Journal of Scientific & Industrial Research
Vol. 79, May 2020, pp. 383-386

A Robust Feature Extraction with Dual Fusion aided Extreme Learning for Audio–Visual Hindi Speech Recognition

Usha Sharma[1]*, Hari Om[2] and A N Mishra[3]

[1]Indian Institute of Technology (ISM) Dhanbad, India

[2]CSE Department, Indian Institute of Technology (ISM) Dhanbad, India

[3]Krishna Engineering College, Ghaziabad, India

Received 23 June 2019; revised 29 January 2020; accepted 17 March 2020

In Automatic Speech Recognition (ASR) based system implementation, robustness to several noisy background situation is a unique challenge. In this paper, for estimating both audio and visual aspect feature in light of different information representation perspectives directs to the robust feature extraction from audio-visual speech image. Further, the authors obtain the bottleneck features from the bottleneck layer of the bottleneck deep neural network (BN-DNN). Further, a familiar powerful texture descriptor of Local Binary Pattern (LBP) and Local Phase Quantization (LPQ) is applied to obtain the visual related features from the face region. Moreover, the categorization is executed utilizing the help of Extreme Learning Machine (ELM) and to reach the global optimum through Jaya optimization algorithm for audio-visual Hindi speech recognition. The proposed scheme is evaluated in MATLAB platform and the implementation is equated with the existing audio-visual speech recognition (AVSR) approaches.

Keywords: Speech recognition, Audio-visual, Jaya optimization, Bottleneck DNN, ELM

Introduction

The audio-visual speech recognition (AVSR) is a segment that has the capability of solving the problems in the speech processing. The main intention is to identify the expressed word regarding visual signals or data.[1,2] Commonly, the two methods are to identify the noticeable speech like the holistic system and the visemic method.[3] The state-of-the-art techniques show the utilization of diverse languages in AVSR like Hindi, Turkish, Arabic, Portuguese, and English etc.[4–6] An automatic isolated digit recognition system (AIDRS) developed by using Hidden Markov Model (HMM) could recognize spoken digit utterance in English effectively.[7] In Hindi AVSR, the bimodal technique was suggested in[8] to increase the power of the Hindi speech recognition structure. The Hindi phoneme recognition structure was introduced in[9] and which was created through the combination of HMM and Convolutional Neural Network (CNN) by employing the audio and visual characteristics. The Coupled Hidden Markov Model (CHMM) based recognition method was presented by Abdelaziz *et al.*[10] They constructed the complete scheme which allowed the visionless prediction of dynamic stream weights for changing the audio and visual speaking recognition regarding CHMMs. An AVSR incorporating the 3D lip data attained from the Kinect was suggested by Wang *et al*.[11]

Experimentation

In this article, a system based on an effective audio and visual aspect is suggested for Hindi speech recognition, that depends on a new dual fusion assisted strong feature extraction approach with Jaya algorithm enhanced the greatest learning machine oriented categorization for constant audio and visual Hindi speech recognition. Additionally, the authors have obtained the bottleneck audio features from the DNN frankly as well as qualified for the language detection and substituted the whole final hidden layer with a bottleneck layer. Moreover, a low dimensional nonlinear conversion of the input features is represented by the bottleneck features.

In the feature extraction phase, the Local Binary Pattern (LBP) and Local Phase Quantization (LPQ) texture descriptors are utilized for extracting the visual connected features from the standardized visual images. The LBP is one of the strong texture descriptors which is able to extract the flat areas,

*Author for Correspondence
E-mail: ushasharma1529@gmail.com

edges and bright/dark marks on visual images that holds the important information for the visual speech recognition. The LPQ is one of the most potent texture descriptors which depend on the Fourier phase spectrum of the blur invariance property. The illumination-invariance property of the LPQ ensures the significant implementations in the demanding situations of the changing illumination. The DCT features of the high variance, as well as the middle features, are utilized according to considerably participate in handling the feature selection method. The computational charge and the development of the recognition charge are the outcomes of the DCT. The speech recognition work was then carried out by using modeling with the Extreme Learning Machine (ELM) through the Jaya algorithm. The complete process is depicted in Fig 1.

Subsequently, by employing the decision level fusion system under very low SNR, the feature level fusion procedure is accomplished. As per the present SNR situations, the audio and visual feature is merged with the assistance of the adaptive weights that could correspondingly yield a finer overall recognition execution. In the MATLAB platform, the suggested method is executed and its functioning is equated with the current audio and visual AVSR recognition methods.

Results and Discussion

The audio only speech data indicated as D1 and audio and visual dataset signified as D2 includes audio-visual speech have been recorded. The sound from the audio and visual dataset (D2) is separated by employing the standard online video to audio converter, and the new video series is changed into the quantity of casings by MATLAB tool. The common classification plans are to calculate the possibility of the suggested method, for example, ANN, SVM (Support Vector Machine) and HMM. They have employed the dual level integration procedure before arranging the audio and video strategy. Nevertheless, the suggested ELM classifier obtained higher accuracy. In developed ASR modelling system, the bottleneck feature is attained via the BN-DNN. From feature extraction phase, 13D MFCC, 13D SDC and 9D LPC in total, and 35D stacked feature set have been attained. In present recorded dataset, total 1640 training samples from the various speakers have been considered, belonging to the age group of 20–35 years and got total 57400(35x1640) features, which was finally given to DNN for obtaining bottleneck feature.

The BN-DNN model which is utilized in this investigation contains 4 hidden layers. The fourth hidden layer is the bottleneck layer, and the quantity of units in other layers is 1024. The input data is a 35-dimensional stacked feature set.

Feature level Fusion

The research results were attained in viewpoint of Accuracy, F1 Score, Recollection and Precision as for Audio Modality from dataset D1 (AM-D1), Audio Modality from dataset D2 (AM-D2), Video Modality from dataset D2 (VM-D2) and Audio-Video Modality from dataset D2 (AM-AVM-D2) modalities. From the results, it is observed that the precision develops vividly while the audio and the visual modalities are used with one another. The suggested ELM classifier delivers greater precision which is equated with disparate processes.

Classifier level fusion

The investigational outcomes acquired in view of Accuracy, F1 Score, Recall and Correctness regarding Audio Modality from dataset D1 (AM-D1), Audio Modality from dataset D2 (AM-D2), Video Modality

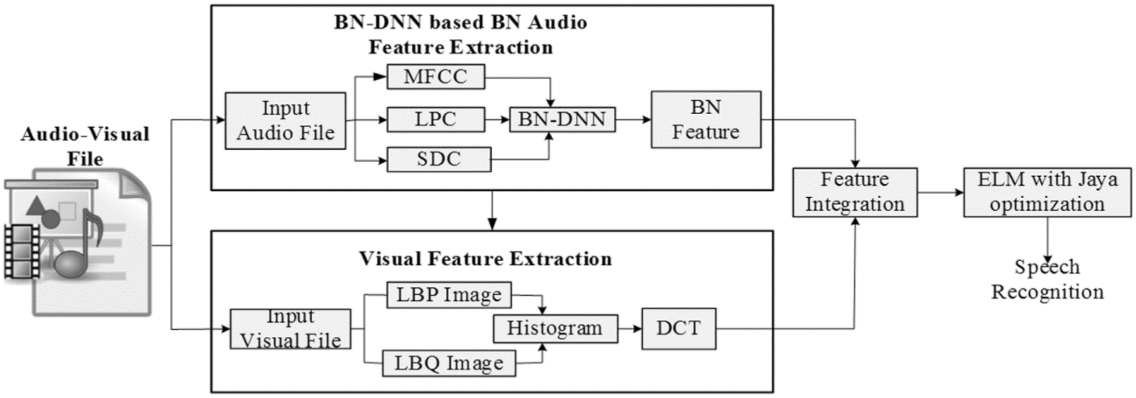


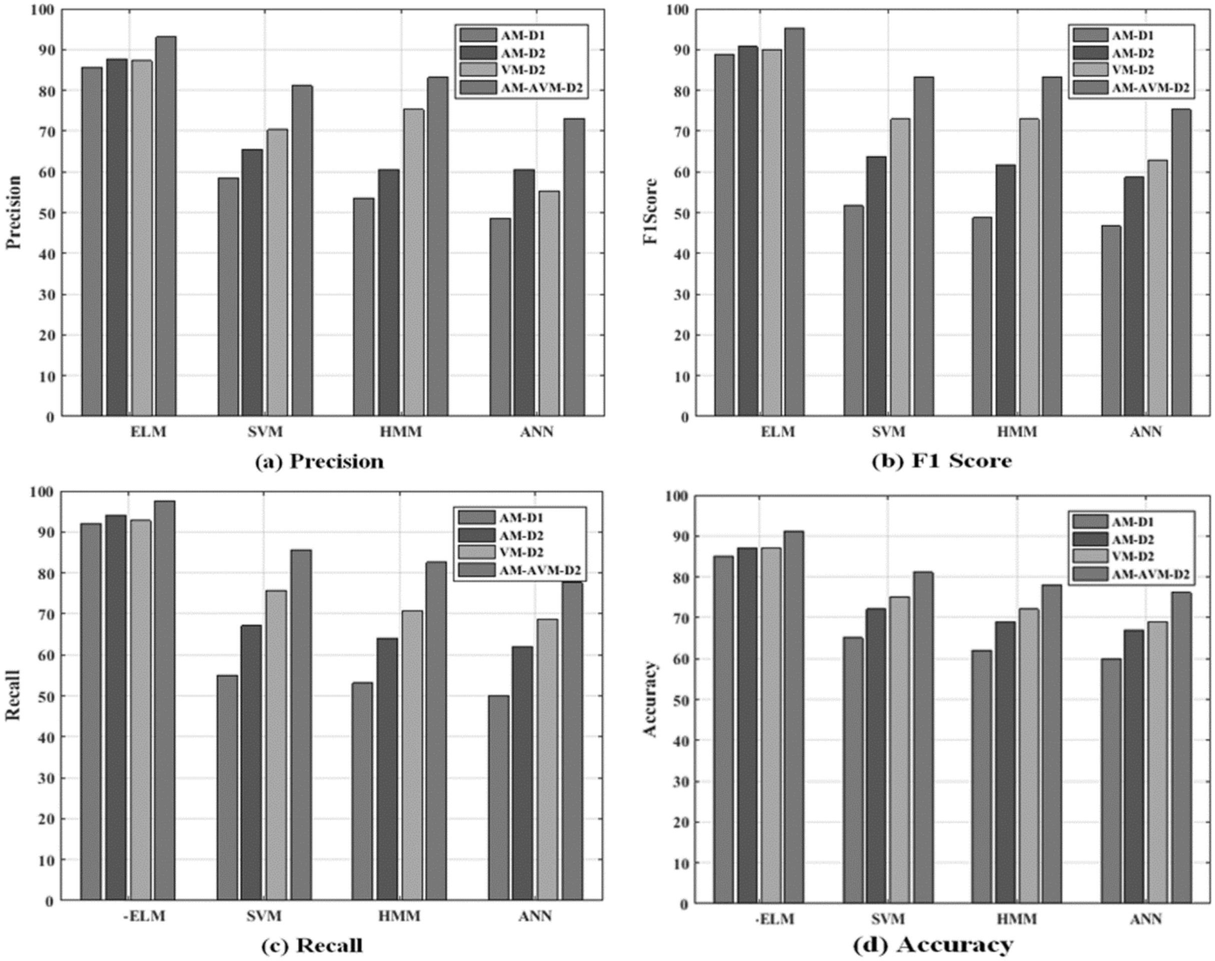
Fig. 1 — Process flow of proposed Hindi audio-visual speech recognition

from dataset D2 (VM-D2), and Audio-Video Modality from dataset D2(AM-AVM-D2) modalities are illustrated in Fig. 2. It is observed that the precision develops severely while the audio and visual modalities are applied with one another. Eventually, the suggested ELM classifier delivers supreme precision that is matched with the disparate structures, for instance, HMM, SVM.

The Babble and White noise are at the disparate SNR levels of 0, 5, 10 and 20 dB are being applied to create the noisy database. The ELM has been employed as the classifier for all the recognition research. The recognition functioning for the clean database is utilized for video characteristics extrication by disparate lip localization techniques. The recognition functioning for Babble noise and White noise is indicated in Table 1.

Performance comparison of different classifiers

The same training and test sets were matched by employing the ELM, ANN, SVM and HMM. The ELM exceeded the ANN that depends on the precision (Table 2). Nevertheless, the minor variations in the accuracy between the ELM and SVM



(a) Precision

(b) F1 Score

(c) Recall

(d) Accuracy

Fig. 2 — Classifier level fusions

Table 1 — Recognition accuracy with babble noise and white noise

SNR levels (dB)	Audio only (D1)		Audio only (D2)		Audio +Video (D2)	
	babble noise	white noise	babble noise	white noise	babble noise	white noise
20	88	87	87	88.5	91	91.5
10	86	86.5	85	87	88	88.5
5	83	85	82.5	86	87.5	87
0	82.1	83	81	82	85	86

Table 2 — Performance of different classifiers

Classifier	Accuracy (%)
Proposed	91
SVM	88
HMM	85
ANN	78

classifiers are observed. The ELM exceeded HMM, ANN and SVM by a huge side depends on the training time (Table 2). The objective is to improve an actual multimodal speech recognition, thus selected the ELM as a classifier that delivered the most elegant execution based on the training time and the precision.

Conclusions

This paper covers the Hindi AVSR for developing automatic speech recognition under different noisy situations. The audio feature extraction task is executed utilizing the different acoustic features namely MFCC, LPC and SDC features. Following conclusions are drawn from the work.

- The bottleneck feature is attained via bottleneck deep neural network (BN-DNN) to discover the speech in an effective way and the suggested Extreme Learning Machine (ELM) classifier yields more accuracy.
- The ELM as a classifier delivered the most elegant execution based on the precision.
- The ELM exceeded HMM, ANN and SVM significantly in terms of accuracy.

References:

1 Anguera X, Bozonnet S, Evans N, Fredouille C, Friedland G & Vinyals O, Speaker diarization: A review of recent research, *IEEE T Audio Speech*, **20(2)** (2012) 356–370.

2 Benzeghiba M, Mori R D, Deroo O, Dupont S, Erbes T, Jouvet D, Fissore L, Laface P, Mertins A, Ris C & Rose R, Automatic speech recognition and speech variability: A review, *Speech commun*, **49(10)** (2007) 763–786.

3 Gurban M & Thiran J P, Information theoretic feature extraction for audio-visual speech recognition, *IEEE T Signal process*, **57(12)** (2009) 4765–4776.

4 Biswas A, Sahu P K & Chandra M, Multiple camera in car audio–visual speech recognition using phonetic and visemic information, *Comput Electr Eng*, **47** (2015) 35–50.

5 Biswas A, Sahu P K & Chandra M, Admissible wavelet packet features based on human inner ear frequency response for Hindi consonant recognition, *Comput Electr Eng*, **40(4)** (2014) 1111–1122.

6 Pandey H M, Jaya a novel optimization algorithm: What, how and why?: 6th IEEE International Conference on In-Cloud System and Big Data Engineering, (2016) 728–730.

7 Nimje K & Shandilya M, Automatic isolated digit recognition system: an approach using HMM, *J Sci Ind Res*, **70** (2011) 270–272.

8 Varshney P, Farooq O & Upadhyaya P, Hindi viseme recognition using subspace DCT features, *Int J Appl Pattern Recog*, **1(3)** (2014) 257–272.

9 Noda K, Yamaguchi Y, Nakadai K, Okuno H G & Ogata T, Audio-visual speech recognition using deep learning, *Appl Intell*, **42(4)** (2015) 722–737.

10 Abdelaziz A H, Zeiler S & Kolossa D, Learning dynamic stream weights for coupled-HMM-based audio-visual speech recognition, *IEEE T Audio Speech*, **23(5)** (2015) 863–876.

11 Wang J, Zhang J, Honda K, Wei J & Dang J, Audio-visual speech recognition integrating 3D lip information obtained from the Kinect, *Multimedia Syst*, **22(3)** (2016) 315–323.