception of speech and music. Wiley, New Jersey
23. Kaynak MN, Zhi Q, Cheok AD, Sengupta K, Jian Z, Chung KC (2004) Analysis of lip geometric features for audio-visual speech recognition. IEEE Trans Syst Man Cybern A Syst Hum 34(4):564–570
24. Gravier G, Potamianos G, Neti C (2002) Asynchrony modeling for audio-visual speech recognition. In: Proceedings of the second international conference on human language technology research. Morgan Kaufmann Publishers Inc., Burlington, pp 1–6.
25. Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2(3):18–22
26. Togneri R, Bennamoun M, Sui C (2014) Multimodal speech recognition with the AusTalk 3D audio-visual corpus. Tutorial at Interspeech.