Audio-Visual Automatic Speech Recognition Using PZM, MFCC and Statistical Analysis

Saswati Debnath, Pinki Roy *

National Institute of Technology, Silchar, Assam (India)

Received 3 March 2020 | Accepted 15 October 2020 | Early Access 1 September 2021

Abstract

Audio-Visual Automatic Speech Recognition (AV-ASR) has become the most promising research area when the audio signal gets corrupted by noise. The main objective of this paper is to select the important and discriminative audio and visual speech features to recognize audio-visual speech. This paper proposes Pseudo Zernike Moment (PZM) and feature selection method for audio-visual speech recognition. Visual information is captured from the lip contour and computes the moments for lip reading. We have extracted 19th order of Mel Frequency Cepstral Coefficients (MFCC) as speech features from audio. Since all the 19 speech features are not equally important, therefore, feature selection algorithms are used to select the most efficient features. The various statistical algorithm such as Analysis of Variance (ANOVA), Kruskal-wallis, and Friedman test are employed to analyze the significance of features along with Incremental Feature Selection (IFS) technique. Statistical analysis is used to analyze the statistical significance of the speech features and after that IFS is used to select the speech feature subset. Furthermore, multiclass Support Vector Machine (SVM), Artificial Neural Network (ANN) and Naive Bayes (NB) machine learning techniques are used to recognize the speech for both the audio and visual modalities. Based on the recognition rate combined decision is taken from the two individual recognition systems. This paper compares the result achieved by the proposed model and the existing model for both audio and visual speech recognition. Zernike Moment (ZM) is compared with PZM and shows that our proposed model using PZM extracts better discriminative features for visual speech recognition. This study also proves that audio feature selection using statistical analysis outperforms methods without any feature selection technique.

Keywords

Audio-visual Speech Recognition, Lip Tracking, Pseudo Zernike Moment, Mel Frequency Cepstral Coefficients (MFCC), Statistical Analysis, Incremental Feature Selection (IFS).

DOI: 10.9781/ijimai.2021.09.001

I. Introduction

Human speech production and perception are bi-modal, like audio, visual features can also be extracted from speech. Research in Audio-Visual Automatic Speech Recognition (AV-ASR) is promising when a speech signal is affected by acoustic noise, different environments, and recording channels [1]. Speech is one of the ancient ways to express ourselves and speech recognition develops the methodologies that enable recognition of spoken word into text. There are many real-world applications where speech recognition is applied to authenticate [2], [3], especially for remote access of a system [4]. Audio alone also gives a good performance in a clean environment, but in a noisy environment, the signal may degrade [1]. Therefore, adding visual features make the system more robust, because visual features are less sensitive to noise [5]. But visual speech recognition is a challenging problem because visual features provide very less information as compared to an acoustic signal [5]. Research is going on in this area to find more and more robust and specific features that convey more accurate visual features. Visual features can be appearance-based, shape-based, and appearance and shape features. All the visual feature extraction methods include determination of the region of interest (ROI), face detection, and lip tracking. The audio-visual integration mechanism also plays a very crucial part in the AV-ASR research where two types of fusion can be done, feature fusion and decision level fusion [6]. Decision fusion is useful for individual analysis of time frames and phone segments. Frame level or feature level fusion is difficult because of frame mismatch and asynchrony of audio-visual data [6]. The fusion of audio-visual modality ensures better and convenient recognition than a single modality. AV-ASR can be applied to build a robust and secure authentication system, silent speech recognition system for deaf people, etc. But it introduces the challenging task of localization of mouth and lip tracking. Prashant Borde et al [5] have introduced the application of shape-based visual features for isolated word recognition. They have used Zernike Moment (ZM) [7] for visual feature extraction. This paper is primarily focused on building a speech recognition model that utilizes both audio and visual features i.e. audio-visual speech recognition based on audio-visual features and integration method. Here, we have used the shape-based visual features and the features are extracted from the lip contours.

The major contributions of this paper are as follows:

- Proper articulation is the most important lip-reading condition i.e. quality of speech of a speaker and angle of view. Thus, Pseudo

* Corresponding author.
E-mail addresses: dnsaswati@gmail.com (S. Debnath), pinkiroy2405@gmail.com (P. Roy).

Please cite this article in press as:
S. Debnath, P. Roy. Audio-Visual Automatic Speech Recognition Using PZM, MFCC and Statistical Analysis, International Journal of Interactive Multimedia and Artificial Intelligence, (2021), http://dx.doi.org/10.9781/ijimai.2021.09.001

Zernike Moment (PZM) [8] is proposed here for the extraction of visual features from the lip contour. The proposed algorithm extracts the shape based visual features to calculate the lip geometry of a speaker.

- Mel Frequency Cepstral Coefficients (MFCCs) [9] are widely used cepstral features extracted from the audio signal. In this study, the significance of MFCCs is calculated using different statistical algorithms. ANOVA, Kruskal-wallis, and Friedman tests are used in the proposed model to analyze different cepstral features and their significance. After the statistical analysis of features, the IFS method is used to select the features subset from the speech signal incrementally.
- To meet the final objective of AV-ASR, this paper proposes threshold-based decision fusion which improves the system performance.

Comparison of results is also explained in this paper based on the research paper published by Prashant Borde et al. Visual speech features are extracted using PZM and ZM using vVISWa [12] and 'CUAVE' [49] datasets and the results are compared for both the feature extraction techniques. Similarly, this paper also compares the results of audio speech recognition. The paper is organized as follows: Section II gives the literature review of AV-ASR, the proposed model is introduced in section III. Database description and experimental results are given in section IV and V. In section VI, we conclude our paper.

II. Literature Review

A. Audio-visual Speech Recognition

Audio-visual speech recognition is an active research area. To improve recognition performance in noisy environments visual information is added to automatic speech recognition.

Prashant Borde et al. [5] in 2014 have introduced Zernike features for visual speech recognition. The work described audio-visual speech recognition, which included face as well as lip detection, visual feature extraction, audio feature extraction, and recognition. The system was divided into two phases- the recognition of visual speech and the recognition of audio speech. Viola-Jones algorithm has been used for mouth localization or ROI detection. After extraction of ROI, the authors have used Zernike Moments (ZM) and Principal Component Analysis (PCA) for visual speech recognition. For audio speech recognition, they have used MFCC features. However, ZMs are sensitive to noise, and extracted features are scale as well as rotation variant.

Kuniaki Noda et al. [13] proposed AV-ASR, using a deep learning architecture, and introduced a connectionist-HMM. The system has three phases, in the first and second phase, deep de-noising autoencoder, as well as Convolutional Neural Network (CNN), have been used for acquiring noise-robust audio feature and for visual feature extraction, respectively. After that, a multi-stream HMM has been utilized here for integrating individual audio-visual features. De-noised MFCC features were used as an audio feature while CNN was used to predict the phoneme levels from the corresponding mouth area of the input image. After feature extraction, both the audio and visual features have been provided as an input to the Multi-stream HMM (MSHMM) for integration, which leads to a recognition of isolated words. However, the visual speech features extracted by CNN are not translation, rotation, and scale-invariant. Thus, the proposed method failed to meet the robustness due to illumination variation.

An experiment on continuous Audio-visual recognition was performed by Jeffrey B.Mulligan et al. [14]. They deployed the N-best approach for decision fusion. The recognizer that has been developed so far gives the best result in noise-free environments, but results degrade when it comes under noisy conditions. The authors have shown that their proposed system improves the robustness in all the situations where the audio signal is distorted. Data from both the audio-visual modality was first processed separately and then they combined them.

Visual speech information from the speaker's mouth region has been successfully shown to improve noise robustness of automatic speech recognizer by Gerasimos Potamianos et al. [15]. Thus, it has been promising to extend the usability into the human-computer interface. The authors have designed the visual front-end, based on a cascade of linear image transform. They have also added audio-visual speech integration. New work on a feature and decision fusion combination, the modeling of audio-visual speech asynchrony, and incorporating modality reliability estimates to the bi-modal recognition process have been analyzed. They also briefly touched upon the issues of audio-visual speaker adaptation. The experiments were carried out using three multi-subject bi-modal databases, ranging from small to large vocabulary recognition tasks, recorded at both visually controlled and challenging environments.

Namrata Dave [16] in 2015 has presented a lip-localization based visual feature extraction method. The proposed method segments the lip region from the image. To synchronize the lip movements with input audio they have segmented the lip region. Thus, the author has presented a color-based approach for the localization of lips. The main goal of their work was to synchronize lips with the input speech. Therefore, synchronizing with audio, viseme visual features have been extracted from the input video frame. HSV and YCbCr color models along with various morphological operations have been used. However, color-changing features are not very effective in AV-ASR research because they are sensitive to noise and illumination. Poor illumination does not give very good performance in a color model. Illumination affects the pixels values of an image. The color model also increases the experimental complexity.

Alin G et al. [17] proposed lip geometry and optical flow for capturing mouth movement. The method combined appearance-based features with the statistical approach for lip reading. However, the audio-only speech recognition has still lacked in robustness issues in a noisy condition while the video information is more reliable in real-time. The optical flow analysis captured the motion information of the speaker's mouth region. For the classification, they have used the Hidden Markov Model (HMM). A different noisy environment is a strong requirement for developing a robust speech recognition system. In this proposed method, the appropriate weights measurement is a very crucial part for different data medium. The author also mentioned that the system's accuracy could decrease because of a large number of features.

The lip movement of an individual speaker has been added to the acoustic features of speech for AV-ASR. Stéphane Dupont and Juergen Luettin [18] proposed a system that consists of three modules: the visual module, an acoustic module, and a sensor fusion module. Lip contour and grey level information were used as visual speech features. The acoustic features Perceptual linear prediction (PLP) and noise-robust RASTA PLP have been extracted from the speech signal. The system combined the visual and audio features using a multistream HMM. The appearance-based model for noise-robust audio-visual speech recognition has been introduced by the authors.

Continuous audio-visual digit recognition using N-best decision fusion has been introduced by Georg F. Meyer et al. [6]. The main contribution of the paper was decision fusion in audiovisual continuous speech recognition at the utterance level and proposal of an algorithm called N-best decision fusion. For the audio feature, they

have taken 12 orders cepstral coefficients (MFCC) and calculated the word error recognition (WER) rate. For video feature recognition, lip shape has been measured.

In [45], the authors introduced visual speech recognition by calculating the Gray Level Co-occurrence Matrix (GLCM) and Gabor convolve algorithm for discriminative feature extraction of the lip. They have collected a dataset of three Indian languages of English, Kannada, and Telugu for the experiments. GLCM provides the statistical texture features of lip movements. In this work, the authors have used four GLCM features such as contrast, energy, entropy, and correlation for the calculation of lip parameters. The mean and variant of the filtered image have also been calculated by using the Gabor filter. Thus, the main objective of this work was to analyze the texture of different images of lip movements.

B. Audio Speech Recognition

The work in [19] was carried out for the automatic recognition of English digits in 2010. The main objective of this study was to design and execute an English digits recognition system with the help of Matlab; using Hidden Markov Model. The framework perceives the speech waveform by making an interpretation of the speech waveform into an arrangement of high- light vectors using the popular technique MFCC.

Hindi Number Recognition [20] system was carried using Gaussian Mixture Models and MFCC. In the primary stage vowel acknowledgment models are created, which is supervised learning and in the subsequent stage, testing of the prepared models has been performed. Spectral components are separated from the discourse signals of the digits (0-9) and these elements are utilized to prepare Gaussian mixture models.

In [21] an idea is proposed that was a digit recognition system using Reservoir Computing (RC) which is a concept of machine learning. It is a non-linear dynamical system. It computes the state likelihood in HMM through two-layer Recursive Neural Network. The input hidden layer repetitively interfaces non-straight neurons with a settled number of non-prepared coefficients which is called a store (reservoir). They tested multilayer systems with 8000 and 16000 neurons. Later they performed a systematic evaluation using AEF (Advanced front end) where they replaced MVN features with AEF features and obtained significant gains in GMM-HMM recognizer.

S. Lokesh et al. [22] discovered a bidirectional recurrent neural network-based automatic Tamil speech recognition system in 2018. Bidirectional recurrent neural network (BRNN) with a self-organizing map (SOM) is used for the classification of Tamil speech. Savitzky–Golay filter is used for pre-processing to remove noise. For feature extraction, they have used discrete cosine transform and perceptual linear predictive coefficients. Using their proposed BRNN-SOM method 93.6 % accuracy was achieved for Tamil speech recognition.

The selection of feature vector from MFCC and Sequence-based Mapped Real Transform (SMRT) coefficients has been proposed by the author Mini p p et al. [23]. The first feature set was the coefficients extracted from all frames and after feature fusion, feature dimension reduction has carried out using a statistical measure such as energy, sum, mean, standard deviation, and energy distribution. These statistical measures are applied on the time average base to derive the second feature set. To solve the length variation problem of the speech signal, all the statistical measures are applied on the ensemble average base for generating the third feature vector. Furthermore, they have used Support Vector Machine (SVM) for the classification of the speech signal.

In [24] the authors introduced optimal speech feature extraction as well as feature selection using Artificial Bee Colony and Particle Swarm Optimization (ABC-PSO) hybrid algorithm. They have extracted eight types of statistical and acoustic features and ABC-PSO has been proposed for the selection of optimal features. After that SVM has been used to carry out the recognition process.

Nasir Saleem et al. [46] presented a detailed survey on unsupervised single-channel speech enhancement algorithms. The speech enhancement algorithms on unsupervised single-channel perspectives are analyzed and presented. Various methods have been discussed by the author for improving noisy speech. They have reviewed different approaches such as spectral subtraction, wiener filtering, minimum mean square error estimators, signal subspace, etc, and presented the experimental overview of these approaches. The authors have found that these methods show improvement in speech quality but speech intelligibility remains medium, thus, various problems have been introduced in the paper for designing robust single-channel speech enhancement algorithms.

In 2019, Nasir Saleem et al. [47] proposed speech enhancement using deep neural network (DNN) in complex noisy environments. They have also used an ideal binary mask (IBM) as a binary classification function during training and the trained DNNs are used for estimating IBM during the enhancement stage. The mean square error (MSE) has been used as an objective cost function at various epochs. The experimental results at different input SNR of this research showed that DNN-based speech enhancement performed better in a complex noisy environments than the competing methods in terms of perceptual evaluation of speech quality (PESQ), segmental signal-to-noise ratio (SNRSeg), log-likelihood ratio (LLR), weighted spectral slope (WSS), short-time objective intelligibility (STOI) and also improved the speech intelligibility an average 6.5% improvement during experiments.

Issues from the literature review:

- In the visual speech recognition approach, features are not translation, rotation invariant in many studies.
- Audio-visual integration is not done in many research works.
- Shape-based feature ZM has minimum feature dimensions also is very sensitive in noise.
- In audio, there are very few algorithms used to select the speech features, the majority of work focused on feature extraction and classification using machine learning.
- In the fusion method, a frame-level fusion mismatches the audio and visual frame.
- Feature fusion is partially valid because there is a different data rate of audio and visual data and differing segmentation [6].
- Many lip-reading systems use vizeme based representations for the visual recognition and phone-based for audio recognition but the co-articulation effects cause the asynchrony between the phone level segmentation of audio and visual data.

III. Proposed Methodology

Audio-visual speech recognition includes two separate processes of recognition: audio speech recognition and visual speech recognition. The proposed methodology of audio-visual speech recognition is shown in Fig. 1. The system includes the following steps.

a) **ROI detection**: Face and ROI have been detected using Viola Jones algorithm.

b) **Visual speech feature extraction**: After ROI detection visual features are extracted using PZM from the lip contour. PZMs are rotation and translation invariant of the image. ZM is also used with our dataset and shows that our proposed model using PZM

gives better recognition than the ZM.

c) **Audio feature extraction**: MFCC feature extraction is used and also extraction of significant features is done using statistical analysis.

d) **Audio feature selection**: Statistical algorithms are used to select efficient audio features. The proposed method is a combination of different statistical analysis and IFS.

e) **Classification**: Classification of audio-visual speech is carried out using multiclass SVM, ANN and NB classifier.

f) **Decision making**: Combined decision is taken from audio and visual speech recognition. Here, we propose a threshold-based decision level fusion to overcome frame level mismatch.

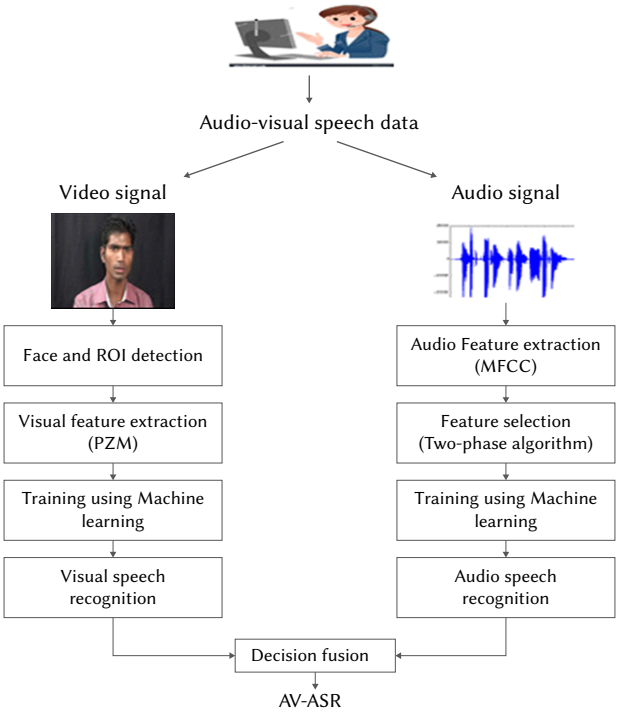

Fig. 1. Proposed model of AV-ASR.

A. Visual Feature Extraction

Visual features can be categorized by appearance-based, shape-based, and appearance and shape-based features. All of these methods include the determination of the ROI, face detection, and lip tracking. For visual feature extraction, we consider the video stream as an input. From each utterance, we have extracted the frames and processed each frame separately to obtain the discriminative features. Visual feature extraction method includes:

- Detection of the face and ROI (speaker's lip contour) using Viola-Jones algorithm.
- Calculate the visual features from lip contour.

1. Viola-Jones For ROI Detection

The Viola-Jones object detection framework facilitates Haar-Like [25] [26] features to be extracted from a face image as the initial step. The reason for using Haar-like features over the raw pixel value of the image is to reduce the in-class variability while increasing the out-of-class variability, which makes the classification easier. The contrast variances between the pixel groups are used to determine relative light and dark areas. It considers neighbouring rectangular regions in the image that is targeted for facial detection. After that, it sums up the pixel intensities in each region and calculates the difference between these sums. This difference is then used to categorize subsections of an image. In a human face, it is common that the region of the eyes is darker than the region of the cheeks. Therefore, a common Haar-like feature for face detection is a set of two adjacent rectangles that lie above the eye and the cheek region [27]. Fig. 2 and Fig. 3 depict pre-processing of visual feature extraction method.

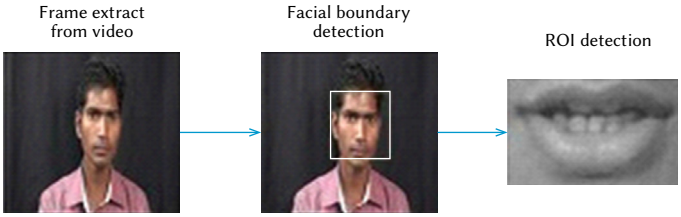

Fig. 2. Pre-processing of visual feature extraction from video.

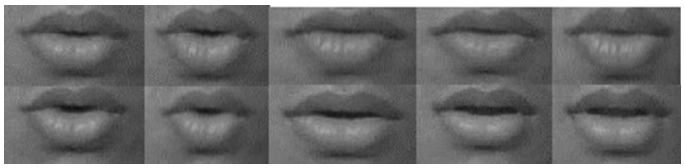

Fig. 3. Lip contour: open and close mouth for a particular word uttered by speaker.

B. Visual Speech Feature Extraction

Zernike Moment (ZM)

ZM is an orthogonal polynomial which is independent of scale and rotation of the image. It has less information redundancy and is used to capture discriminating feature of image frames.

Pseudo Zernike Moment (PZM)

PZMs are orthogonal moments on the unit disk defined by mapping an image onto a set of pseudo-Zernike polynomials [8]. PZMs are also rotation and flipping invariant. ZM polynomials are defined in polar coordinates. The orthogonal moments represent an image with the minimum number of redundant information [8]. PZM of order n and repetition of m can be computed over a unit disk by the following equation [28], [29].

$$A_{n,m} = \frac{n+1}{\pi} \iint_{x^2+y^2 \le 1} f(x,y) V^*_{n,m}(x,y) dxdy \quad (1)$$

Where, n= 0, 1, 2, 3, …., ∞ defines the order, f(x,y) is the function being described, * denotes the complex conjugate, while m is the positive or negative integer depicting the angular dependence, $V^*_{n,m}(x,y)$ is the complex pseudo Zernike polynomial which is defined by:

$$V_{n,m}(x,y) = R_{n,m}(x,y) \exp^{jm\theta} \quad (2)$$

Where (x,y) are defined over the unit disc, here, $R_{n,m}(x,y)$ is the real-valued radial polynomial, n ≥ 0, |m|≤ n, and the radial polynomials $R_{n,m}(x,y)$ are defined as:

$$R_{n,m}(x,y) = \sum_{s=0}^{n-|m|} (-1)^s \frac{(2n+1-s)!\,(x^2+y^2)^{\frac{n-s}{2}}}{s!\;[n+|m|+1-s]!\;\;[n-|m|-s]!} \quad (3)$$

Where, $n = 0, 1, 2, 3, ..., \infty$ defines the order and m is the positive or negative integer value subject to |m|≤n. According to simple enumeration, the set of pseudo-Zernike polynomials contains $(n + 1)^2$ linearly independent polynomials of degree ≤n [28].

PZM has been proven to be more efficient than the conventional ZM because of their feature representation capabilities [29]. The feature vector of ZM has 36 dimensions for maximum 10th order while PZM has 66 dimensions of feature vectors [29]. PZM is more efficient

to recognize the similar image frames since the number of features are more in PZM than the ZM. Also, it has been proven by Mukundan et al. that PZM are less sensitive to noise than the ZM for recognition of the image frame [30], [31]. The proposed visual feature extraction method using Viola-Jones and PZM is given in algorithm 1.

Algorithm 1: Shape-based visual features calculation using PZM

1: Input: Video of a speaker
2: Output: Lip geometry calculation
3: **procedure**: VISUAL SPEECH FEATURE EXTRACTION
4: Extract frame // Read video data
5: N←number of frames
6: bbox←(facedetector,N)
7: *videoFrameN = insertShape*(*videoFrame*,' *Rectangle*', *bbox*);
8: image write "detectedface.jpg"
9: MouthDetect : I←vision.
CascadeObjectDetector('Mouth','MergeThreshold'k);
k=threshold value [detection of object using Viola -Jones]
10: I← detected mouth area from each frame
11: **for** i= 1 to I **do**
12: img ←Load image
13: Normalize co-ordinates to [-squreroot2,squreroot2] and calculate origin or centroid ; x and y two coordinates
14: x1 = 2/(squreroot (2)*(d(1)-1)); d= size of image (dimension)
15: y1 = 2/(squreroot (2)*(d(1)-1));
16: [x,y]=meshgrid(1/ squreroot (2):x1:1/ squreroot (2),1/ squreroot (2: -y1:
-1/ squreroot (2));
17: $x^2 + y^2 \le 1$; // Compute unit circle
18: pixels inside the unit circle [*cimage, cindex*] = *p*1(*img, m*); m=zeros of d
19: z = p1(x+i*y,m);
20: p ← compute z;
21: q ← angle(z);
22: Compute order n and repitition m
23: **for** n= 1:length (l) **do** //n=order of PZM
24: n1 = l(n);
25: **for** r=1:length(m1) **do**
26: *V* = *pzpolynomial*(n1, *m*1(*r*), *p*, *q*);
27: PZp1 = *cimage* * *conj*(*V*);
28: *PZM* = 2 * (*n*1 + 1) * *sum*(*sum*(*pzp*1))/(*d*(1)^2 * *pi*);
29: pzm (u,z)= round(p(PZMoment(I,i,j)));
Magnitude of each component is evaluated and rounded off to nearest integer.
30: **end for**
31: **end for**
32: **end for**
33: end

C. Audio Feature Extraction and Statistical Analysis

In this step, 19 MFCCs are extracted from each input audio data. Here, we have used 19 MFCC features because the increasing level of spectral information comes from the higher order of coefficients and we need more information from more coefficients to select efficient features. The main aim of feature selection is to select the effective feature subset that can increase the classification accuracy while reducing the irrelevant and redundant features [32]. The feature extraction and selection methods are described below:

MFCC:

The most popular acoustic feature extraction technique MFCC was first introduced in 1980 by David and Mermelstein [9]. Today most of the speech recognition system focuses on the short-term spectral features which are captured from a short frame of the speech signal. The MFCC feature extraction consists of the following major steps:

- Framing and windowing: At first, each speech signal breaks down into short time duration by splitting the signal into several frames instead of analyzing the complete signal at once. After framing the signal, a window function is multiplied with each frame of the speech signal. We perform windowing in order to avoid unnatural discontinuities in the speech segment and the distortion in the underlying spectrum.
- Discrete Fourier Transforming: For extracting spectral information from a discrete frequency band Discrete Fourier Transform (DFT) is used. Fast Fourier Transform (FFT) is the most commonly used algorithm to compute the DFT. Here we are using FFT to convert the signal from time domain to frequency domain for preparing the next stage (Mel frequency warping).
- Mel-Frequency Warping: The FFT of the signal gives the magnitude, frequency response of each frame. After getting FFT the Mel filter bank includes the following calculations.
- The Mel scale: The result of FFT is the information about the amount of energy that the signal contains at each frequency band. For a given frequency f we can use the following formula to compute the Mels in Hz: Mel scale [9] is defined as:

$$m_f = 1125 \ln\left(1 + \frac{f}{700}\right) \quad (4)$$

Low frequency components of the speech signal carries much more information compared to the high frequency components. Mel scaling is performed in order to place more emphasis on the low frequency components. Since Mel filter banks are non-uniformly spaced on the frequency axis, so we have more filters in the low frequency regions and less number of filters in high frequency regions.

- Cepstrum: In the final step, the log mel spectrum is converted back to time. The result is called the mel frequency cepstrum coefficients (MFCCs) [33]. For the given frame analysis, the cepstral representation of the speech spectrum is a good representation of the local spectral properties of the signal. Since the mel spectrum coefficients are real numbers (and so are their logarithms), Discrete Cosine Transform (DCT) is performed to convert them into time domain [34].
- As per the procedure above, for each speech frame of about 20 ms with overlap of about 10 ms, a set of mel-frequency cepstrum coefficients are computed. This set of coefficients is called an acoustic vector. These acoustic vectors are used to represent and recognize the speech. Therefore, each input utterance is transformed into a sequence of acoustic vectors. We have extracted 19 mfc coefficients for each frame. From these 19 coefficients, feature 1 is the energy value of speech and feature 2 represents the broad shape of spectrum, features 3 to feature 7 represent the pitch information, features 8 to 13 represent the shape of spectrum and 14 to feature 19 provide the pitch or tone information which are the lowest dimensions of DCT coefficients. All these features are not equally important; therefore, we have calculated the F-statistics using statistical algorithm. The highest F-value of feature represents the most significant speech feature.

D. Audio Feature Selection

Analysis of Variance (ANOVA): ANOVA is a very effective

statistical method to test the difference in means between groups [10]. It assesses the potential difference in a scale-level dependent variable to a nominal-level variable having two or more categories. There are many advantages for that ANOVA has been used for feature selection [35], [36].

$$F(\xi) = \frac{sB^2(\xi)}{sW^2(\xi)} \tag{5}$$

Where, $sB^2(\xi)$ is the sample variance between groups (also called Mean Square Between, MSB) and $sW^2(\xi)$ is sample variance within groups (also called Mean Square Within, MSW) [16]. Decision is made using F-statistics [16], in one way ANOVA, the F-statistics is the ratio calculated by following equation:

$$F = \frac{variation\ between\ sample\ means}{variation\ within\ the\ samples} \tag{6}$$

Audio feature selection (ANOVA with IFS):

The method is calculated using the following steps:

ANOVA calculation:

- First, find the group mean of MFCC feature of each input data group.
- Find the overall mean i.e. mean of all MFCC feature matrices.
- Calculate within group variation, that is, the total deviation of each feature score from the group mean of each input data.
- Calculate the deviation of each input group mean from the overall mean ($ m2 $), i.e. the between group variation.
- Find the F statistics, the ratio between group variations to within group variation. The ratio calculates the measure of dispersion i.e. how much difference is there from each feature to the mean. Larger the value of F defines the more significant speech feature. These steps are performed for all the speech features.

IFS calculation for cepstral features is as follows:

- The feature subset starts from the feature with the highest F value in the ranked feature set.
- A new feature subset is produced when the feature with the second-highest F value is added.
- Until all the candidate features are added, this process continues from the highest F value to the lowest F value.

Kruskal-wallis test: Kruskal-wallis is a rank based nonparametric test which is used to evaluate the significant difference between two or more groups of an independent variable on an ordinal or continuous dependent variable. It is a Chi-square distribution. The test is performed to obtain the Chi-square value of each feature and rank those feature in a decreasing order. Null hypothesis is important for this test, if the Chi-square distribution score is less than the critical chi-square value then null hypothesis is accepted or has same group otherwise it is rejected or in a different group [37].

- Chi-square distribution uses categorical data, the mean (μ) of the distribution is the number of degrees of freedom (f) i.e. μ= f and variance (σ) of the distribution is σ=2 * f.
- The Chi-square is calculated using the following equation:

$$Chi - square = \sum \frac{(O-E)^2}{E} \tag{7}$$

Where, O is the observed value, E is the expected value.

- It calculates whether there is any statistically significant difference between different coefficients of cepstral features of different classes. The less Chi-square value indicates that less statistical difference. The higher Chi-square value indicates the more statistical difference between cepstral features of different classes. Therefore, the features with higher Chi-square values are selected using this Chi-square calculation. After feature ranking IFS is used here to select features for the classifier. The feature with highest Chi-square value is selected first, and then second highest value and so on.

Friedman test: Friedman test is a non-parametric test and alternative to one-way ANOVA with repeated measures. It is used to test differences between groups when the dependent variable being measured is ordinal [38]. The probability distribution is a Chi-square distribution and according to Chi-square value, features are ranked in order. Although Kruskal-wallis and Friedman test both calculate the Chi-square value but difference is there. Kruskal-wallis is the non-parametric test equivalent to one-way ANOVA while Friedman is a non-parametric test equivalent to two-way ANOVA.

E. Classifiers

Multiclass classifier SVM:

SVM [11], [39], [42] is a supervised machine learning that estimates decision surfaces directly rather than modeling a probability distribution across the training data. The kernel functions of SVM can be linear, Radial Basis Function (RBF), and polynomial function, which are used in this experiment. In a characteristics way, SVMs are two-class classifier but it can be used for multi-class classification problems. The methods used for multi-class classification are one-versus-all and one-versus-one. In this experiment, we have used the one-versus-all method for multi-class classification. We have 10 classes in this experiment, thus we generate 10 binary classifiers for 10 respective classes. After that, class levels are generated and the training dataset is created for each classifier of each class. After training, we have passed the test dataset. If the input test data belongs to a particular class, then the classifier generates a positive response and all the other classifiers provide a negative response.

Artificial neural networks (ANN): ANN [39], [40] is used for pattern classification because of its capability of nonlinear, non-parametric relationships between input data and output. Multi-layer feed-forward neural networks are the most popular neural networks which are trained using backpropagation learning algorithm. In this work, a multi-layer feed-forward ANN with the sigmoidal activation function is used with different hidden units.

Naive Bayes (NB): In NB [41], 'kernel' and 'normal' functions are used to model the feature distribution to decide the best distribution, and performed testing using fourfold cross-validation. It calculates the probability using Bayes theorem [41].

The audio-visual features are fed into ANN, SVM and NB classifiers to classify the speech. The performance is calculated by total number of words correctly recognized during the testing phase.

$$RR = \frac{C_s}{T_s} \times 100\% \tag{8}$$

Where, RR denotes recognition rate and C_s and T_s represent the correctly identified test sample and total supplied test sample, respectively. Audio speech recogn ition accuracy is computed using the same method.

F. Decision Fusion

It is important to integrate audio and visual recognition to obtain the final objective of AV-ASR. For the integration of two types of fusion such as feature fusion and decision, fusion can be done. Decision fusion is useful for isolated word recognition, where each recognized token is considered as a decision. Here, we propose a threshold-based fusion of audio-visual speech at the decision level. Based on the recognition accuracy of the system we consider the threshold for each system. According to that threshold of the individual system, we calculate

the system performance. If accuracy is greater than or equal to the threshold, then consider that input data is acceptable. When one of the recognition systems recognizes the respective input data of audio and visual speech, based on the threshold the system considers that speech gets recognized. In this work, we have processed audio and visual data separately. After the recognition of audio-visual speech, we have used decision fusion for the integration of audio-visual speech. Audio-visual speech has a different data rate and it creates a synchronization problem when encounters frame as well as feature level fusion [6]. Thus, to avoid the asynchrony of audio and visual data we have developed decision fusion. Many types of research have introduced a dynamic weighting method for AV-ASR integration. However, the decision fusion method improves performance while overcoming the issues of frame and feature level fusion. The decision fusion proposed in this paper is given in algorithm 2.

Algorithm 2: Decision fusion algorithm of audio-visual speech recognition

1. Input: Output of audio speech recognition and visual speech recognition

2. Output: Combined decision

3. Set threshold for audio speech recognition, and visual speech recognition

4 X = Audio speech recognition threshold

Y = Visual speech recognition threshold

5 if (*audio speech recognition* ≥ *X*) || (*visual speech recognition* ≥*Y*)

then

recognition=1

(accept) otherwise

0 (reject)

end

IV. Dataset Description

We have used an audio-visual English digit database 'vVISWa' and 'CUAVE' for AV-ASR.

'vVISWa' dataset: Prashant Borde et al. [12] published a paper about 'vVISWa' dataset in 2016. English 10 digits (zero, one, two, three, four, five, six, seven, eight, and nine) have been recorded. The database has been recorded in a lab environment. Dataset consists of 10 speakers, 6 male, and 4 female. Each speaker uttered each word 10 times. So, one word is uttered 100 times by a different speaker. Each individual word is uttered without any head movement.

'CUAVE' dataset: E.K. Patterson et al. [49] published the details of 'CUAVE' audio-visual database for multimodal human-computer interface. The 'CUAVE' dataset is used here for the experiments and to compare the results. The dataset consists of ten English digits from 0 to 9 of 36 speakers (18 male and 18 female speaker). Each digit is uttered five times by each speaker; thus, the total 1800 words has been used. The database was recorded in an isolated sound booth at a resolution of 720 x 480 with the NTSC standard of 29.97 fps. The audio is 16-bit, stereo, at a sampling rate of 44 kHz. There is also word-level labelling at millisecond accuracy, done manually for all sequences of the database.

V. Experimental Result and Analysis

A. Visual Speech Recognition

After detecting the ROI, PZMs are calculated for the lip region. PZM measures how lips are moving for a particular speech and based on these visual features, we classify the speech spoken by the speaker. The steps of calculating lip movements using PZM are given below:

1. First, consider the open area of the mouth for each frame of the lip.
2. Take the origin of the lip contour.
3. The pixel coordinates of lips are normalized to the range of a unit circle. i.e. $x2 + y2 \leq 1$
4. Calculate the angle and coordinates of each point. These are the features of visual speech.
5. Repeat steps 1 to 4 for every frame of a particular word and generates a feature matrix.

Calculate the angle and coordinates of each point. These are the features of visual speech. We have extracted frames from the video for each utterance. From each frame face and ROI is detected using the Viola-Jones algorithm. ZM and PZM [28] are calculated for every frame of lip contour to extract the discriminating feature of visual information. The PZMs have effective mathematical calculation to capture the different movements of the image. Here, we have taken 10 discriminative frames of lip contour and calculated the pseudo-Zernike feature of each frame. We have extracted 19 coefficients for each frame and 10 frames for a single utterance, therefore 10x19 feature matrix for a single digit. After extracting the visual features, we have applied multiclass SVM, ANN, and NB machine learning to train the system. The performance is calculated by the total number of words correctly recognized using visual features during the testing phase. The performances of visual speech recognition with different classifiers are presented in Table I, Table II, and Table III using 'vVISWa' dataset.

We have carried out all the experiments using 'CUAVE' dataset also and presented the recognition rate in Table IV, Table V and Table VI for visual speech recognition using different classifier.

TABLE I. Visual Speech Recognition Using ZMs, PZMs and ANN for 'VVISWa' Dataset

Exp. No.	Hidden layer	No. hidden nodes	Accuracy (%) using ZMs	Accuracy (%) using PZMs
1	2	30,20	70.00	73.34
2	2	40,30	69.54	72.98
3	2	50,40	69.23	71.56
4	2	60,50	70.12	72.89
5	2	70,60	70.00	72.10

TABLE II. Visual Speech Recognition Using ZMs, PZMs and NB for 'VVISWa' Dataset

Exp. No.	Distribution function	Accuracy (%) using ZMs	Accuracy (%) using PZMs
1	Normal	68.20	72.00
2	Kernel	70.34	74.65

TABLE III. Visual Speech Recognition Using ZMs, PZMs and SVM for 'VVISWa' Dataset

Exp. No.	Kernel function	Accuracy (%) using ZMs	Accuracy (%) using PZMs
1	**Radial basis function (RBF)**	**68.00**	**75.23**
2	Linear	61.54	67.45
3	Polynomial	66.12	70.56

TABLE IV. Visual Speech Recognition Using ZMs, PZMs and ANN for 'CUAVE' Dataset

Exp. No.	Hidden layer	No. hidden nodes	Accuracy (%) using ZMs	Accuracy (%) using PZMs
1	2	30,20	72.15	73.44
2	2	40,30	68.24	73.08
3	**2**	**50,40**	**73.12**	**75.34**
4	2	60,50	72.27	73.90
5	2	70,60	71.00	73.00

TABLE V. Visual Speech Recognition Using ZMs, PZMs and NB for 'CUAVE' Dataset

Exp. No.	Distribution function	Accuracy (%) using ZMs	Accuracy (%) using PZMs
1	Normal	67.20	70.00
2	**Kernel**	**73.34**	**75.00**

TABLE VI. Visual Speech Recognition Using ZMs, PZMs and SVM for 'CUAVE' Dataset

Exp. No.	Kernel function	Accuracy (%) using ZMs	Accuracy (%) using PZMs
1	**Radial basis function (RBF)**	**72.15**	**76.03**
2	Linear	64.00	67.54
3	Polynomial	65.21	71.60

B. Audio Speech Recognition

We have extracted19-dimensional MFCC features for the experiment. The statistical test is carried out to rank the MFCC features based on the F-statistics which has been discussed in section III D. The feature with the highest F value is placed first, that is ranked one, followed by the feature with second-highest value that is ranked two, and so on. The F-statistics is calculated by equation (6). The performance of the system is measured using equation (7). We have calculated the recognition rate using all the MFCC features and also the features subset resulting from the feature selection method. The recognition rate is carried out using SVM, ANN, and NB and IFS gradually concatenates the features for all the classifiers.

Table VII, Table VIII, and Table IX show the F-statistics value of speech features after statistical analysis for 'vVISWa' dataset. The performance of the system after feature selection technique using SVM, ANN, and NB are depicted in Table X, Table XI, and Table XII. Table XIII shows the performance of audio speech recognition using MFCC. From the experiment, it has been observed that using 'vVISWa' dataset, the SVM classifier with kernel function 'RBF' gives the highest accuracy that is 96.42 % for 12 cepstral features. The highest accuracy is obtained using ANOVA with IFS feature selection method. All the experiments are carried out using 'vVISWa' dataset.

TABLE VII. F-value After Statistical Analysis of MFCC Using ANOVA

Feature set	F-value	Feature set	F-value
f1	719.63	f11	127.68
f2	477.40	f12	31.50
f3	249.84	f13	65.62
f4	253.10	f14	110.67
f5	249.15	f15	35.55
f6	115.42	f16	85.87
f7	154.19	f17	50.45
f8	22.15	f18	21.76
f9	47.34	f19	45.56
f10	226.65		

TABLE VIII. F-value After Statistical Analysis of MFCC Using Kruskal-wallis

Feature set	F-value	Feature set	F-value
f1	634.64	f11	127.68
f2	376.00	f12	22.10
f3	265.23	f13	47.65
f4	364.78	f14	58.11
f5	188.25	f15	25.62
f6	60.12	f16	19.27
f7	91.19	f17	42.55
f8	62.15	f18	20.27
f9	32.84	f19	18.45
f10	87.65		

TABLE IX. F-value After Statistical Analysis of MFCC Using Friedman Test

Feature set	F-value	Feature set	F-value
f1	671.34	f11	127.68
f2	423.37	f12	46.20
f3	286.23	f13	35.75
f4	364.78	f14	22.87
f5	210.15	f15	29.56
f6	92.32	f16	17.68
f7	64.99	f17	28.34
f8	63.65	f18	16.07
f9	56.34	f19	16.35
f10	58.75		

TABLE X. Audio Speech Recognition Rate Using MFCC, Feature Selection and SVM for 'VVISWa' Dataset

Exp. No.	Feature selection method	No. of features	Kernel function	Accuracy
1	ANOVA+IFS	12	Radial basis function (RBF)	96.42
2	ANOVA+IFS	13	Linear	78.11
3	ANOVA+IFS	14	Polynomial	80.66
4	Kruskal-wallis+IFS	16	Radial basis function (RBF)	95.31
5	Kruskal-wallis+IFS	14	Linear	77.78
6	Kruskal-wallis+IFS	17	Polynomial	85.65
7	Friedman+IFS	12	Radial basis function (RBF)	93.45
8	Friedman+IFS	13	Linear	77.57
9	Friedman+IFS	14	Polynomial	90.34

TABLE XI. Audio Speech Recognition Using MFCC, Feature Selection and ANN for 'VVISWa' Dataset

Exp. No.	Feature selection method	No. of features	Hidden layer and nodes	Accuracy
1	ANOVA+IFS	13	2 (30,20)	92.06
2	ANOVA+IFS	11	2 (40,30)	94.78
3	ANOVA+IFS	14	2 (50,40)	90.88
4	ANOVA+IFS	13	2 (60,50)	93.96
5	ANOVA+IFS	13	2 (70,60)	90.34
6	Kruskal-wallis+IFS	12	2 (30,20)	92.18
7	Kruskal-wallis+IFS	13	2 (40,30)	91.68
8	Kruskal-wallis+IFS	13	2 (50,40)	91.85
9	Kruskal-wallis+IFS	19	2 (60,50)	90.75
11	Friedman+IFS	10	2 (30,20)	89.67
12	Friedman+IFS	12	2 (40,30)	92.89
13	Friedman+IFS	16	2 (50,40)	93.17
14	Friedman+IFS	14	2 (60,50)	92.78
15	Friedman+IFS	19	2 (70,60)	91.56

TABLE XII. Audio Speech Recognition Using MFCC, Feature Selection and NB for 'VVISWa' Dataset

Exp. No.	Feature selection method	No. of features	Distribution function	Accuracy
1	ANOVA+IFS	15	Normal	93.72
2	ANOVA+IFS	13	Kernel	94.23
3	Kruskal-wallis+IFS	17	Normal	94.01
4	Kruskal-wallis+IFS	16	Kernel	94.57
5	Friedman+IFS	19	Normal	91.72
6	Friedman+IFS	15	Kernel	92.87

TABLE XIII. Audio Speech Recognition Using MFCC and SVM

Exp. No	Classifier	Accuracy (%) using 'vVISWa' dataset	Accuracy (%) using 'CUAVE' dataset
1	SVM ('RBF' kernel function)	93.86	94.55
2	ANN (Hidden layer-2 and hidden nodes- 50,40)	93.67	93.35
3	NB (kernel distribution function)	92.86	94.00

The 'CUAVE' digit dataset is also used for the audio speech recognition and recognition accuracies are presented in Table XIV, Table XV and Table XVI. The highest accuracy achieved by this dataset is 98.0 % for 13 number of features using ANOVA with IFS.

TABLE XIV. Audio Speech Recognition Using MFCC, Feature Selection and SVM for 'CUAVE' Dataset

Exp. No.	Feature selection method	No. of features	Kernel function	Accuracy
1	**ANOVA+IFS**	**13**	**Radial basis function (RBF)**	**98.00**
2	ANOVA+IFS	12	Linear	75.23
3	ANOVA+IFS	13	Polynomial	82.12
4	Kruskal-wallis+IFS	15	Radial basis function (RBF)	96.32
5	Kruskal-wallis+IFS	14	Linear	77.00
6	Kruskal-wallis+IFS	16	Polynomial	82.02
7	Friedman+IFS	12	Radial basis function (RBF)	95.45
8	Friedman+IFS	13	Linear	76.57
9	Friedman+IFS	15	Polynomial	92.81

TABLE XV. Audio Speech Recognition Using MFCC, Feature Selection and ANN for 'CUAVE' Dataset

Exp. No.	Feature selection method	No. of features	Hidden layer and nodes	Accuracy
1	ANOVA+IFS	14	2 (30,20)	91.32
2	ANOVA+IFS	12	2 (40,30)	93.45
3	ANOVA+IFS	13	2 (50,40)	94.75
4	**ANOVA+IFS**	**13**	**2 (60,50)**	**96.55**
5	ANOVA+IFS	14	2 (70,60)	96.00
6	Kruskal-wallis+IFS	13	2 (30,20)	91.08
7	Kruskal-wallis+IFS	13	2 (40,30)	92.52
8	Kruskal-wallis+IFS	14	2 (50,40)	93.00
9	Kruskal-wallis+IFS	18	2 (60,50)	94.00
11	Friedman+IFS	10	2 (30,20)	90.22
12	Friedman+IFS	12	2 (40,30)	92.00
13	Friedman+IFS	15	2 (50,40)	93.55
14	Friedman+IFS	13	2 (60,50)	93.00
15	Friedman+IFS	19	2 (70,60)	92.11

TABLE XVI. Audio Speech Recognition Using MFCC, Feature Selection and NB for 'CUAVE' Dataset

Exp. No.	Feature selection method	No. of features	Distribution function	Accuracy
1	ANOVA+IFS	14	Normal	93.55
2	ANOVA+IFS	13	Kernel	95.74
3	Kruskal-wallis+IFS	16	Normal	93.00
4	Kruskal-wallis+IFS	15	Kernel	94.17
5	Friedman+IFS	17	Normal	90.00
6	Friedman+IFS	16	Kernel	93.65

C. Audio-Visual Speech Recognition Fusion

Here, we have considered decision level fusion for combining two systems because feature level fusion encounters the frame mismatch. The individual word recognition rate has been calculated for both audio and visual speech, after that using decision logic we integrate two modalities for the better result.

If one recognition system fails to recognize the input digit, then we can consider the result using another system. Decision fusion provides a better recognition rate for the overall system because each individual word is recognized as a token. The decision has been taken based on logic such that if the audio signal recognition rate is more than 90%, we have considered that audio speech is recognized. If the visual speech recognition rate is more than 70%, we have considered that visual speech is recognized. Thus, when the accuracy is greater than or equal to the threshold, then it is considered that the input data is acceptable. Based on the threshold, when one of the recognition systems recognizes the respective input data of audio and visual speech, the system considers that speech gets recognized. The proposed decision fusion method is represented by Algorithm 2.

D. Comparison of Results and Analysis

We have experimented separately for both audio and video data. The performances of our proposed model for audio speech recognition are 96.42 % and 98.00 % using 'vVISWa' and 'CUAVE' dataset respectively. For visual speech recognition, we have used lip tracking and 75.23 % accuracy is achieved using the proposed visual speech recognition model for 'vVISWa' dataset. For 'CUAVE' dataset, the recognition accuracy of visual speech is 76.03 %. We have compared the results of the existing model using ZM and our proposed model. Prashant Borde et al. [5] introduced ZM and MFCC features for audio-visual speech recognition respectively and achieved 63.88% accuracy

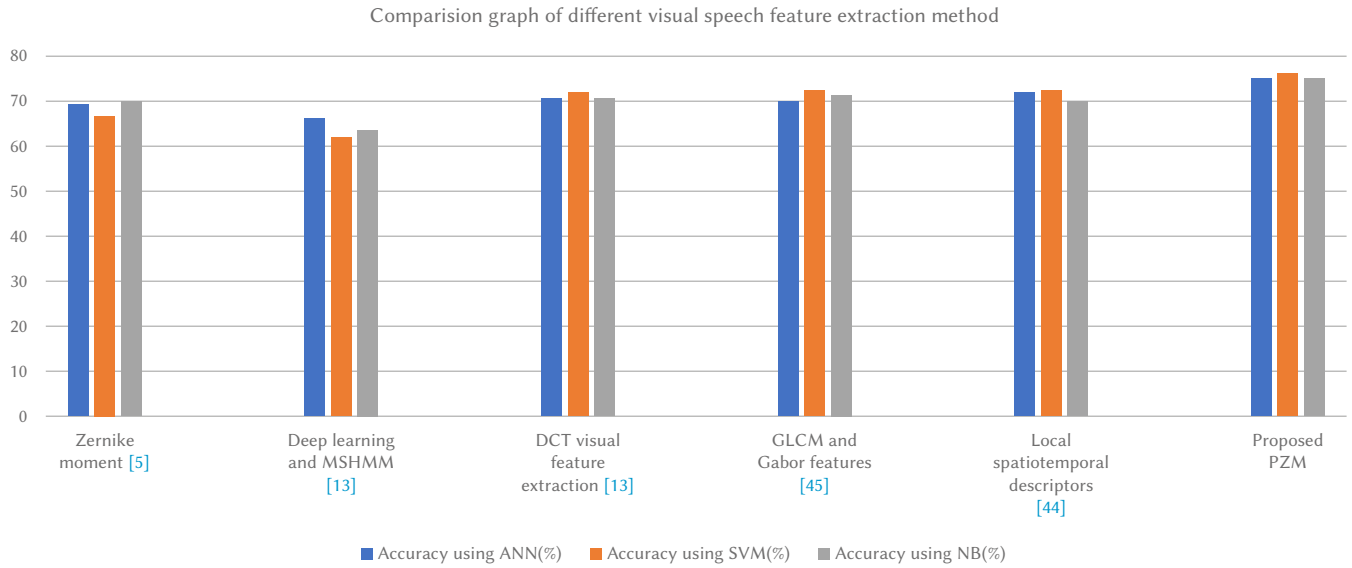

Fig. 4. Comparison of proposed visual speech recognition with existing method ('vVISWa' dataset).

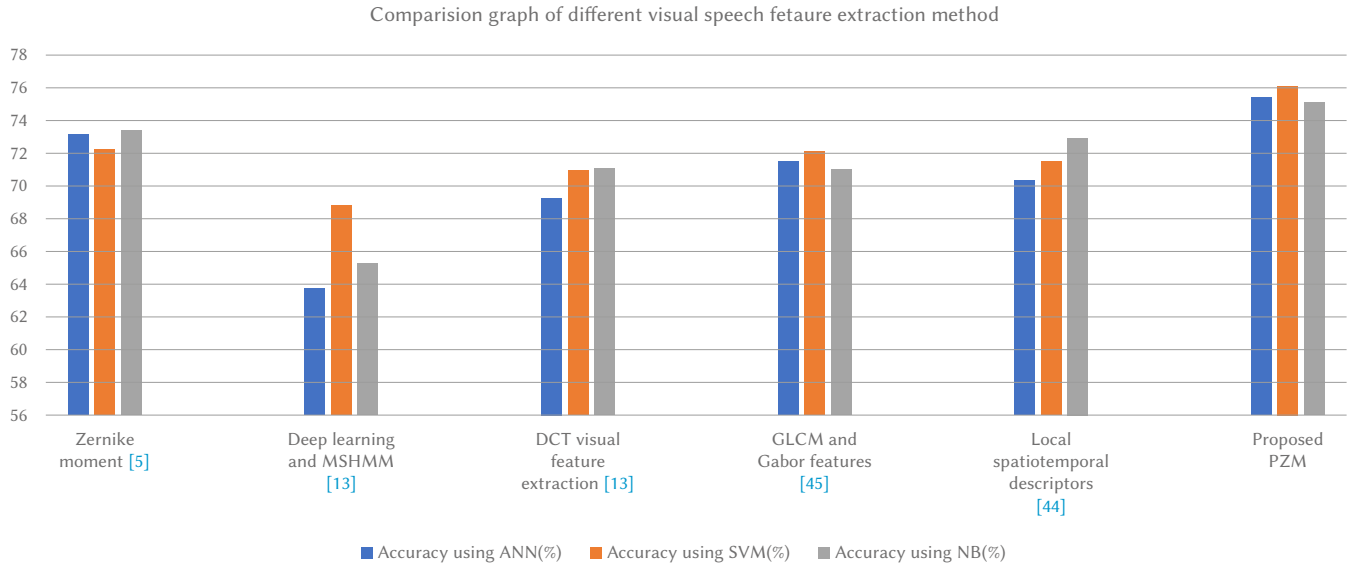

Fig. 5. Comparison of proposed visual speech recognition with existing method ('CUAVE' dataset).

for visual speech recognition. In this experiment, it has been observed that our proposed model using PZM gives better recognition accuracy than ZM for the both 'vVISWa' and 'CUAVE' English digit datasets, which is depicted in Fig. 4 and Fig. 5. Using PZM the system gives a better recognition rate because PZM has more feature dimensions and better feature representation capability, also PZM is less sensitive to noise than the ZM. We have also used local spatiotemporal descriptors [44] and GLCM and Gabor features [45] for lip reading in visual speech recognition. These are appearance-based features and capture the texture description of lip images. Appearance-based features consider all pixels within the ROI that are informative for speech recognition. However, the appearance-based features are sensitive to illumination, orientation variation, and position of the head [48]. Thus, these features are not very efficient for visual speech recognition. Shape-based features calculate the width and height of the lip contours of the speaker's lips. The proposed shape-based features extracted by PZM are illumination, rotation, translation, and scale-invariant. Therefore, in this research, we have introduced the shape-based feature extraction using PZM. The recognition rate of visual speech using GLCM and Gabor features are presented and compared in table XI. In audio speech recognition, when we select features based on feature selection algorithm it gives more accuracy than without any feature selection method. Using ANOVA, Kruskal-wallis and Friedman test statistical algorithm we have extracted the important features and model the system accordingly. The feature selection algorithm is important to remove the redundant features as well as to rank the significant features. All the individual classifiers select the feature subset using the feature selection method. In this paper, we have considered Zernike moment based visual speech recognition [5], deep learning and MSHMM [13], DCT visual feature extraction [43], local spatio-temporal descriptors [44], and GLCM and Gabor features [45] for comparison of visual speech recognition using 'vVISWa' and 'CUAVE' dataset and MFCC and HMM [19], MFCC and GMM [20], MFCC, SMRT and SVM [23] and optimal feature selector based on ABC-PSO [24] are taken for the comparison of audio speech recognition using our proposed model. Fig. 4 and Fig. 5 depict the comparison of the proposed visual speech recognition with existing methods for 'vVISWa' and 'CUAVE' dataset respectively. Fig. 6 and Fig. 7 represent the comparison graph of audio speech recognition of the proposed model and the existing models for both 'vVISWa'

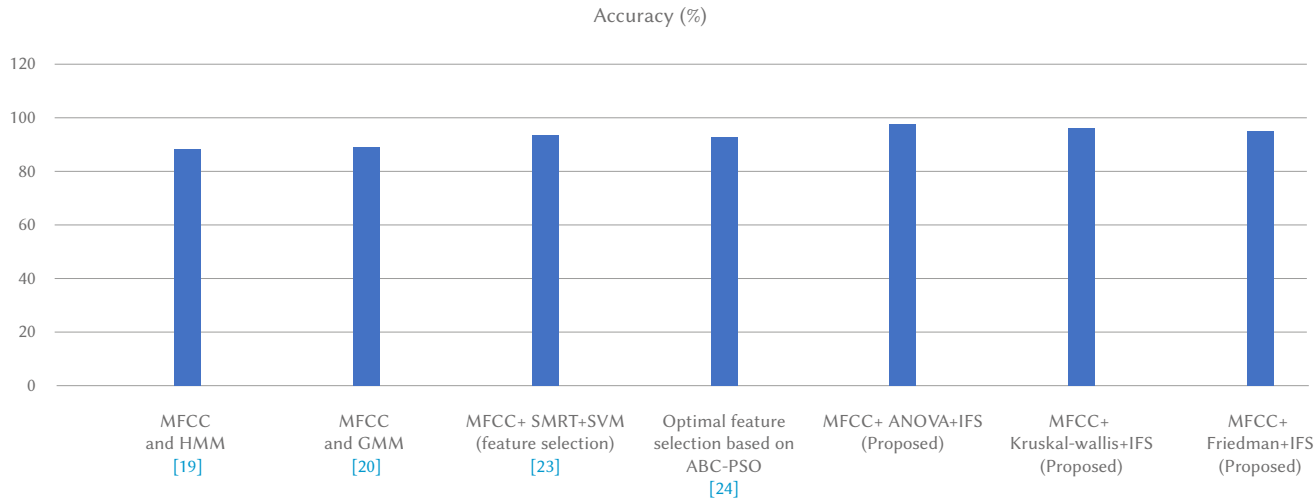

Fig. 7. Comparison of proposed audio speech recognition with existing method ('CUAVE'dataset).

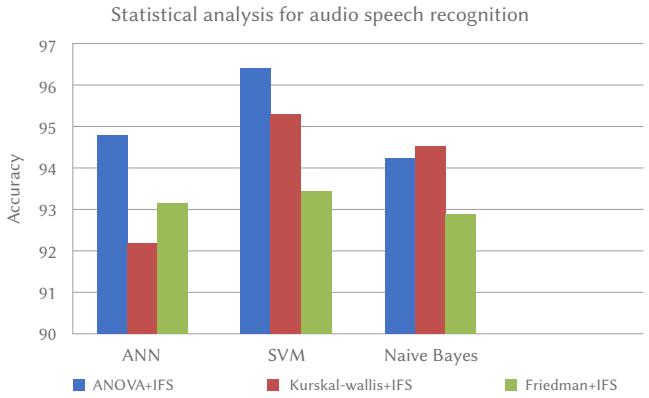

Fig. 6. Performance of different statistical analysis and classifiers for audio speech recognition ('vVISWa' dataset).

and 'CUAVE' dataset respectively. If both the systems are analysed individually, for only audio speech it gives a good recognition rate in a lab environment, but in a noisy environment, the audio signal may get corrupted, also the signal may degrade because of the different environment channels. In visual speech recognition, the face may not be properly detected all the time because of the camera and different lightning issues, also, a person's lip may not move properly all the time. But when we combine both the systems, it helps to overcome the shortcomings of individual recognition. The proposed system combines audio-visual recognition at the decision level and recognizes the speech based on audio-visual features.

From the experiment it has been shown that our proposed model gives better recognition accuracy than the existing methods. 'CUAVE' dataset provides better recognition accuracy than the 'vVISWa' dataset for both audio and visual speech.

VI. Conclusion and Future Work

In AV-ASR, the primary research on developing algorithms for the lip-reading, representation of visual features, and the integration of audio-visual information are the most promising areas. By watching the speaker's lip movement along with his voice can improve speech intelligibility especially in a noisy background and for hearing impaired people. Though it is an emerging field of research it still lacks proper visual articulations for visual speech recognition. Thus, the extraction of proper visual articulation attracts the interest of researchers in AV-ASR. Different types of lip-reading conditions provide very significant information regarding visual speech. This research has proposed shape-based visual speech features used for classification by the machine learning algorithms. The system includes two individual recognition: visual speech recognition and audio speech recognition. Visual speech recognition comprises of face detection, ROI detection, and lip tracking. A new visual feature extraction method using PZM has been proposed to track the lip movement. The PZM is an efficient orthogonal moment that describes a discriminating feature of an image or frame. In an audio speech, MFCC has been used and statistical algorithm along with IFS for selecting the significant features is proposed. The proposed method ranks the features based on their statistical significance and select features subset for the individual classifier. This paper compares the results using a feature selection method and without any feature selection method. After recognition combining the two modalities of audio-visual speech at the decision phase it gives the final outcome of AV-ASR. We use the threshold-based decision fusion and the threshold has been taken based on the average accuracy of individual recognition. The research can be extended in the future to develop a system using more specific audio-visual speech feature in a real-time environment. In the real-time, sometimes features may not be recognized properly because of noise, improper articulations. Thus, it is essential to capture more speaker-independent visual features.

References

[1] N. Moritz, K.Adiloglu, J. Anemuller, S. Goetze, B. Kollmeier, "Multi-Channel Speech Enhancement and Amplitude Modulation Analysis for Noise Robust Automatic Speech Recognition", Computer Speech & Language, vol. 46, pp. 558-573, 2017.

[2] D. Rudrapal, S. Das, S. Debbarma, N. Kar, N. Debbarma, "Voice Recognition and Authentication as a Proficient Biometric Tool and its Application in Online Exam for P.H People", International Journal of Computer Applications (0975 8887), vol. 39, no. 12, 2012.

[3] S. Singh and M. Yamini, "Voice based login authentication for Linux," 2013 International Conference on Recent Trends in Information Technology (ICRTIT), 2013, pp. 619-624, doi: 10.1109/ICRTIT.2013.6844272.

[4] Z. Saquib, N. Salam, R. Nair, N. Pandey, "Voiceprint Recognition Systems for Remote Authentication-A Survey", International Journal of Hybrid Information Technology, vol. 4, no. 2, 2011.

[5] P. Borde, A. Varpe, R. Manza, P. Yannawar, "Recognition of Isolated Words using Zernike and MFCC features for Audio Visual Speech Recognition", International Journal of Speech Technology, 2014, doi: 18.10.1007/s10772-014-9257-1.

[6] G. F. Meyer, J. B. Mulligan, S. M. Wuerger, "Continuous audiovisual digit

recognition using N-best decision fusion", Information Fusion, vol. 5, 2004.

[7] H. Marouf and K. Faez, "Zernike Moment-Based Feature Extraction For Facial Recognition Of Identical Twins", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), vol. 3, no. 6, 2013.

[8] A. Bhatia and E. Wolf, "On the Circle Polynomials of Zernike and Related Orthogonal Sets", Proceedings of the Cambridge Philosophical Society, vol. 50, no.1, pp. 40-48, 1954.

[9] S. B. Davis, P. Mermelstein, "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357-365, 1980.

[10] H. Ding, P. M. Feng, W. Chen, H Lin, "Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis", MolBioSyst, vol. 10, no.8, pp. 22292235, 2014.

[11] A. Ganapathiraju, J. E. Hamakerand, J. Picone, "Applications of Support Vector Machines to Speech Recognition", IEEE Transactions on Signal Processing, vol. 52, no. 8, 2004.

[12] P. Borde, R. Manza, B. Gawali and P. Yannawar. " Article: vVISWa A Multilingual Multi-Pose Audio Visual Database for Robust Human Computer Interaction", International Journal of Computer Applications, vol. 137, no. 4, pp. 25-31, 2016.

[13] K. Noda, Y. Yamaguchi, K. Nakadai, H. Okuno, T. Ogata, "Audio-visual speech recognition using deep learning", Applied Intelligence, vol. 42, 2014, doi:10.1007/s10489-014-0629-7.

[14] G. Meyer, J. Mulligan, S. Wuerger, "Continuous audio-visual digit recognition using N-best Decision Fusion", Information Fusion, vol. 5, pp. 91-101, 2004.

[15] G. Potamianos, C. Neti, G. Gravier, A. Garg and A. W. Senior, "Recent Advances in the Automatic Recognition of Audio-Visual Speech", Proceedings of the IEEE, vol. 91, no. 9, pp. 1306-1326, 2003, doi: 10.1109/JPROC.2003.817150.

[16] N. Dave, "A Lip Localization Based Visual Feature Extraction Method", An International Journal (ECIJ), vol. 4, no. 4, 2015.

[17] A. G. Chitu, L.J.M Rothkrantz, J. C. Wojdel, W. Pascal, "Comparison Between Different Feature Extraction Techniques for Audio-Visual Speech Recognition", Journal on Multimodal User Interfaces, vol. 1, no. 1, pp 720, 2007.

[18] S. Dupont and J. Luettin, "Audio-Visual Speech Modeling for Continuous Speech Recognition", IEEE Transacctions on Multimedia, Vol. 2, No. 3, 2000.

[19] T. S. Gunawan, A. A. M. Abushariah, M. A. M. Abushariah and O. Othman, "English Digits Recognition System Based on Hidden Markov Models," IEEE 978-1-4244-6235- 3/10/ , 2010.

[20] H. R. Goyal and S. G. Koolagudi, "Hindi Number Recognition using GMM," Global Journal of Computer Applications, vol. 63, no. 21, pp. 25-30, 2013.

[21] A. Jalalvand, F. Triefenbach, K. Demuynck, and J.-P. Marten, "Robust continuous digit recognition using Reservoir Computing," Computer Speech and Language , vol. 30, no. 1, pp. 135-158, 2015.

[22] S. Lokesh, P. M. Kumar, M. R. Devi, P. Parthasarathy, C. Gokulnath , "An Automatic Tamil Speech Recognition system by using Bidirectional Recurrent Neural Network with Self Organizing Map", Neural Computing and Applications , vol. 31, pp. 1521-1531, 2019, doi: https://doi.org/10.1007/s00521-018-3466-5.

[23] Mini P P, T. Thomas, R Gopikakumari, "Feature Vector Selection of Fusion of MFCC and SMRT Coefficients for SVM Classifier Based Speech Recognition System", 978-1-5386-6575-6 /18/$31.00 2018 IEEE.

[24] S. Mendiratta, N. Turk, D. Bansal, "Automatic Speech Recog- nition Using Optimal Selection of Features Based On Hybrid ABC-PSO", 2016 International Conference on Inventive Computation Technologies (ICICT), doi: 10.1109/INVENTIVE.2016.7824866.

[25] R. Lienhart, J. Maydt, "An Extended Set of Haar-like Features for Rapid Object Detection", Intel Labs, Intel Corporation, Santa Clara, n.d. Web. 2014.

[26] E. Gregori. "Introduction To Computer Vision Using OpenCV. Embedded Vision Alliance", 2012 Embedded Systems Conference, 2012.

[27] P.I. Wilson, J. Fernandez, "Facial feature detection using haar classifiers", Texas A & M University, (2014).

[28] C. Teh, R. Chin, "On Image Analysis by the Method of Moments", IEEE Transaction on Pattern Analysis and Machine Intelligence, vol.10, no. 4, pp. 496-513, 1988.

[29] C. Singh,R. Upneja, "Accurate calculation of high order pseudo Zernike moments and their numerical stability", Digital Signal Processing, vol. 27, pp. 95-106, 2014.

[30] K. M. Hosny , "Accurate pseudo Zernike moment invariants for greylevel images", The Imaging Science, vol. 60, doi: 10.1179/1743131X11Y.0000000023.

[31] R. Mukundan, K.R. Ramakrishnan, "Moment Functions in Image Analysis Theory and Applications", World Scientific, Singapore (1998).

[32] G. Chandrashekar, F. Sahin, " A survey on feature selection methods", Computer Electrical Engineering, vol. 40, no. 1, pp. 628, 2014.

[33] B. Soni, S. Debnath, P.K. Das, "Text-dependent speaker verification using classical LBG, adaptive LBG and FCM vector quantization", International Journal of Speech Technology September, vol. 19, no. 3, pp. 525-536, 2016.

[34] M.A. Hossan, S. Memon, M.A. Gregory, "A Novel Approach for MFCC feature extraction", 4th International Conference on Signal Processing and Communication Systems (ICSPCS), pp. 1-5, 2010.

[35] N. Settouti, M.E.A. Bechar, M.A. Chikh, "Statistical comparisons of the top 10 algorithms in data mining for classification task", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 4 no. 1, pp. 46-51, 2016.

[36] B. Niu, G. Huang, L. Zheng, X. Wang, F. Chen, Y. Zhang, T. Huang, "Prediction of substrate-enzyme-product interaction based on molecular descriptors and physicochemical properties", BioMed Research International, vol. 2013, article ID 674215, 2013.

[37] Y. Chan, R. P. Walmsley, "Learning and understanding the Kruskal Wallis one-way analysis of variance by ranks test for differences among three or more independent groups", Physical Therapy, vol. 77 no. 12, pp.1755-1761,1997 .

[38] D.W. Zimmerman, B.D. Zumbo, "Relative power of the Wilcoxon test, the Friedman test, and repeated-measures ANOVA on ranks", Journal of Experimental Education, vol. 62, no. 1, pp. 75-86, 1993.

[39] J.D. Pujari, R. Yakkundimath, A.S. Byadgi, "SVM and ANN based classification of plant diseases using feature reduction technique", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 3, no. 7, pp. 6-14, 2016.

[40] W. Gevaert, G. Tsenov, V. Mladenov, "Neural Networks used for Speech Recognition", Journal of Automatic Control. Vol. 20, pp. 1-7, 2010.

[41] S. Russell, P. Norvig, "Artificial intelligence: a modern approach", 2nd edn. Prentice Hall, Englewood Cliffs. ISBN 978-0137903955, 2003.

[42] M. Islam, A. Roy, R. H. Laskar, "SVM-based robust image watermarking technique in LWT domain using different sub-bands," Neural Computing and Applications, vol. 32, pp. 1379-1403, 2020, doi: https://doi.org/10.1007/s00521-018-3647-2.

[43] A. Jain and G. N. Rathna, "Visual Speech Recognition for Isolated Digits Using Discrete Cosine Transform and Local Binary Patterns Features", 978-1-5090-5990-4/17/, 2017 IEEE.

[44] G. Zhao, M. Barnard, M. Pietikainen, "Lipreading with local spatiotemporal descriptors", IEEE Transactions on Multimedia, vol. 11, no. 7, 1254-1265, 2009.

[45] A. Kandagal, V. Udayashankara, "Visual Speech Recognition Based on Lip Movement for Indian Languages", 2017.

[46] N. Saleem, M. Khattak, E.Verdú, "On Improvement of Speech Intelligibility and Quality: A Survey of Unsupervised Single Channel Speech Enhancement Algorithms", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 2, pp. 78-89, 2020, doi:10.9781/ijimai.2019.12.001.

[47] N. Saleem, M. Khattak, "Deep Neural Networks for Speech Enhancement in Complex-Noisy Environments", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 1, pp. 84-90, 2020, doi: 10.9781/ijimai.2019.06.001.

[48] M. Z. Ibrahim, D. J. Mulvaney, "Robust geometrical-based lip-reading using Hidden Markov models", Eurocon 2013, pp. 2011-2016.

[49] E. K. Patterson, S. Gurbuz, Z. Tüfekci, J.N. Gowdy, "CUAVE: A new audio-visual database for multmodal human-computer interface research", 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002, pp. II-2017-II-2020, doi: 10.1109/ICASSP.2002.5745028.

Saswati Debnath

Saswati Debnath received the B.Tech. degree in Computer Science and Engineering from Government Women Engineering College, Ajmer, Rajasthan Technical University India in 2013 and the M.Tech. degree in Computer Science and Engineering from National Institute of Technology, Silchar, Assam, India in 2015. She received Ph.D. from National Institute of Technology, Silchar, Assam, India in 2020.

Pinki Roy

Dr. Pinki Roy received the B.Tech. and M.Tech. degrees in Computer Science and Engineering from Dr. Babasaheb Ambedkar Technological University, Lonere, Maharastra, India, and the Ph.D. in Computer Science and Engineering from National Institute of Technology, Silchar, Assam,India. Currently, she is an assistant professor with National Institute of Technology, Silchar, Assam, India.