AVSR: AN APPROACH FOR KANNADA LANGUAGE USING HYBRID MODEL

[1]Sujatha H, [2]Shankara C, [3]Basavaraju M
[1]Lecturer, [2]Lecturer, [3]Selection Grade Lecturer
[1,2,3]Department of Electronics & Communication Engineering,
[1,2]Government Polytechnic, Nagamangala, Karnataka, India
[3]Government C.P.C Polytechnic, Mysore, Karnataka, India

Abstract: Communication involves conveying thoughts and ideas to others through speech and facial expressions. It is a vital aspect of human life, facilitating interaction with the world. However, for individuals with hearing impairments, communication can be challenging without the aid of devices like battery-operated hearing aids, hand signals, and facial expressions. In this project, an Audio-Visual Speech Recognition (AVSR) system is introduced, leveraging image-processing capabilities in lip-reading to support speech recognition systems. This approach integrates both audio and visual processing, meaning that separate video and audio recognition processes are combined. Our project utilizes a Kannada audio database—a Dravidian language spoken by over 60 million people in Karnataka—to assist those with hearing impairments. The proposed AVSR model comprises three main components: (a) an audio feature extraction mechanism, (b) a visual feature extraction mechanism, and (c) an audio and visual feature integration mechanism. The integration mechanism merges the outputs from both the audio and visual extraction processes to produce final classification results. In this work, the audio feature mechanism is designed using Mel-frequency Cepstral Coefficients (MFCC) and a Random Forest system, while the visual feature mechanism is developed using Haar-Cascade Detection with MATLAB and a Feedforward Neural Network system. These extracted features are then integrated using a hybrid model that includes both Random Forest and Artificial Neural Network (ANN) techniques. The system's performance, in terms of speech recognition accuracy and robustness, was tested using speech data in a clean environment. On average, the final audio recognition rate is 88%, the visual recognition rate is 57%, and the integrated system achieves an overall recognition rate of 86%.

Index Terms – Hybrid Model, Random Forest, AVSR, Feature Extraction, Lip Localization, Word Recognition.

I. INTRODUCTION

In recent trends, audio-visual speech recognition has become a significant area of focus, enabling computers to edge closer to artificial intelligence by imitating human abilities to understand words and objects. Compared to other recognition systems, audio-visual speech recognition is more efficient, beneficial, and robust, making it a crucial component of human-machine interfaces. Currently, lip reading plays a vital role in recognition systems, where various techniques are employed to enhance model performance. The history of lip-reading dates back to 1954, with the first related work. A notable advancement came from the University of Illinois by Petajan in the 1980s, which became a recognized method. Since then, extensive research has been conducted in this field. Given that audio signals are susceptible to noise, pixel-based methods and artificial neural networks (ANN) were utilized for lip reading in 1989. In 1993, Hidden Markov Models (HMMs) were introduced to lip-reading systems for sentence recognition, achieving an accuracy rate of approximately 25%. The process involves several steps: the first step is detecting the speaker's face and identifying the lip region of interest. The second step is reducing the image data and extracting lip features. The third step involves recognizing the visual data from the extracted lip features and classifying it using a highly efficient classifier.

1.1 Face Detection

Face detection can be visualized as a computer vision task that involves locating faces within an image. It serves as the foundational step in face biometrics and is known for its high efficiency in performance. Predicting words through lip reading is a challenging task for humans, which is why feature-based classifier techniques are employed to address it. Currently, deep learning methods have achieved impressive results on standard face detection datasets.

Face detection is considered the essential step in numerous face-oriented technologies, such as face identification and recognition, and has many beneficial applications. It is a form of object class recognition, where the positions and sizes of all faces in an image are determined. Face detection algorithms focus primarily

on human faces, similar to image recognition, where each part of the image is analyzed and compared. Any changes in facial features within the database can affect the validity of the comparison.

1.2 Lip Localization

Lip-area extraction is a crucial region of interest in improving recognition rates. Various methods have been proposed for extracting facial images from the face, with the Viola-Jones algorithm being a notable example. The Viola-Jones algorithm is object-oriented, converting the image into grayscale to reduce the amount of data to process. It initially detects the face in the grayscale image and then identifies the location in the colored image. The algorithm uses multiple small steps, where several boxes detect face-like features (Haar-like features). The combined data from these boxes helps the algorithm determine the exact location of the face within an image. Similarly, this process is applied to detect the lip region. Direct detection of the lip region without first identifying the face is challenging for the algorithm due to the presence of other facial features such as the nose, ears, and mustache.

1.3 Feature Extraction and Recognition Models

Feature extraction was performed using the Viola-Jones algorithm, effectively identifying the appropriate features of the lips for visual recognition. Some of the extracted features included brightness, contrast, correlation, entropy, mean, variance, and standard deviation. These features were then trained and tested using a Feed Forward Neural Network (FFNN). Similarly, audio features like Mel-Frequency Cepstral Coefficients (MFCC) and Zero Crossing Rate (ZCR) were successfully extracted and trained using a Random Forest (RF).

For modeling, the Random Forest (RF) was used for audio processing, while the Feed Forward Neural Network (FFNN) was employed for visual processing. Hybrid models, which combine two or more methods to interpret and analyze data, were utilized due to their superior accuracy. Thus, a hybrid model combining Random Forest (RF) and Artificial Neural Network (ANN) was used to integrate audio and visual features, resulting in improved overall performance.

1.4 AVSR

Audio-visual speech recognition (AVSR) is a technique that leverages image processing for lip reading to enhance speech recognition systems. As illustrated in Figure 1.1, the fundamental working of an AVSR system emphasizes the integration of audio and video processing of speech signals. The AVSR system comprises two main components: audio processing and video processing.

Audio processing involves feature extraction and recognition, where specific audio features are identified and analyzed. Video processing includes face detection, lip localization, feature extraction, and recognition. By combining these audio and video processing techniques, the AVSR system aims to improve the accuracy and robustness of speech recognition.

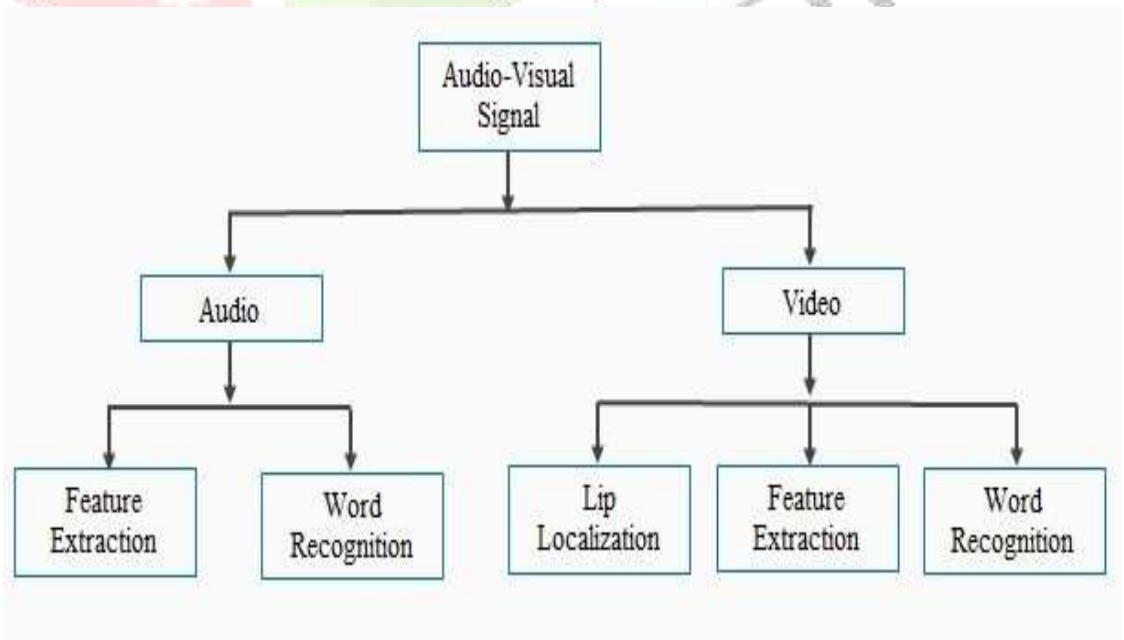


Figure 1.1. AVSR System

II. RELATED WORK

The existing AVSR systems are limited to relatively controlled environments and typically aim for better quality by focusing solely on the acoustic channel. An alternative method, which has shown some success, involves using visual features extracted from the movement of the speaker's mouth to improve recognition rates. This method is known as Visual Speech Recognition (VSR).

The proposed system demonstrates VSR for Indian languages using lip parameters. Effective preprocessing steps, such as denoising and resizing, are employed, along with the Canny edge detection algorithm to identify the true edges of the input image for region of interest (ROI) extraction. Four major features—entropy, energy, contrast, and correlation—are extracted using the Gray Level Co-occurrence Matrix (GLCM) algorithm and classified accurately with an Artificial Neural Network (ANN) classifier.

The proposed system's performance is more accurate compared to conventional methods, achieving a 90% accuracy rate. However, a notable drawback is that combining both audio and video input parameters could further enhance the performance of visual speech recognition, potentially increasing the overall accuracy[1].

New visual features for the audio-visual speech recognition system are extracted from side face images of people, focusing on the lip angle between the upper and lower lip lines and its delta. Significant improvements were noted across all SNR conditions. When combined with the proposed optical-flow features, recognition accuracies further improved. The enhancement was confirmed by using the visual features in conjunction with MLLR-based noise adaptation applied to the audio HMM, achieving an overall accuracy of 87% at 20dB. However, the evaluation could benefit from more general recognition methods, and an optimization method for stream weights could enhance the combination of lip angle and optical flow features [2].

Hybrid speech recognition systems outperform simple HMM-based ones, utilizing SVM as an alternative to neural networks. The integration of SVM and HMM successfully processes multimodal data by employing multi-class SVM classifiers with probabilistic outputs—one for feature fusion and two for decision fusion. Video accuracy stands at around 80% individually, but when combined, the accuracy increases to 91%. This demonstrates that combining audio and visual information results in better performance than using either modality alone, proving their complementarity [3].

An AVSR system uses an automatic lip-geometric-based feature extractor applied to the Mandarin audio-visual database with the WDKNN classifier. The proposed method involves detecting automatic lip feature extraction for Mandarin audio-visual speech recognition. The algorithm first segments and locates the mouth region using normalized RGB space from a color video sequence. Subsequently, it extracts the lip segments from the background using both color and edge information. Key lip points defining lip position are detected, and relevant visual speech parameters are derived to form the input dataset for the recognition engine. Improvements were made by utilizing both geometric and motion visual features, enhancing visual front-end accuracy compared to previous studies that only used lip geometric features. Under 15 dB noisy conditions, accuracy ranges were 68.7% to 88.7% for geometric features, 64.4% to 85.5% for motion features, and 83.2% to 94.3% for combined geometric and motion conditions [4].

Localizing the lip region for Myanmar consonant recognition involves processing lip motion sequences from video input streams. The technique effectively localizes lip movements and achieves an accuracy of 87.6% using CIELAB color transformation, Moore neighborhood tracing algorithm, and a linear SVM classifier. Despite these results, the method could benefit from a more advanced feature recognition algorithm such as SVM for improved accuracy [5].

In visual speech recognition with sequences of colored images of variable lengths, the preprocessing method involves concatenating the first k images of each sequence into a 2D grid, classified by a VGGNET pretrained on faces. A smaller model was then trained on these concatenated images, and the final model incorporated multiple LSTM layers to handle variable-length sequences. While the pretrained VGGNET performed well, the expected performance of the LSTM model was not fully realized, achieving an accuracy of 76%. Potential improvements could include using recurrent networks and addressing the lengthy training times associated with updating VGGNET [6].

A novel visual feature extraction approach for a VSR system employs a CNN to recognize phonemes from mouth area image sequences in video streams. This method achieves 58% recognition accuracy by classifying phoneme sequences collected from six speakers. The system utilizes speaker-dependent models for phoneme recognition and a common model for isolated word recognition, with an overall accuracy of 60%. However, it struggles with larger datasets and increasing the number of speakers [7].

Feature sets for a speaker-independent lip-reading system were evaluated using GMM-HMM and DNN-HMM hybrid models. The system tested geometric shape features of the lips, resulting in a 20.4% reduction in error rate. The introduction of DBNFS features, combining the advantages of previous sets, showed a 15.4% improvement in accuracy, reaching 79.6% [8].

A multi-stream asynchronous DBN model for audio-visual speech recognition handles the asynchrony between audio and visual streams at the word level. This model, using Gaussian Mixture Models, compares favorably with multi-stream HMMs by addressing asynchrony limitations. The system achieves a 75% accuracy across an SNR range of 0 dB to 15 dB. However, the drawback is the high processing time required for large vocabulary recognition [9].

In an AVSR system utilizing dynamic weighting schemes, weights control the contribution of each stream to the recognition task. The system, using HMM-ANNs or HMM-GMMs models, achieved 87% accuracy based on transition probabilities common to both HMM architectures. For high and medium SNRs, accuracy is about 72%. The system's main drawback is confusion with the silence class, which hinders performance; thus, the weights need to be adjusted to minimize these confusions and focus on accurate speech recognition [10].

The study presents a speaker identification system that utilizes Random Forest (RF) classifiers, with Mel Frequency Cepstral Coefficients (MFCC) and Relative Phase Space (RPS) as key feature extraction methods. This system, applicable in fields like biometric authentication, security, forensics, and human-machine interaction, details the entire process from preprocessing speech files to final speaker identification. The RF classifier achieved an 80% accuracy rate, demonstrating that spectral features like MFCC are highly effective and challenging to surpass with novel alternatives [11].

The paper explores speaker identification through voice features and acoustics, emphasizing its relevance in Human-Computer Interaction (HCI), biometrics, security, and the Internet of Things (IoT). It evaluates various preprocessing, feature extraction, and machine learning techniques on audio recorded in natural, unconstrained settings to find the most effective combinations for speaker recognition. By applying preprocessing methods such as trimming, splitting, noise reduction, and vocal enhancements, and extracting Mel Frequency Cepstral Coefficients (MFCC), the study uses a Random Forest algorithm to achieve optimal classification performance. The results demonstrated a 5.41% increase in F1 scores when models were trained with MFCC coefficients alone, compared to those using MFCC delta-delta coefficients [12].

The study focuses on Tamil speech recognition, where words are segmented into syllables using short-term energy (STE) to minimize corpus size. The syllable segmentation algorithm employs STE on the continuous speech signal. For feature extraction, a combination of Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC) is used. MFCC captures detailed linguistic features, while LPC effectively estimates vocal tract parameters and reduces speech bit rate, enhancing signal transmission efficiency. Evaluated with a Random Forest classifier, this combined approach achieves an improved recognition accuracy of 82%, demonstrating the effectiveness of integrating MFCC and LPC for better performance [13].

The paper presents an innovative near real-time speaker recognition architecture that improves performance by combining multiple feature extraction techniques, including Gabor Filters (GF), Convolutional Neural Networks (CNN), and statistical parameters into a unified matrix. This system is designed for secure voice-based user interface access and integrates with existing Natural Language Processing (NLP) systems. The study explores various feature extraction methods and evaluates classifiers like Support Vector Machine (SVM), Random Forest (RF), and Deep Neural Network (DNN). The results indicate that the hybrid feature extraction methods and the RF classifier are most effective for speaker recognition, achieving an accuracy of 75% [14].

The study introduces a new CNN-RF network model that integrates Convolutional Neural Networks (CNN) with Random Forest (RF) to enhance speech emotion recognition. In this model, the CNN serves as the feature extractor to derive emotion features from normalized spectrograms, which are then classified using the RF algorithm. The CNN-RF model outperforms traditional CNN approaches in recognizing emotional states from speech. Additionally, the model was implemented in NAO robots to improve their ability to interpret human emotions such as happiness, anger, sadness, and joy. This advancement enables more sophisticated human-computer interactions by allowing NAO robots to better understand and respond to human emotional cues. The enhanced Record Sound command box, developed as part of this study, successfully demonstrates the effectiveness of the CNN-RF model in practical applications [15].

This comprehensive study evaluates the use of Artificial Neural Networks (ANN) in speech recognition, examining various neural network methods and their relative advantages and disadvantages. The paper highlights the importance of preprocessing audio signals to enhance recognition accuracy by eliminating irrelevant variations through noise filtering, smoothing, end point detection, framing, windowing, reverberation cancellation, and echo removal. Feature extraction techniques, such as Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC), are discussed for their robustness and reliability in handling speaker variations and environmental conditions. The study concludes that Recurrent Neural Networks (RNN) offer superior speech recognition rates compared to Multi-Layer Perceptrons (MLP), although RNNs involve more complex and dynamically sensitive training algorithms that can pose challenges [16].

This study explores the recognition of individual speakers' voices by analyzing their continuous speech waveform distributions using a combined approach of artificial neural networks (ANN) and Gaussian mixture models (GMM). A feed-forward multilayer ANN with 30 hidden neurons was employed for discriminative classification, while GMMs provided scores to match speech features. The decision-making system utilized correlation coefficient analysis to evaluate how well the speech features from the ANN and GMM matched. Experiments with speech utterances from 30 speakers (20 males and 10 females) revealed average recognition rates of 77% for 5-word utterances and 43% for 20-word utterances with trained speech. For unknown utterances, the recognition rate dropped to 18% for 20-word utterances [17].

This study evaluates the effectiveness of Decision Trees (DT) and Convolutional Neural Networks (CNN) as classifiers for emotion recognition from English and Kannada audio data. The analysis involved extracting Mel-Frequency Cepstral Coefficients (MFCC) features from the audio signals and training, testing, and evaluating models with various parameters. The research finds that CNN outperforms DT in recognizing emotions, achieving a classification accuracy of up to 72%, compared to 38% and 52% for DT with fewer emotions. The paper concludes that CNN is more effective for emotion classification in speech data, demonstrating its superior performance over DT in this context. The Speech Emotion Recognition system has potential applications in psychiatric diagnosis, lie detection, call center interactions, customer feedback, and voice messaging [18].

This paper investigates the differentiation of English alphabet sets (E-set to AH-set) and phonemes using neural network techniques. It specifically employs Recurrent Neural Networks (RNNs) to analyze these differences and enhance understanding of phoneme and word recognition. The study utilized RNNs and backpropagation through Multilayer Perceptron (MLP) to train models with speech data from six speakers (both male and female) in a controlled environment. The primary goal is to decode the linguistic message from speech signals into text by identifying sound sequences. The research highlights the effectiveness of Mel-Frequency analysis for feature extraction and demonstrates that RNNs outperform MLPs in classifying speech signals, even with simplified models and a limited character set [19].

Speech recognition systems convert spoken language into text using a series of stages and technologies. The process begins with capturing speech through a microphone in a quiet environment, where the audio is stored in a database. The next step involves pre-processing, which segments the audio into 20-millisecond chunks for easier handling. These chunks are then converted into a digital format and inputted into a Recurrent Neural Network (RNN), the core model for speech recognition in this system. The effectiveness of the Speech-to-Text (STT) engine relies heavily on the quality of sampling and pre-processing. The RNN, equipped with memory, uses previous predictions to influence future ones, making it well-suited for processing sequential data like speech. This paper highlights RNNs as superior to Multilayer Perceptrons (MLPs) in handling speech signals, although their training remains complex. RNNs are identified as highly effective for voice-controlled technologies [20].

This paper presents an advanced audio-visual fusion strategy that enhances recognition accuracy by automatically aligning audio and visual modalities, rather than merely concatenating features. Tested on the TCD-TIMIT and LRS2 datasets, which are used for large vocabulary continuous speech recognition, the method was evaluated under various noise conditions. The approach leverages state-of-the-art Sequence-to-Sequence (Seq2Seq) architectures, demonstrating that it can significantly improve performance. Results showed relative accuracy gains ranging from 7% to 30% over acoustic-only models, depending on the noise level. The fusion strategy not only promises broad applicability across multimodal tasks but also offers a straightforward implementation, easily adaptable to attention-based Seq2Seq models using existing code. This makes it a flexible and effective tool for integrating multimodal data in various applications [21].

Audio-Visual Speech Recognition (AVSR) leverages both visual information, such as lip movements, and acoustic signals to enhance recognition, particularly in noisy environments. Unlike acoustic signals, visual signals remain unaffected by noise, making them a valuable resource for improving speech recognition performance under challenging conditions. The process begins with recording audio and visual signals using a microphone and camera, respectively. Salient features are then extracted from each modality. The two types of information can be integrated in two primary ways: feature fusion and decision fusion. Feature fusion involves concatenating features from both modalities into a single composite vector, which is used by a classifier for recognition. Decision fusion, on the other hand, involves separate recognition processes for each modality, with the final decision made by combining the outputs of the individual classifiers. The paper describes a neural network-based fusion method that effectively uses the reliability of both modalities to achieve robust performance across various noise conditions [22].

III. PROPOSED METHODOLOGY

The design flow for the Audio-Visual Speech Recognition (AVSR) system, illustrated in Figure 3.1, includes both processing and recognition stages to convert audio-visual signals into text. The visual component focuses on lip reading, involving three key steps: lip localization, feature extraction, and recognition, all of which require neural network-based training and testing. In contrast, the audio component involves feature extraction and classification, with audio recognition generally outperforming visual recognition. Figure 3.2 details the audio-visual processing workflow. The audio processing begins with pre-processing to clean the signal by removing background noise and unwanted silence, and trimming audio samples to 1 second. Spectral subtraction is used to estimate and remove noise from the speech signal. In the feature extraction phase, Mel-frequency cepstral coefficients (MFCCs), introduced by Davis and Mermelstein in the 1980s, are computed to capture spectral features of the audio signal. Classification is performed using a Random Forest Classifier, which employs 120 decision trees to determine the output. The final prediction is achieved by comparing the test dataset features with the trained model to produce the text output.

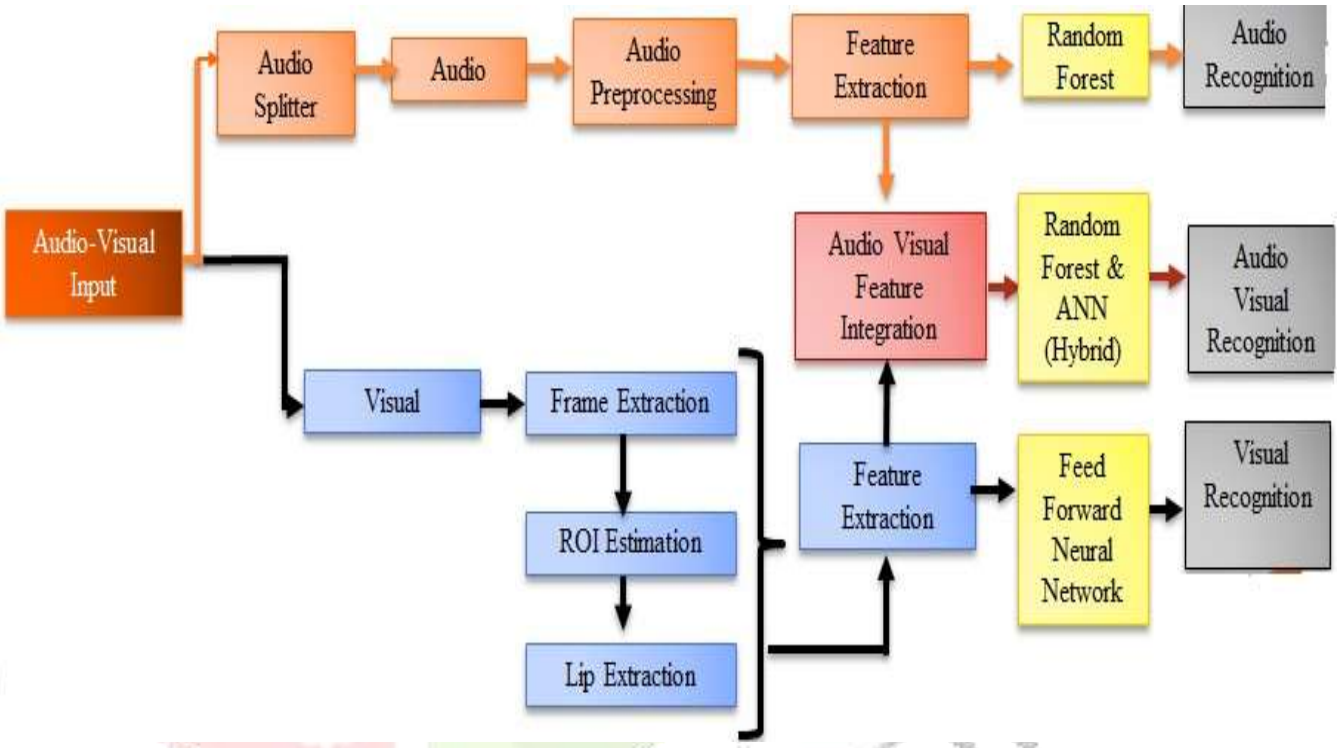


Figure 3.1. Block Diagram of proposed system

3.1 Audio model

The implementation of audio recognition for the Kannada dataset was carried out using Matlab 2018b and involves several key steps. The process is divided into training and testing phases. Initially, the dataset is imported as a .wav file, and the audio data is read. The first phase, preprocessing, focuses on eliminating unwanted elements like background noise. Next, feature extraction is performed using the Mel-Frequency Cepstral Coefficients (MFCC) method to obtain feature vectors that represent distinct characteristics of the speaker, such as pitch and amplitude. The dataset is then split into training and testing sets, with 80% used for training and 20% for validation. The MFCC features from the training set are classified using a Random Forest (RF) classifier, and the trained model is saved as a .mat file. The RF algorithm, known for its effectiveness in handling large datasets and mitigating overfitting through averaging predictions from multiple decision trees, has been identified as one of the top-performing classifiers among 179 tested models in terms of accuracy and precision. For testing, the model is evaluated against the test set, and performance metrics such as the error function and confusion matrix are plotted. Additionally, a function is defined to predict text from the given audio.

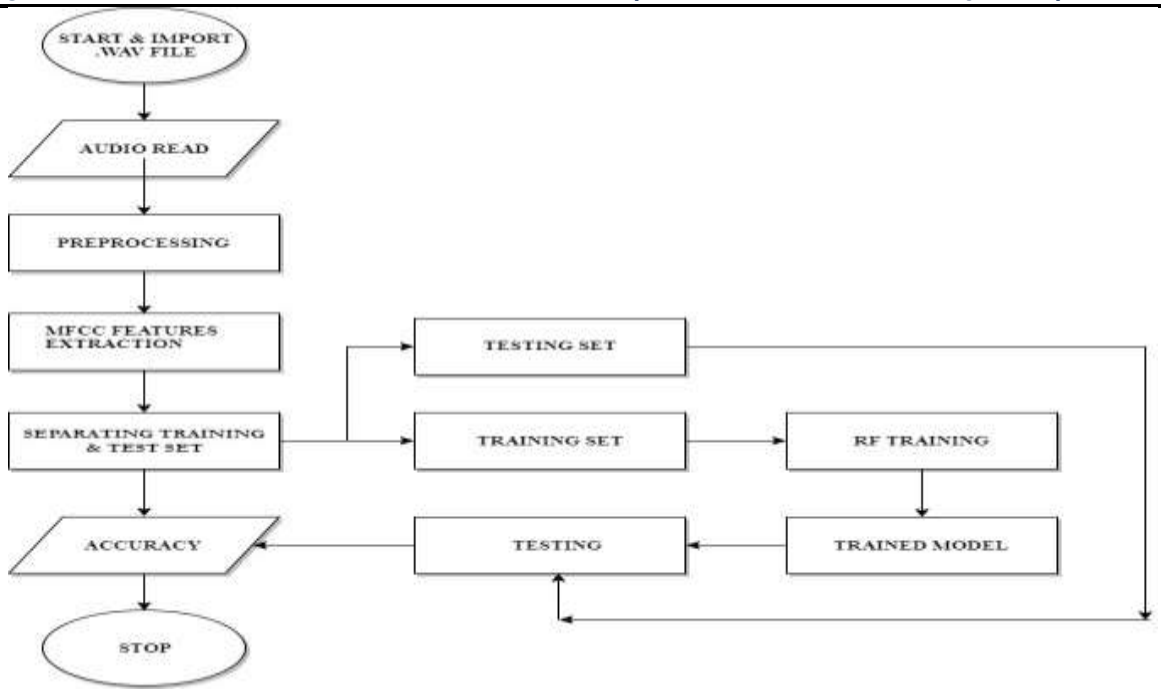


Figure 3.2. Simple block diagram of audio recognition part showing all implementation

3.2 Random Forest

Random forests are an ensemble learning technique that aggregates multiple decision trees to improve classification and prediction performance. These forests are created by growing a collection of decision trees from random subsets of the data, leveraging the diversity among them to enhance overall accuracy. Decision trees, the core component of this method, are well-suited for ensemble approaches due to their tendency to have high variance but low bias. In a decision tree, each node represents a test on a particular attribute, each branch signifies the outcome of this test, and each leaf node (terminal node) corresponds to a class label, as illustrated in Figure 3.3. The ensemble method of random forests benefits from the averaging of these trees' predictions, which mitigates overfitting and improves generalization.

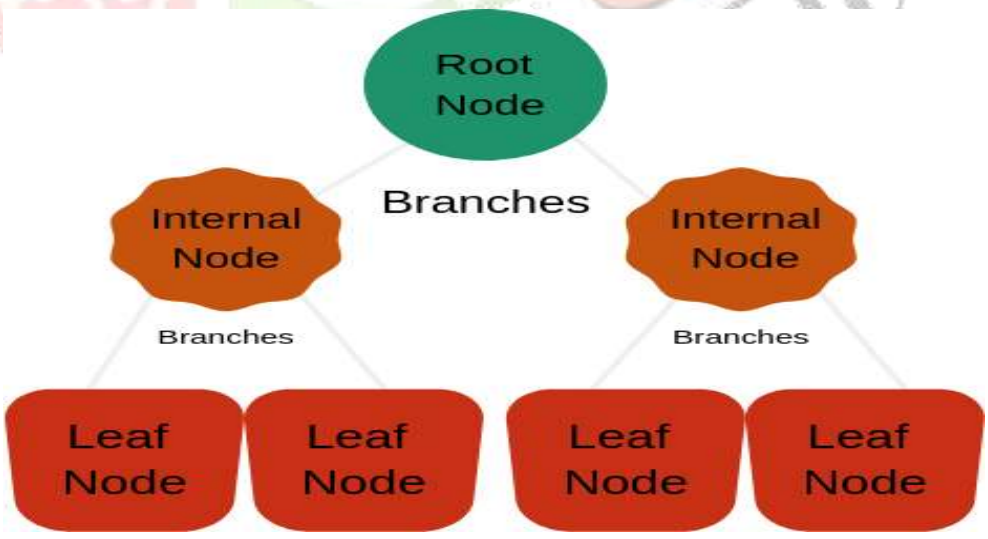


Figure 3.3. Representation of Decision Tree

3.3 Video Model

The initial and crucial step in processing video signals involves importing all essential libraries into Matlab. After the libraries are successfully loaded, the next task is to organize the dataset appropriately. This involves creating two separate directories for training and testing data. These directories facilitate dynamic access to the video samples stored in each. Following data organization, the focus shifts to lip region identification. This step uses the Viola-Jones Algorithm to extract the lip region from the video frames. The extracted lip regions are then

used to train a Feedforward Neural Network (FFNN) algorithm. Figure 3.4 illustrates the directory structure containing the training and testing files.

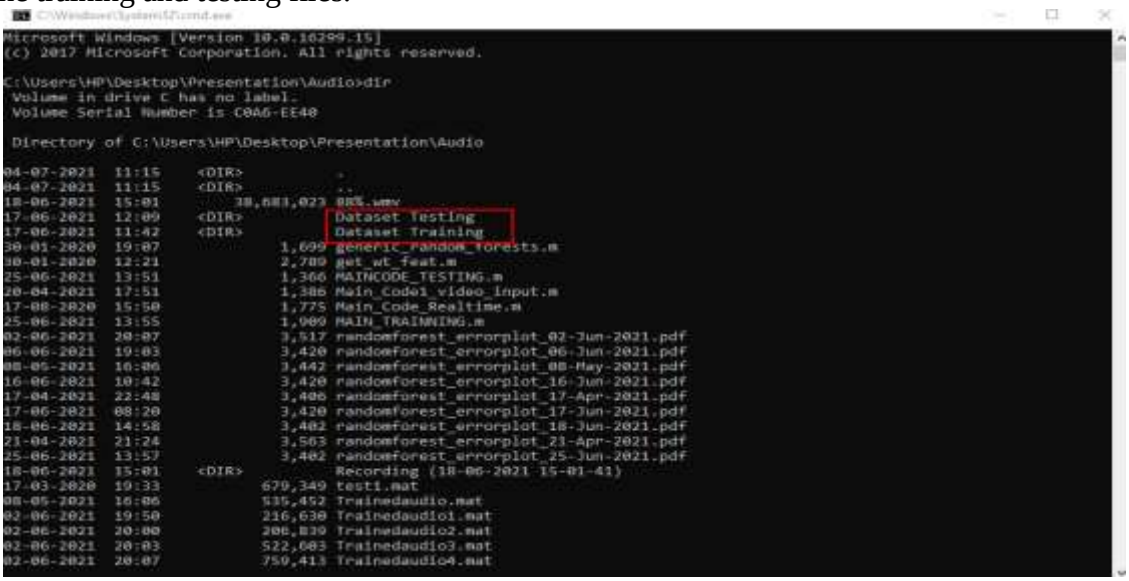


Figure 3.4. Directory containing both training and testing files

3.4 FFNN (Feed Forward Neural Network)

Feedforward neural networks are widely utilized in machine learning, particularly in image processing tasks. These networks are termed "feedforward" because data flows exclusively in one direction—from input nodes through hidden layers to output nodes. Unlike recurrent networks, feedforward networks do not have connections that allow information to loop back from the output to the network's earlier layers. This unidirectional flow ensures that the network processes inputs in a straightforward manner, as depicted in Figure 3.5. The activation of nodes in a feedforward network is typically governed by functions such as the sigmoid function, illustrated in Figure 3.6.

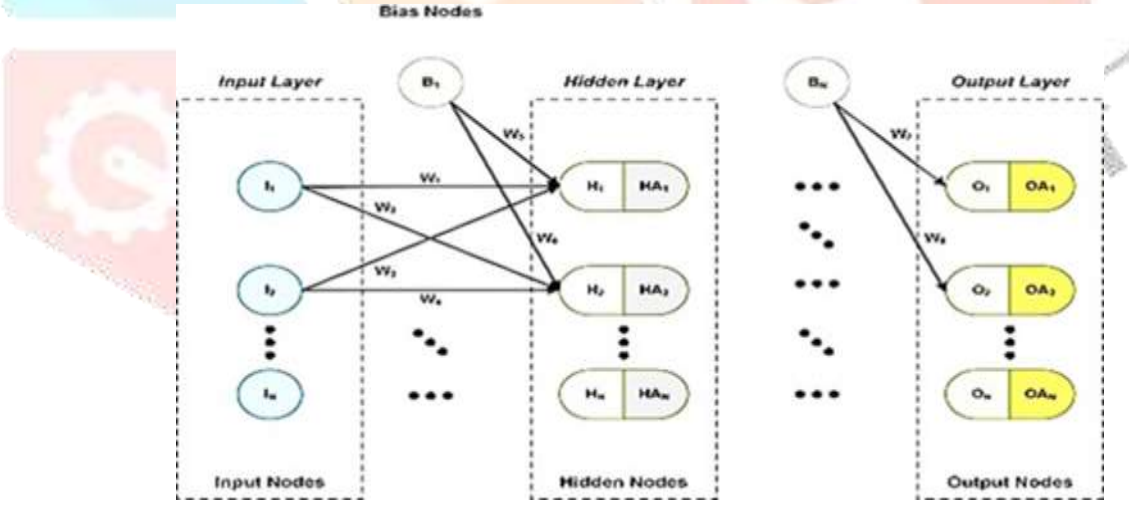


Figure 3.5. Feed Forward Neural Network

Where,
W: The weight of a connection.
I: Input node (the neural network input).
H: Hidden node (a weighted sum of input layers or previous hidden layers). Hidden node activated (the value of the hidden node passed to a predefined function).
O: Output node (A weighted sum of the last hidden layer).
OA: Output node activated (the neural network output, the value of an output node passed to a predefined function).
B: Bias node (always a contrast, typically set equal to 1.0).
e: Total difference between the output of the network and the desired value(s) (total error is typically measured by estimators such as mean squared error, entropy, etc.)

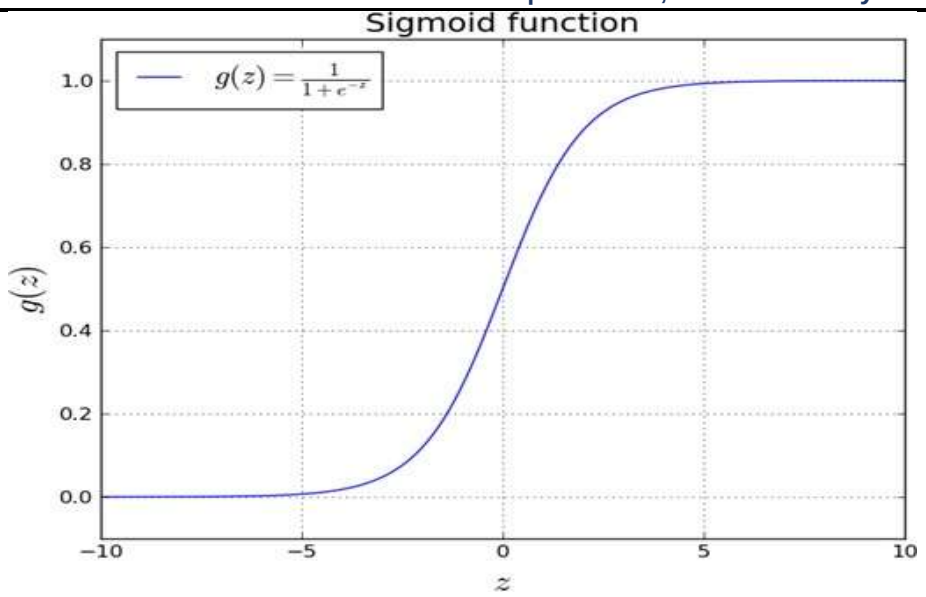


Figure 3.6. Presentation of sigmoid activation function

3.5 Artificial Neural Network

Artificial Neural Networks (ANNs) are computational frameworks designed to emulate the functioning of biological neural systems. Modeled after the nervous systems of animals, ANNs excel in tasks such as machine learning and pattern recognition. These networks consist of interconnected nodes, or "neurons," that process input data through a series of weighted connections, similar to biological synapses and dendrites. Each node applies an activation function to the values received, influencing the output of the network. Neural networks, depicted in Figure 3.7, operate as directed graphs with nodes connected by weighted arcs. Beyond classification tasks, ANNs are also adept at regression for predicting continuous variables. Their versatility makes them valuable in various fields, including economics and forensics, where they can handle large datasets and complex pattern recognition challenges.

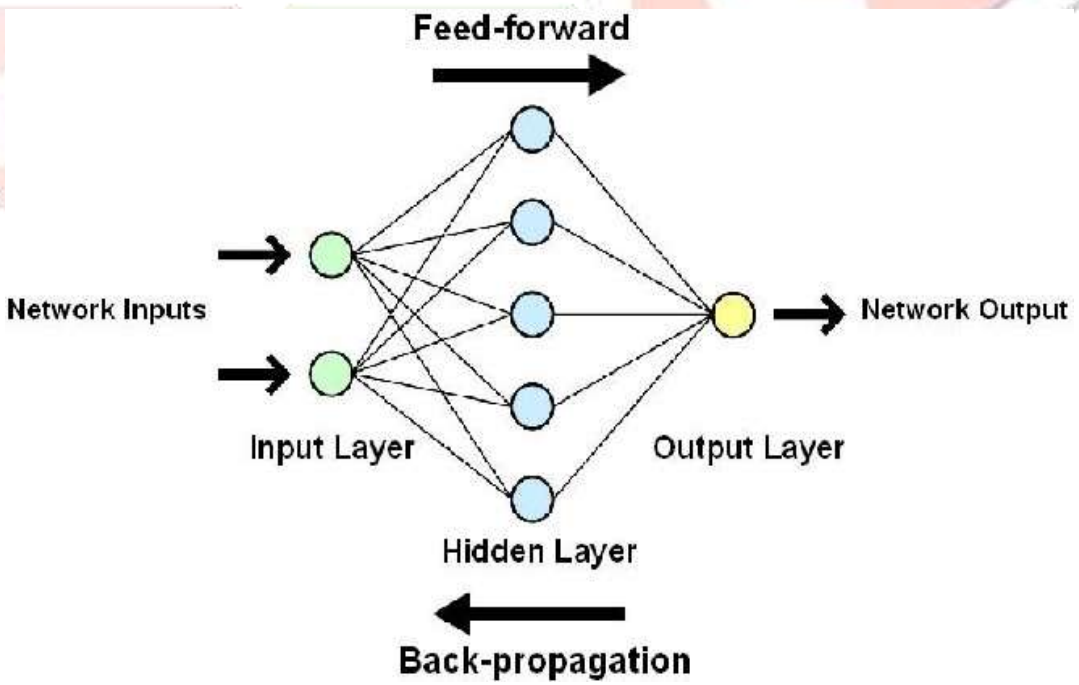


Figure 3.7. Representation of Artificial Neural Network

IV. RESULTS AND DISCUSSION

4.1 Matlab

MATLAB (an abbreviation *of "matrix laboratory*") is a proprietary multi- paradigm programming language and numeric computing environment developed by MathWorks. MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages. Although MATLAB is intended primarily for numeric computing, an optional toolbox uses the MuPAD symbolic engine allowing access to symbolic computing abilities. An additional package, Simulink, adds graphical multi-domain simulation and model-based design for embedded systems. The MAtlab libraries (AddOns) which are used in this project are; DeepLearning Toolbox, Machine learning toolbox, Image processing toolbox, Audio Processing Toolbox and Signal Processing Toolbox

4.2 Audio Model

The audio recognition was conducted separately for 5 different Kannada words like" Guruthu", "Howdu","Kelida","Illa" and "Neenu" and the results are compared original dataset as shown in Figure 4.1.

Totally 400 datasets have been used out of which 250 is used for training and 150 is used for the testing.

Figure 4.1: Five Different Kannada words

A total number of 30 samples for each word has been collected that could be seen in the below Figure 4.2

Figure 4.2. Total 30 samples per word

An accuracy of 80.5% was achieved with respect to Kannada words recognition with the ANN algorithm which was satisfactory so another algorithm was used which made the recognition accuracy better, Random Forest algorithm gave an accuracy of 88% which could be seen in Figure 4.3, corresponding confusion plot and predicted word could be seen in Figure 4.3.

Figure 4.3. Accuracy of ANN and Random Forest are shown

4.2.1 Confusion Plot

It can be concluded that Random Forest may be used for audio recognition to achieve good recognition rates or accuracy of around 90% and better performance. The confusion matrix depicted in Figure 4.4 shows the graph of predicted class v/s the true class. The diagonal elements represents the correctly classified words.

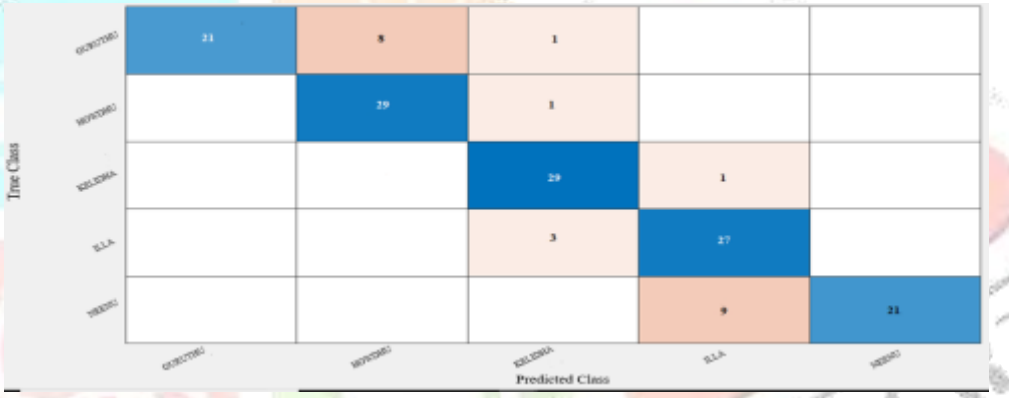


Figure 4.4. Confusion plot of audio processing

4.2.2 Random Forest Error Plot

It is a plot of Number of trees grown v/s Out-of-bag classification error. Random forest error plot finding - As the number of trees increases the error rate decreases. Total of 120 decision trees are grown for the minimum error rate as shown in Figure 4.5.

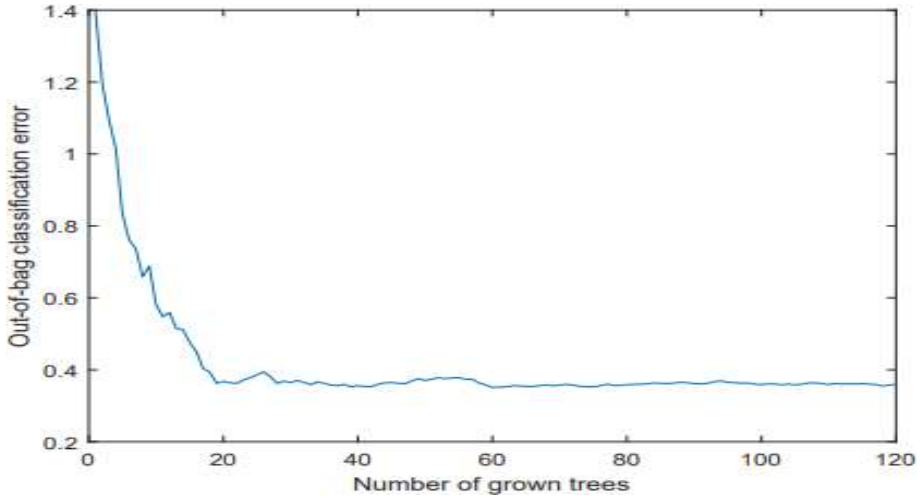


Figure 4.5. Random forest error plot of audio processing

4.2.3 Validation Plot

The performance plot represents the variation of number of Epochs v/s the mean squared error. The best validation performance is 0.9345 at epoch 1. As the number of epochs increases the error rate decreases as shown in Figure 4.6.

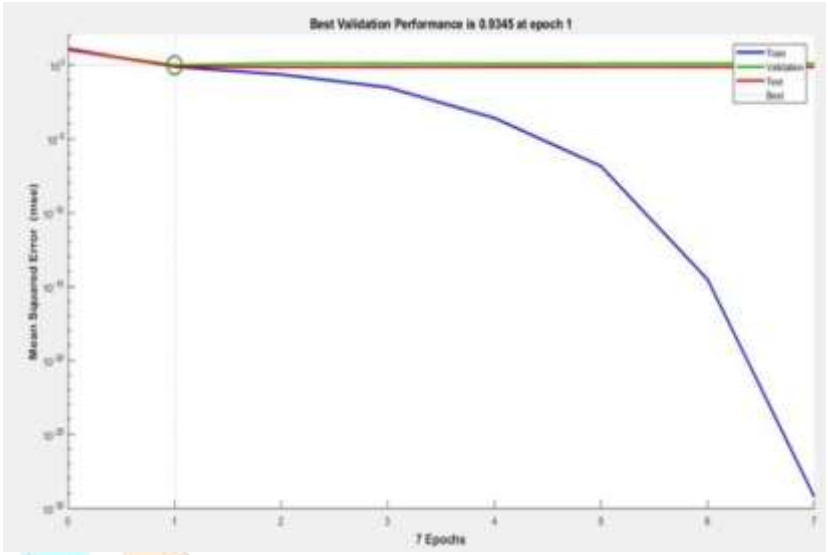


Figure 4.6. Performance validation plot of audio processing

4.2.4 Gradient Descent

Gradient Descent is an optimization technique used to locate a local minimum of a differentiable function. In machine learning, it is employed to determine the parameter values (coefficients) that minimize a cost function as effectively as possible. Figure 4.7 illustrates how the gradient, Mu, and validation checks vary with the number of epochs.

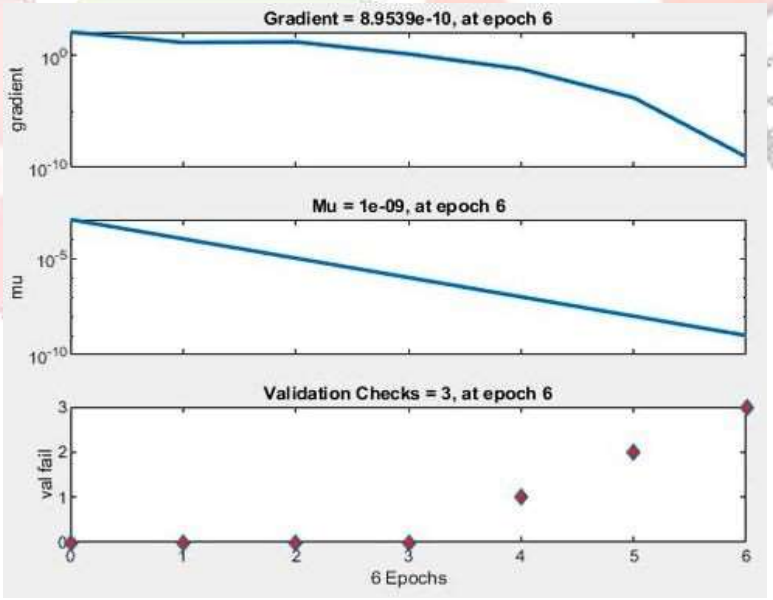


Figure 4.7. Gradient Descent plot of audio processing

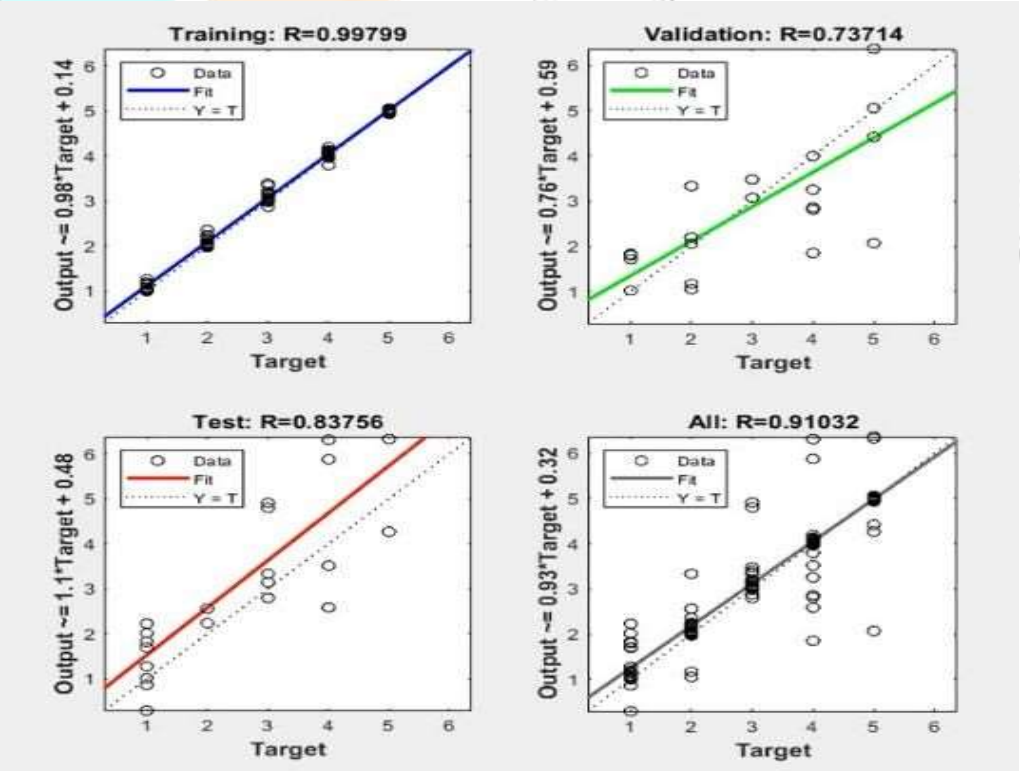
The gradient obtained is 0.00159 at epoch 7. The Mu is 1e-08 at epoch 7. The validation Checks obtained is 6 at epoch 7.

4.2.5 Regression plot of training, testing, validation and overall

Figure 4.8 displays a regression plot that depicts the linear relationship between an independent and a dependent variable, providing a visual representation of the relationship's strength and the dispersion of results. Meanwhile, Figure 4.9 presents a comparative table of the audio processing results.

Method	MFCC, ANN,GMM [1]	SSVAD (Spectral Subtraction with voice activity detection), MMSESPZC (Min. Mean Square Error Spectrum power Estimator based On zero crossing)[2]	LSTM (Long Short term memory) & CNN (Convolution Neural Network) [3]	MFCC & ZCR for feature extraction, Random Forest method for classifier
Accuracy	77%	82.36%	82%	88%
Dataset	Local Dataset & an MySQL Dataset	TIMIT Speech Database	TCD-TIMIT & LRS2	Custom

Figure 4.8: Regression plot of audio processing



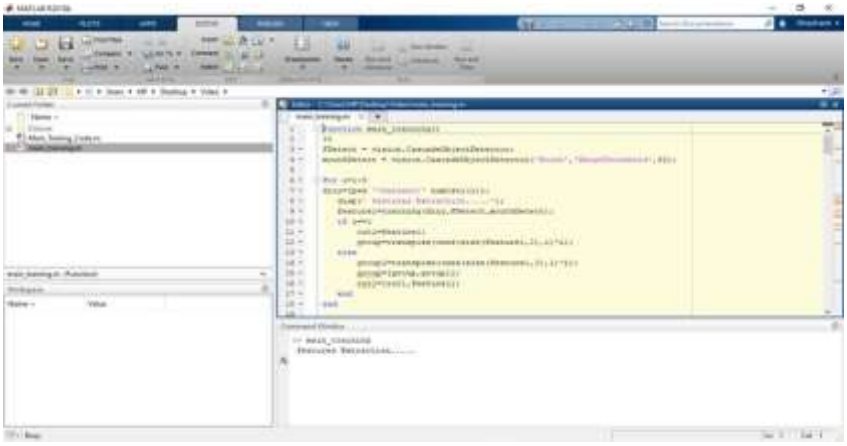


Figure 4.9. Comparison Table of Audio processing

4.3 Visual Recognition

This section discusses the results of implementing the Feedforward Neural Network (FFNN) algorithm for recognizing five distinct Kannada words. The training process involved video samples of these five words, with a total of 400 datasets utilized—250 for training and 150 for testing, as illustrated in Figure 4.10.

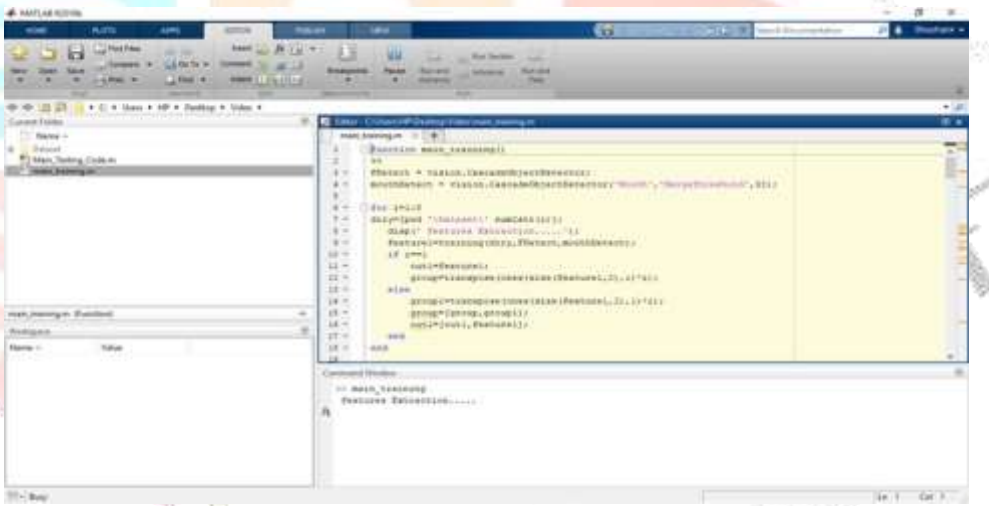


Figure 4.10. Training the dataset

The datasets are trained using a feedforward neural network, with performance metrics evaluated over 9 epochs. The resulting plots include performance metrics, regression analysis, and gradient variations, illustrating the network's learning and accuracy.

4.3.1 Confusion Plot

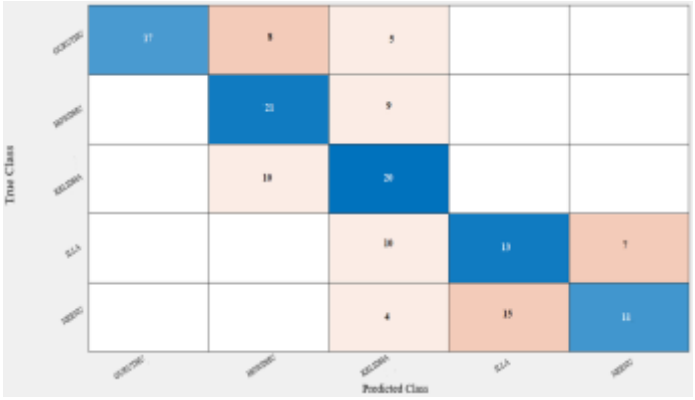


Figure 4.11. Confusion Plot of visual processing

The confusion matrix illustrates the results of video processing, highlighting the accuracy of word classification. Figure 4.11 details both correctly and incorrectly classified words, providing insights into the performance of the video recognition system.

4.3.2 Performance Plot

The performance plot depicts the relationship between the number of epochs and the mean squared error. Figure 4.12 shows that the optimal validation performance, with a mean squared error of 2.2518, occurs at epoch 3. As the number of epochs increases, the error rate decreases, indicating improved model performance over time.

Figure 4.12. Validation performance plot

4.3.3 Gradient Descent

This graph illustrates the relationship between epochs and the metrics of Gradient, Mu, and Validation Checks. At epoch 9, the gradient value is 0.2146, Mu is recorded at 0.01, and the validation checks total 6, as depicted in Figure 4.13.

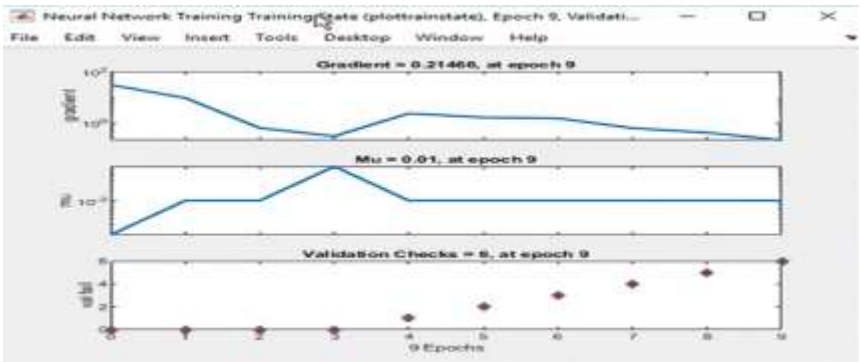


Figure 4.13. Gradient Descent plot

4.3.4 Regression plot of training, testing, validations and overall:

The Regression plot depicts the relationship between target values and output values across training, validation, and test datasets, as well as the overall performance. As illustrated in Figure 4.14, this plot visually represents the linear correlation between independent and dependent variables, allowing for an assessment of the strength and dispersion of the relationship. Additionally, Figure 4.15 provides a comparison table detailing the results of visual processing, highlighting differences and performance metrics across various conditions.

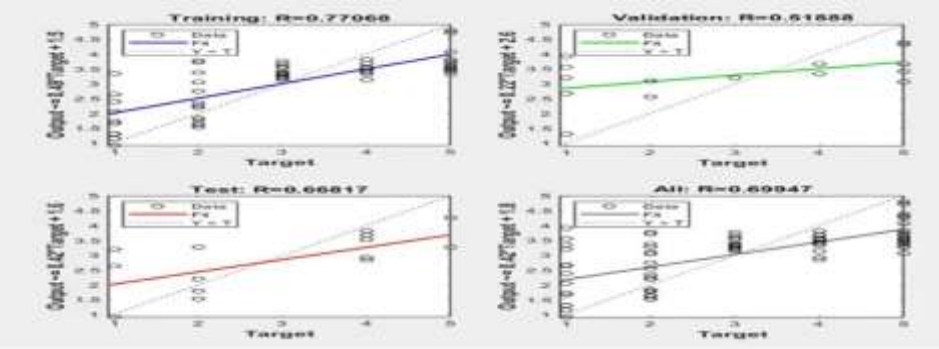


Figure 4.14. Regression plot of visual processing

Method	LSTM & CNN[4]	CNN, AlexNet[5]	DNN & HMM Viola-Jones[6]	DNN & HMM Viola-Jones[6]
Accuracy	62.70%	64.6%	45.63%	57%
Dataset	OLUV'S	Miracl-VC1	CUAVE	Custom

Figure 4.15. Comparison Table of visual processing

4.3.5 *Conclusion*

The Audio-Visual Speech Recognition (AVSR) system, developed within a simulated environment using a limited dataset and resources, demonstrated promising results. Specifically, the audio recognition component, employing Random Forest, achieved an accuracy of 88%, while the visual recognition component, using a Feedforward Neural Network (FFNN), achieved an accuracy of 57%. When combining both modalities with a hybrid model, the system reached an overall accuracy of 86%. These results reflect the system's performance in a controlled setting with available resources. However, there is potential for further improvement by integrating advanced hybrid models and expanding the dataset. In particular, the system could be enhanced by incorporating real-time data inputs and refining algorithms to handle diverse and dynamic conditions more effectively. Future developments may focus on improving sentence prediction capabilities in natural environments where audio samples are not always available.

V. ACKNOWLEDGMENT

REFERENCES

[1] Amaresh P Kandagal V. Udayashankara (2071) Visual speech recognition based on lip movement for Indian languages", International journal of computational intelligence research, Volume 2, Issue 6, 1 ISSN 2229-5518

[2] Leticia Ria Aran, Farrah Wong and Lim Pei Yi (2017) A review on methods and classifiers in lip reading", IEEE 2nd International conference on automatic control and intelligent systems Volume 5, Issue 9, 1 ISSN 2939-5518

[3] Tomoaki Yoshinaga, Satoshi Tamura, Koji Iwano, and Sadaoki Furui (2014) Audio- visual speech recognition using new lip features extracted from side-face images and research workshop on robustness issues in conversational interaction" IEEE 2nd Inter- national conference on computational research Volume 5, Issue 6, 1 ISSN 1229-5418

[4] Mihai Gurban and Jean-Philippe Thiran (2019)" Audio-visual speech recognition with a hybrid SVM-HMM system" IEEE 2nd International conference on computational research Volume 7, Issue 6, 1 ISSN 1229-5418 ,

[5] Tsang-long Pao, Wen-Yuan Liao, Tsan-Nung Wu, Chingylin (2019)" Automatic visual feature extraction for mandarin audio-visual speech recognition proceedings", IEEE 3rd International conference on computational research Volume 5, Issue 6, 1 ISSN 3229-5418

[6] Thein and Kalyar Myo San (2018) "Lip localization technique towards an automatic lip reading approach for Myanmar consonants recognition", international conference on information and computer technologies Volume 7, Issue 5, 1 ISSN 1229-5417.

[7] Amit Garg, Jonathan Noyola Sameep Bagadia (2019) "Lip reading using CNN and LSTM" international conference on machine learning, Volume 5, Issue 3, 1 ISSN 1289-5417.

[8] Mohammad Hasan Rahmani Farshad Almasganj (2017) "Lip-reading via a DNN- HMM hybrid system using combination of the image-based and model-based features", 3rd international conference on pattern recognition and image analysis, Volume 4, Issue 1, 1 ISSN 1259-5417

[9] Guoyun LV, Dongmei Jiang, Rongchun Zhao, Xiaoyue Jiang, H. Sahli (2015) "Multi- stream asynchrony dynamic Bayesian network model for audio-visual continuous speech recognition", international conference on pattern recognition and image analysis, Volume 1, Issue 3, 1 ISSN 1229-5517.

[10] Mihai Gurban, and Jean-Philippe Thiran (2012) "On dynamic stream weighting for audio-visual speech recognition", IEEE transactions on audio, speech, and language processing, vol. 20, no. 4.

[11] Takeshi Saitoh Ryosuke Konishi (2010) "Profile lip reading for vowel and word recognition", International conference on pattern recognition, vol. 18, no. 5.

[12] Kenichi Kumatani and Rainer Stiefelhagen (2018) "State synchronous modeling on phone boundary for audio visual speech recognition and application to muti-view face images", International conference visual recognition, vol. 12, no. 3

[13] Khalid Daoudi Alexandros Potamianost (2017) "Unsupervised stream weight computation in a segmentaion task: application to audio-visual speech recognition" IEEE international conference on signal processing and communications, vol. 11, no. 5

[14] Khadar Nawas K1, Manish Kumar Barik1 (2021) "Speaker Recognition using Random Forest" Institute of Technology, Chennai, Conferences 37, 01022 ICTSD-2021.

[15] M Subba Rao, G Bhagya Lakshmi, P Gowri, K Bharath Chowdary (2020) "Random forest based automatic speaker recognition system", The International journal of analytical and experimental modal analysis Volume XII, Issue IV, ISSN NO:0886-9367.

[16] Nivetha S, Rathinavelu A, Gayathri S (2020)"Speech Recognition System for Iso- lated Tamil Words using Random Forest Algorithm", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-9 Issue-1.

[17] Parashar Dhakal, Praveen Damacharla , Ahmad Y. Javaid and Vijay Devabhatuni "A Near Real-Time Automatic Speaker Recognition Architecture for Voice-Based User Interface", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2378-3779, Volume-11 Issue-1.

[18] Li Zheng, Qiao Li, Hua Ban, Shuhua Liu (2018) "Speech Emotion Recognition Based on Convolution Neural Network combined with Random Forest", Conference Paper DOI: 10.1109/CCDC.2018.8407844.

[19] Bhushan C. Kamble (2016) "Speech Recognition Using Artificial Neural Network",Int'l Journal of Computing, Communications Instrumentation Engg. (IJCCIE) Vol. 3, Issue 1 ISSN 2349-1469 EISSN 2349-1477.

[20] Aaron nichie (2013) "Voice recognition using artificial neural networks and gaussian mixture models", International Journal of Engineering Science and Technology (IJEST),ISSN : 0975-5462 Vol. 5 No.05.

[21] Navya Damodar, Vani H Y, Anusuya M A (2019) "Voice Emotion Recognition using CNN and Decision Tree" ,International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-12.

[22] Dr.R.L.K.Venkateswarlu, Dr. R. Vasantha Kumari, G.Vani JayaSri (2011) "Speech Recognition By Using Recurrent Neural Networks", International Journal of Scientific Engineering Research Volume 2, Issue 6, 1 ISSN 2229-5518.