Turkish Journal of Computer and Mathematics Education Vol.11 No.3(2020),2853-2864
Research Article

A NOVEL SPEECH RECOGNITION SYSTEM USING FUZZY NEURAL NETWORK

B. Kishore Babu[1], Dr. Rakesh Mutukuru[2]

[1]Research Scholar, Department of Electronics and Communication Engineering, Shri Venkateshwara University, Gajraula, Uttar Pradesh,India

[2]Research Supervisor & Professor, Department of Electronics and Communication Engineering, Shri Venkateshwara University, Gajraula, Uttar Pradesh, India

ABSTRACT: An important field in digital speech processing is human voice. Speech recognition in humans has long been a hot topic in signal processing and artificial intelligence research. Natural Language Processing (NLP) offers an interdisciplinary subfield called speech recognition that makes it easier for machines to recognize spoken language and transform it into text. To communicate with machines, speech recognition machines has the potential to be helpful. Technology that can have communicates in real time has been made possible. However, there remain additional issues, such as speaker variance due to other factors like age, gender, signal speed, pronunciation variations, noise from the surrounding area etc. Classification of age and gender is important for speech processing. A lot of work has been done to enhance each of these phases in order to get better and more accurate results. The main goal of this analysis is on the integration of machine learning into the speech recognition system. Hence, this analysis presents a novel speech recognition system using Fuzzy neural network. Pre-processing, speech signal segmentation, speech feature extraction, and speaker recognition are the different phases of a speech recognition system. It uses a fuzzy neural network to identify the speaker's age and gender. Accuracy, F1-score, and Precision are used to evaluate the fuzzy model's performance.

KEYWORDS: Speech Recognition, Natural Language Processing, Machine Learning and Fuzzy Neural Network.

I. INTRODUCTION

The capacity for communication is one of the most fundamental aspects of human behavior. Humans communicate verbally and in writing with each other through their native languages. Speech, or the vocalized form of human communication, is the representation of human communication in written form.

Language and speech technology have advanced to a level where a high-quality human-computer interaction system has been created. The most basic and natural form of human communication is speech. It can accurately and quickly transmit crucial information. Speech is a type of communication found in nature, and along the way of evolution, humans have progressively developed speech-corresponding characters that have grown to be important social communication tools [1].

Transforming natural human speech into linguistic text that computers can understand is the goal of speech recognition technology. Recognizing different word pronunciations was the first goal of speech recognition technologies. However, individual word identification is no longer sufficient to meet the growing constantly requirements for human–computer interaction. The environment, speaker pronunciation habits, and the quantity of words to be recognized all contribute to the difficulty of speech recognition for sentences made up of plural words [2].

In computer-controlled applications, speech recognition is becoming more and more important. Voice biometrics, including voice frequency, flow, and accent, are evaluated by speech recognition techniques. A new method for humans to interact with machines will be made possible by this technology. While speech recognition is common for humans, computers find it difficult to recognize voices, particularly in real-time applications.

Intelligent human–computer interaction and machine translation both benefit from accurate speech recognition [3]. Applications involving speech classification and recognition several difficulties since speech signals and settings contain a number of non-linearities and disturbances. The implementation of classical computing systems can present with difficulties. Since speech varies from person to person, they need more efficient computing models than only regular, conventional models [4].

The context has a significant impact on the speech signal. There will always be some variation in the speech signal even the same speaker uses the same word repeatedly. Body language and more simple sentences are excellent tools for human communication, especially in two-way conversations. Developing an appropriate ASR (Automatic Speech Recognition)

system is complicated by a number of factors, including unclear word boundaries, noisy signals, regional and geographic dialects, and speaker variability [5].

There are several uses for speaker recognition. Control access to services like voice mail, a few examples include mobile banking, remote computer access, and information security. A reliable speaker recognition system should be constructed by looking into the impact of the feature extraction process. Additionally, as the characteristics of the acoustic signal change from male to female, it's essential to identify an appropriate feature extraction technique that takes into consideration all of these variances [6].

Although signals for speech have a high viability, interest in automatic speech detection is growing. It is true that authors can use a number of accents, dialects, and pronunciations to convey their views. They also vary between the sexes and from person to person [7]. Words with accents or incorrect pronunciations that cause the algorithm to become confused the challenges associated with speech recognition [8].

In clean environments, the majority of ASR programs function effectively. They do not function properly, though, when there is noise around. The unwanted elements found in speech signals are referred to as noise. Noise of any type complicates the ASR process. As an example, it is much simpler to recognize someone's voice in a quiet room than it is in a noisy environment. Because of this, a number of study found that ASR accuracy for a degraded speech signal is remained insufficient [9].

Since classifying gender has become second nature to humans and is still an important topic in the cognitive sciences, gender identification or recognition is a pressing matter on both a cultural and societal level. Human voice alone is a simple and accurate way to determine someone's gender. Male and female speakers are distinguished by timbre analysis and low-level pitch analysis. There exist several acoustical differences between the sexes due to the anatomo-physiological features of their respective speech production systems, including glottal function, formant frequencies, mean fundamental frequency, long-term average spectrum changes. Men's voices tend to be deeper and louder than women's, with women communicating at a pitch that is typically one octave higher. According to reports, an adult female's pitch average range is approximately 120–350 Hz, whereas an adult male's range is approximately 100–200 Hz. However, compared to male voices, a lower spectral tilt was indicated with greater aspiration noises in female voices. The reason behind women's voices appearing more "breathy" than men's is the larger opening at the back of their vocal cords, which permits more air to pass through. However, due to the high intra-subject variability and significant overlap in the range of acoustical values between male and female voices, depending just on pitch, formants, or other speaker-dependent acoustical aspects would not provide a reliable approximation [10].

Using a number of speech analysis methods and classification algorithms, gender detection from speech signals is still an important area of research. Some aspects of a person's voice, such as pitch level and utterances, play an essential role in determining their age and gender. Features of a speaker's speech or voice, such as their identity, age, gender, and emotions, might reveal information about person [11].

Everybody has an accent, and our voices change a lot, so one of the most difficult things about developing an ASR system contributes to for those differences in speech patterns. The accents of those who speak multiple languages vary more than those who speak just one language. Finding enough resources to teach the ASR model becomes more difficult when other factors are included, such as speaking speed, gender, social habits, and dialects [12].

Audio processing is one of the numerous sectors that has been changed by the introduction of machine learning and deep learning. Although complicated patterns and characteristics can be directly learned by Neural networks (NNs) from the data, they have demonstrated remarkable performance in speech improvement and separation tasks [13].
In order to support the system in unpredictable settings and solve Speech Recognition problems, a fuzzy model is implemented. This analysis presents a unique fuzzy neural network-based speech recognition system. The structure of this analysis is as follows: In Section II, the Literature Survey is explained. The section 3 presents novel speech recognition system using Fuzzy neural network. The section IV evaluates the result analysis of presented model. The section V describes the conclusion.

II. LITERATURE SURVEY

Tang Z., Li L., Wang D. and Vipperla R. et. al., [14] extends Joint Training for Speech and Speaker Recognition Using Multitask Recurrent Model-Based Collaborative Approach. The output of one task is back propagated to the other tasks in a collaborative joint training method based on multitask recurrent neural network models. The goal of jointly learning speaker recognition and automatic speech is coordinated with this general jointly learning technique. A thorough analysis demonstrates in comparison to single-task systems, multitask recurrent neural network models perform better on automatic speech and speaker identification tasks. Examined is the efficiency of this type of multitask collaborative learning, and examined are the impacts of various training configurations.

Manamela P. J., Manamela M. J., Modipa T. I., Sefara T. J. and Mokgonyane T. B. et. al., [15] explains that machine learning algorithms are used to automatically recognize the emotions in Sepedi speech. In this research, a Speech emotion recognition (SER) system that can recognize and classify six basic emotions from speech in South Africa's official language, Sepedi, is examined: anger, sadness, disgust, fear, happiness, and neutral. Next, using the pyAudioAnalysis program, 34 speech features were collected from the spoken corpora in order to train and compare several algorithms using 10-fold cross-validation. The data-mining program WEKA (Waikato Environment for Knowledge Analysis) was used to carry out the experiments. According to the results, Auto-WEKA performs better than all of the conventional algorithms (SVM, KNN, and MLP). When compared to a TV (Television) broadcast speech corpus, the recorded speech corpus produced better recognition accuracy.

Meltzner G. S., Deng Y., Heaton J. T., De Luca G., Roy S. H. and Kline J. C. et. al., [16] explains that laryngectomy patients might use silent speech recognition as an alternative communication tool. Eight sEMG (surface Electro-Myo-Graphic) sensors were placed on the face (4) and neck (4) while examining words that are part of a vocabulary of 2500-words. For every one of the 39 frequently used phonemes in English, the phoneme-based recognition models were trained using an alternative collection of phrases; this based on phoneme identification from speech in movement, the remaining sentences were utilized to evaluate the models word recognition abilities. When limiting the sensor set to four places per person (n = 7), word error rates increased to 13.6% from an average of 10.3% for the entire eight-sensor set (averaging 9.5% for the top four individuals). With great potential to further enhance recognition performance, this study offers an attractive proof-of-concept for sEMG-based alaryngeal speech recognition.

D. T. Grozdić and S. T. Jovicic et. al., [17] provides Deep Denoising Autoencoder with Inverse Filtering for Whispered Speech Recognition. This work examines the problems with whispered speech detection in mismatched situations and describes at the acoustic characteristics of whispered speech in order to improve whisper recognition, and then proposes a Deep Denoising Auto-Encoder (DDAE) based novel robust cepstral features and preprocessing technique. Additional comprehensive analysis of cepstral distances, cepstral coefficient distributions, confusion matrices, and inverse filtering experiments demonstrate that voice in speech stimuli is the main cause of word misclassification in mismatched train/test setups. The new framework significantly improves the accuracy of whisper recognition, this is predicated on the TECC (Teager-Energy-based Cepstral) and DDAE characteristics. As a result, whisper detection accuracy increases by an absolute 31%. In the neutral/whisper scenario, 92.81% of words were recognized.

Kolbaek M., Yu D., Tan Z. -H. and J. Jensen et. al., [18] offers Deep Recurrent Neural Networks for Multitalker Speech Separation with Utterance-Level Permutation Invariant Training. This study contrasts two approaches for training an automatic speech recognition system with one that uses clean speech and one that uses noisy speech. Fourteen different types of noise were used in the speech recognition accuracy measure comparison. These were the sounds of computers and home appliances, street, transportation, educational settings, and lobby areas. It has been determined to extent noised speech training strategies are better to the competitive technique. Acceptable identification accuracy can be achieved with noised speech training at a signal-to-noise ratio is at least 10 dB, the study found that training with clean speech results in the same recognition accuracy at a minimum signal-to-noise ratio of 20 dB.

C. Kurian et. al., [19] describes Malayalam automatic voice recognition technology's use of text corpora and speech databases. The creation of speech corpus for various Malayalam speech recognition tasks is presented in this research. The creation of a transcription file and pronunciation dictionary, the other two necessary resources for developing a speech recognition system, is also created. For various recognition tasks, a speech corpus covering around eighteen hours has been gathered. For every assignment, a thorough explanation of the speech and text corpus collection is provided. Each identification task's text and speech corpus size is specified in detail. Following with sufficient examples, pronunciation dictionaries and transcriptions made using these text, speech corpuses are discussed. A full set of phones that are ready for the identification task are displayed in table manner. As a result, they have produced a pronunciation

dictionary with 2480 items and a whole speech corpus of 62000 words. It has been predicted that this work will address the initial challenges in the speech recognition process, including the establishment of pronunciation dictionaries are the corpus gathering of speech and text.

Sharma P., Abrol V. and Sao A. K. et. al., [20] provide features for speech recognition based on deep sparse representation. Obtaining a feature representation in order to recognize speech, it makes use of a multilayer decomposition with several layers, commonly known as the Deep sparse representation (DSR). The proposed structure, which uses a single sparse layer as opposed to several, utilizes a dense layer placed between two sparse layers to help with effective implementation. Experiments show that a number of speech recognition tasks, the suggested feature performs better than existing features. However, for other applications such as voice conversion, emotion identification in speech, etc., it is necessary to evaluate the performance of several models using various dictionary learning strategies at various levels.

B. Wu et al., [21] describes an extensive deep learning approach for simultaneous acoustic modeling and speech Dereverberation for robust speech recognition. By simultaneously learning the front-end speech signal processing and the back-end acoustic modeling, they provide an integrated end-to-end Automated speech recognition (ASR) paradigm. The objective is achieved through two methods: (i) The quality of loud and reverberation speech can be enhanced by a DNN (Deep Neural Network)-based speech dereverberation architecture that can handle a wide range of reverberation times; and (ii) Using the data collected and processed with multichannel microphone arrays to improve ASR performance with DNN-based multicondition training that takes into consideration both clean-condition and multicondition speech. This suggested framework is tested using the most current REverberant Voice Enhancement and Recognition Benchmark (REVERB) Challenge problem. Using the suggested DNN-based pre-processing technique and clean-condition training, the authors achieved the best single-system Word error rate (WER) of 13.28% on the 1-channel REVERB simulated data.

Kim M., Kim Y., Yoo J., Wang J. and Kim H. et. al., [22] described the KL-HMM Regularized Dysarthric Speech Recognition Speaker Adaptation. The hidden Markov model (KL-HMM) based on Kullback-Leibler divergence is implemented, wherein the emission probability of the state is controlled using a categorical distribution and the phoneme posterior probabilities produced by a deep neural network-based acoustic model are utilized. Using a database of several hundred words, the proposed speaker adaption approach is tested for thirty speakers had 12 mildly dysarthric, eight moderately dysarthric and thirty speakers, the proposed method outperformed the conventional deep neural network-based speaker adaption system by a large range on both dysarthric and non-dysarthric speech, as demonstrated by 10 non-dysarthric control speakers.

Zong Y., Zheng W., Zhang T. and Huang X. et. al., [23] explains the use of domain-adaptive least-squares regression for Cross-Corpus Speech Emotion Recognition. A Domain-adaptive least-squares regression (DaLSR) model is used to suggest a unique cross-corpus Speech emotion recognition (SER) technique. By using this method, the labeled training data set from the source speech corpus and an extra unlabeled data set from the target speech corpus are mixed to train the DaLSR model concurrently. To evaluate the effectiveness of the suggested approach in resolving the cross-corpus SER issue, they carry out extensive tests on three emotional speech corpora. Other advanced transfer learning techniques that are frequently applied to cross-corpus SER issues are compared with the outcomes. The results of the experiment demonstrate that the suggested approach outperforms the most advanced techniques in terms of recognition accuracy. On the other hand, the efficiency of transfer learning techniques in cross-corpus SER may be significantly impacted by an unbalanced numbers of class data samples.

Sailor H. B. and Patil H. A. et. al., [24] Enhanced Speech Recognition Using Novel Unsupervised Auditory Filterbank Learning with Convolutional RBM (Restricted Boltzmann Machine). They have experimented with filterbank features (called ConvRBM-BANK) and cepstral features (called ConvRBM-CC). They observed that the Word error rate (WER) improved by 7.21–17.8% when compared to Mel Frequency cepstral coefficient (MFCC) features and by 1.35–6.82% when compared to Mel Filterbank (FBANK) features on a large vocabulary continuous speech recognition test. Using ConvRBM-CC features, a Hybrid Deep Neural Network-Hidden Markov Model (DNN-HMM) system produced a relative improvement in WER of 4.8-13.65% on the Aurora 4 multi-condition training database. In comparison to FBANK features, they obtain an absolute reduction in WER on AURORA 4 test sets of 1.25–3.85% using ConvRBM-BANK features. With a relative improvement of 3.6-4.6% on average for bigram 5k and tri-gram 5k language models, a context-dependent DNN-HMM system substantially enhances performance.

Zheng W., Xin M., Wang X. and Wang B. et. al., [25] explains the novel method of using incomplete sparse least square regression to recognize emotions in speech. An innovative approach to speech emotion recognition based on the Least Square Regression (LSR) model, wherein the linear relationship between speech components and the related emotion labels is characterized by a novel Incomplete Sparse LSR (ISLSR) model. Two types of speech data sets are used to train the ISLSR model: labeled and unlabeled. The purpose of using unlabeled data sets is to improve the model's compatibility so that it may be effectively applied to speech data that is not included in the sample. ISLSR's ability to handle feature selection is another innovation. The ISLSR approach provides an average precision of 60.25% and an average recall of 60.50%, respectively.

III. A NOVEL SPEECH RECOGNITION SYSTEM

A novel speech recognition system using Fuzzy neural network is presented in this section. Figure 1 displays the speech recognition system's block diagram. Linguistic information is carried through speech signals, which allow the exchange of ideas and information. Using a microphone, they gather the continuous speech in the first step. Background noise is removed as part of the data pre-processing stage. The data is removed of background noise, leaving only voice samples as the input for the subsequent speech recognition procedure.

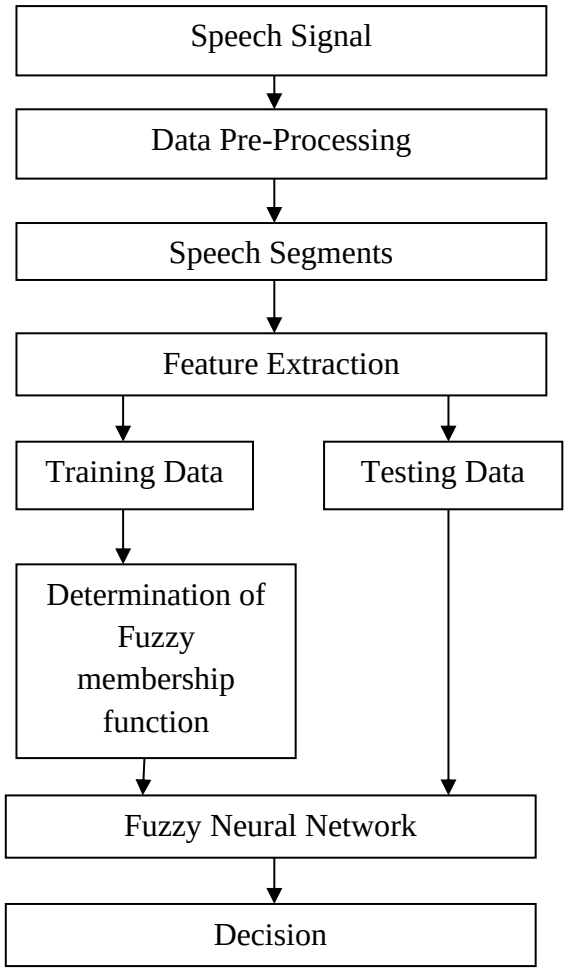

Figure 1: Block Diagram of Novel Speech Recognition System

speech signal Least Mean Square (LMS) filter for noise reduction. After the noisy speech is sent through a FIR (Finite Impulse Response) filter, its coefficients are found by reducing the clean signals Mean square error (MSE), is the desired signal is obtained. Pre-processing of recorded speech samples includes windowing and framing of the speech signals, as noise and silence reduction.

The technique of dividing down the speech signal into smaller components is known as speech segmentation. In continuous speech, it requires segment recognition and further processing to produce identifiable qualities. It is employed to determine the segment borders should begin and end. For different automated speech processing algorithms, it is important. Speech segmentation comes an two different forms: automatic and manual. This system segmentation is done automatically. To identify the speech word segments, a simple dynamic threshold-based technique is used after speech feature sequences have been computed.

Extraction of Mel-Frequency Cepstral Coefficients (MFCCs): The most often utilized characteristics for human speech analysis and recognition are MFCCs. This frequency scale is used to extract cepstral coefficients, which are referred to as MFCCs.

One of the most effective and widely used techniques for extracting speech features. These characteristics are preferred for speech recognition applications that simulate the response of the human auditory system since the frequency bands are equally spaced. The following are the MFCC's primary steps: (i) To reduce spectrum distortion, the speech flow is first windowing and then transformed using the Fourier transform.

(ii) Using triangle overlapping windows, the values of the windowed Fourier frequency component are mapped to MelScale and the powers of the spectrum.
(iii) The power logs for every Mel frequency are calculated.
(iv) The real values are computed using the DCT (Discrete Cosine Transform) method.
(v) The amplitudes of the resultant spectrum are calculated taking into consideration the unique properties of the MFCCs.

Training data are used to determine the fuzzy rule base and the fuzzy membership function at the fuzzification layer. Subsequently, the weight matrix gets started using the data. Training phase is the term for this procedure.

Membership functions of the input variables are found in the fuzzification layer. Because there are n input variables, X_i generates P_i fuzzy sets by examining the data histograms for each input variable. As a result, the total nodes in this layer are equal to $\sum P_i$ nodes. This layer produces the fit values of the input values to the associated membership functions.

In the analysis, three different types of input variable membership functions were used: type 1 (leftmost membership function), type 2 (middle membership function), and type 3 (rightmost membership function). These types of membership functions are as follows: Type 1: In terms of membership functions, MFleftmost is the leftmost membership function. The function is of the L type. MFleftmost exhibits an x1 peak and an x2 valley. Type 1: Of all membership functions, MFleftmost is the leftmost membership function. The function is of the L type. MFleftmost exhibits an x_1 peak and an x_2 valley. The following is an expression of the type of leftmost membership functions:

$$MFleftmost(x) = \begin{Bmatrix} 1\ if\ x \le x_1 \\ \dfrac{x_1 - x}{x_2 - x_1} if x_1 < x \le x_2 \\ 0\ if\ x > x_2 \end{Bmatrix} \quad (1)$$

Type 2: A membership function type known as MFmiddle is situated in the middle of the left and right membership function types. These functions are the triangle type. Between the left valley (at x_1) and the right valley (at x_3), this type only has one peak (at x_2). The following is an expression of this type of membership function.

$$MFmiddle(x) = \begin{Bmatrix} 0\ if\ x \le x_1 \\ \dfrac{x - x_1}{x_2 - x_1} if x_1 < x \le x_2 \\ \dfrac{x_3 - x}{x_3 - x_2} if x_1 < x \le x_3 \\ 0\ if\ x \ge x_3 \end{Bmatrix} \quad (2)$$

Type 3: The rightmost membership function is called MFrightmost. That is a Gama function type. The particular type in presents an x_1 valley and an x_2 peak. The rightmost membership function's type is expressed as

$$MFrightmost(x) = \begin{cases} 0 \; if \; x \leq x_1 \\ \dfrac{x - x_1}{x_2 - x_1} if x_1 < x \leq x_2 \\ 1 \; if \; x > x_2 \end{cases} \quad (3)$$

The fuzzy rules antecedent parts, which take the form of separate fuzzy words logical AND are found in the antecedent layer. $N = \prod_{i=1}^{n} P_i$ nodes are needed for this layer, and all possible combinations of fuzzy sets are utilized. Hence, there are n incoming links on each node in the antecedent layer. A weight that indicates the degree of utility of a related fuzzy set can be assigned to each incoming connection. This layer's nodes just compare the input values and take the minimum of them.

Next, the output of this layer is used to determine the values of the input values for the next layer, the weights of connections between the antecedent and succeeding layers are connected. Fuzzy rule consequent components are contained in the consequent layer. The antecedent layer and the consequent layer are fully connected. However, the weight assigned to each relationship may indicating the level of reliability associated with each one. This stage follows the inference's max-min compositional rule. So, when j^{th} consequent note $B_j(j = 1, 2, \ldots, 7)$ is connected to N antecedent nodes $A_1, \ldots, A_n$ with weight w_{ij}'s, the fuzzy set that results from the j^{th} subsequent node has the following membership function defined

$$\mu'_{Bj} = \min\left[\max\left\{\min(w_{ij}, output\ (A_j))\right\}, \mu_{Bj}(y)\right](4)$$

The j^{th} consequent node has the μ_{Bj} membership function, whereas the i^{th} antecedent node's output is represented by output (A_i). Every subsequent node's output takes the shape of a fuzzy set. The defuzzification layer creates a final, fuzzy sets of incoming results are combined to provide a different conclusion. The max-min compositional rule of inference is utilized by the fuzzy inference system to determine the crisp values. As a result, the resultant y is calculated as

$$y = f^{-1}(MF_{max})(5)$$

A hybrid computing model known as a Fuzzy neural network (FNN) that combines neural network and fuzzy logic concepts. Fuzzy logic, which can process imprecise and uncertain data, neural networks, which it aims to combine, and which are producing complex patterns from data. A neuro-fuzzy system, also known as a fuzzy neural network is a learning system that uses neural network approximation to determine the parameters of a fuzzy system, such as fuzzy sets or fuzzy rules. To determine a person's age and gender in this examination, a fuzzy neural network is used.

In this case, the speech signal's male and female features are identified using fuzzy logic. In general, fuzzy logic involves three essential steps. Fuzzification, fuzzy rule generation, and defuzzification are all included. Fuzzy data is created during the fuzzification process from system data. Triangle membership function plays a role in the fuzzification process. The process of creating fuzzy rules comes further.

Our fuzzy logic receives three inputs: energy Entropy (E), Short time energy (S), and Zero crossing rate (Z). The output of the fuzzy logic is the percentage of different male and female features that are present in the given voice signal. Male, female/male, female, aged/young person, and large are the three different sets of fuzzified input variables. The output variable is fuzzified into five sets. The speech signal in female/male can be assigned to any gender. Speech signals in the old/young domain to either the aged or the young.

STE (Short Time Energy): It is believed that the sudden increase in energy signal in a speech signal is known as the STE. The signal is first divided into S windows in order to compute STE. Each window's function is then computed. The equation that follows is used to compute the STE.

$$S = \sum_{r=-\infty}^{\infty} y(r)^2 . h(s - r) \; (6)$$

The STE is computed using the equation above. The testing results show that females have a high and continuous energy entropy output, males have a low output. The human voice signal's amplitude changed over time. A human voice's utilized part has a greater amplitude value while its unvoiced part has a lower amplitude.

Pitch: Regarding gender or age-based speech recognition systems, pitch is the most important characteristic. The pitch detection algorithm can be used to calculate pitch. It is demonstrated that the female voice has a higher pitch than the male voice using the estimated values from the pitch detection method. The resonant frequency of the vocal folds determines the pitch of a human voice, according to the researchers. The pitches of the adult male and female pitches are significantly lower than those of the small children, whose pitches are substantially higher. Adult male voice frequencies are approximately 125 Hz, adult female frequencies are approximately 210 Hz, and children's voice frequencies are more than 300 Hz.

Zero Crossing Rate (ZCR): The most crucial factor taken into consideration in this approach is the ZCR. The number of time domain zero crossings to the frame length is the ratio that defines the ZCR. The method for determining the zero crossing rate is given in equation 7.

$$Z = \frac{1}{2N}\sum_{i=1}^{N-1} sgn\{x(i)\} - sgn\{x(i-1)\} (7)$$

Where sgn{(x(i)} stands for the sign function, i.e.

$$sgn\{x(i)\} = \begin{Bmatrix} 1; x(i) > 0 \\ 0; x(i) = 0 \\ -1; x(i) < 0 \end{Bmatrix} (8)$$

ZCR is determined for each signal using the equation above. Based on the testing results, they found that female speech had a greater ZCR than male speech.

Energy Entropy (EE): The sudden differences in a speech signal's energy level are referred to as EE in speech signals. First, k frames are extracted from the voice signal, and the EE is then computed using the normalized energy of each frame. The following formula can be used to determine energy entropy:

$$E = -\sum_{i=0}^{k-1} \sigma^2 . log_2 (\sigma)^2 (9)$$

The normalized energy is denoted by σ^2. According to the test results, females have a high energy entropy that lasts for only a few seconds, whereas male energy entropy is low and distribution.

Training fuzzy logic next step, following the creation of fuzzy rules. Table 1's rules are utilized to train the fuzzy logic. It is necessary to create training datasets in order to train fuzzy logic. The training dataset consists of {[E_{max}, E_{min}], [S_{max}, S_{min}], [Z_{max}, Z_{min}]} developed as input. The acquired fuzzy logic is prepared for use in real-world scenarios upon training completion. When testing, the fuzzy logic will produce an output that indicates if a feature belongs to a male, female, aged, or young people if they offer the E, S, and Z values as input.

Here, the percentage of male and female features in a particular speech stream is determined using fuzzy neural networks. An input layer, a hidden layer, and an output layer are the three layers that made up a neural network. One variable is included in the output layer, n variables are in the hidden layer, and three variables are in the input layer. The neural network receives three inputs: zero crossing rate, short time energy, and energy entropy. The training stage and the testing stage are the two phases of neural network activity. A training dataset is created for neural network training. {[Emax,Emin], [Smax,Smin], [Zmax,Zmin]} is the generated input training dataset.

The trained fuzzy neural network is prepared for a wide range of real-world uses. Testing neural networks comes further when training is finished. The speech signal that is provided as input during testing indicates if the speaker is a male, female, old, or young individual.

From the results obtained to the presented approach, i.e., fuzzy neural network, the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values are determined. Performance metrics such as accuracy, precision, and F1-score are calculated using the four numbers mentioned above.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 (10)$$

$$Recall = \frac{TP}{TP+FN} \times 100 \text{ (11)}$$

$$Precision = \frac{TP}{TP+FP} \times 100 \text{ (12)}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \text{ (13)}$$

IV. RESULT ANALYSIS

In this section, result analysis of a novel speech recognition system using Fuzzy neural network is demonstrated. The precision, accuracy, and F1-score of the model is being used to evaluate its performance. The evaluation of performance is displayed in Table 1.

Metrics/ML Methods	**Naïve Bayes (NB)**	**Fuzzy Neural Network(FNN)**
Precision (%)	91	95.6
Recall (%)	90.56	95.4
Accuracy (%)	91.2	96.23
F1-score (%)	90.4	95.5

The Fuzzy neural network performance is compared with Naïve Bayes classifier. Compared to NB, FNN has shown better performance for speech recognition and identification of speaker gender as age. The Figure 2 shows recall and precision performance graph.

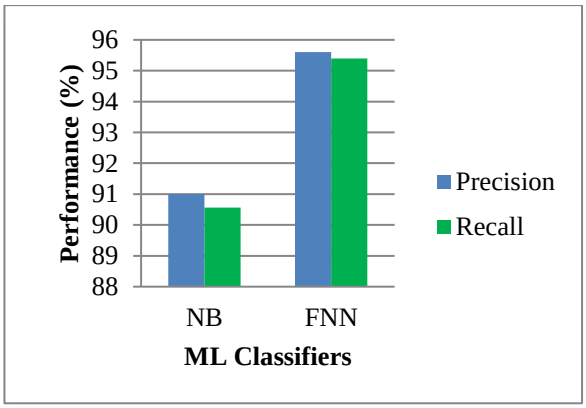

Figure 2: Performance comparative graph

In figure 2, x-axis indicates ML classifiers for speech recognition and y-axis indicates performance values in terms of percentage. The FNN classifier has obtained better Precision and Recall for speech recognition than NB classifier. The Figure 3 shows accuracy comparison.

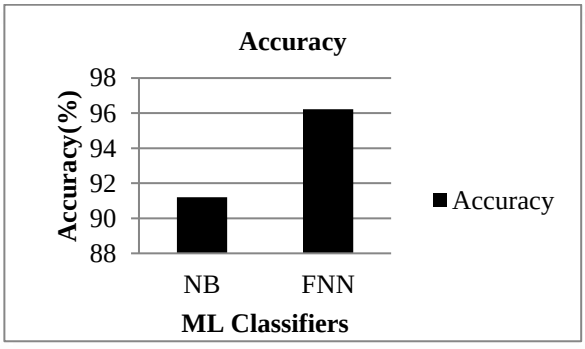

Figure 3: Accuracy Comparison

The FNN has achieved high accuracy than Naïve Bayes for speech gender classification and age estimation. The Figure 4 shows the F1-score comparative graph.

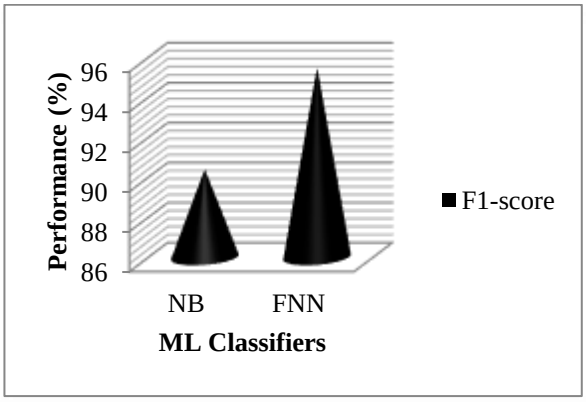

Figure 4: F1-score Comparative Graph

Compared to NB classifier, FNN has achieved better F1-score. Hence this model has achieved better performance for speaker gender classification and age identification.

V. CONCLUSION

A novel speech recognition system using Fuzzy neural network is presented in this study. In this work, the age and gender of the speaker are determined using a fuzzy neural network. First, the speech signal is gathered and processed with pre-processing in order to remove noise from the environment. The pre-processed data is segmented into small segments to detects the speech word segments. The features are retrieved using the Mel-Frequency Cepstral Coefficients (MFCCs). The training and testing sets of data were separated by the extracted features. Fuzzy member function training is applied to the data. The information is tested using the fuzzy neural network. The Fuzzy neural network recognizes the speech, classifies the gender of the speaker and identifies the speaker as aged or young person. The precision, accuracy, F1-score, and recall of the model used to evaluate its performance. Compared to earlier ML classifiers, FNN has obtained better performance for speech recognition. This model has effectively classified the speaker as male or female and the speaker is aged/young. Hence this approach will be used in real time for various speech recognition applications.

VI. REFERENCES

[1] K. Žmolíková *et al.*, "SpeakerBeam: Speaker Aware Neural Network for Target Speaker Extraction in Speech Mixtures," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800-814, Aug. 2019, doi: 10.1109/JSTSP.2019.2922820.

[2] Livieris, Ioannis E., Emmanuel Pintelas, and Panagiotis Pintelas. 2019. "Gender Recognition by Voice Using an Improved Self-Labeled Algorithm" Machine Learning and Knowledge Extraction, vol. 1, no. 1, pp. 492-503, 2019, doi:10.3390/make1010030

[3] P. Agrawal and S. Ganapathy, "Modulation Filter Learning Using Deep Variational Networks for Robust Speech Recognition," in IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 2, pp. 244-253, May 2019, doi: 10.1109/JSTSP.2019.2913965.

[4] Noraini Seman, Ahmad Firdaus Norazam, "Hybrid methods of Brandt's generalised likelihood ratio and short-term energy for Malay word speech segmentation," Indonesian Journal of Electrical Engineering and Computer Science, vol. 16, no. 1, pp. 283-291, October 2019, doi: 10.11591/ijeecs.v16.i1.pp283-291

[5] Sunanda Mendiratta, Neelam Turk, Dipali Bansal, "A Robust Isolated Automatic Speech Recognition System using Machine Learning Techniques," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, no. 10, pp. 2325-2331, August 2019, doi: 10.35940/ijitee.J8765.0881019

[6] F. Tao and C. Busso, "Gating Neural Network for Large Vocabulary Audiovisual Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 7, pp. 1290-1302, July 2018, doi: 10.1109/TASLP.2018.2815268.

[7] Dr. A. S Umesh , Prof. Ramesh Patole , Prof. Krishna Kulkarni, 2019, Automatic Recognition, Identifying Speaker Emotion and Speaker Age Classification using Voice Signal, International Journal Of Engineering Research & Technology (IJERT), vol. 08, no. 11,November 2019, doi:10.17577/IJERTV8IS110123

[8]V. Mitra, W. Wang, C. Bartels, H. Franco and D. Vergyri, "Articulatory Information and Multiview Features for Large Vocabulary Continuous Speech Recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 5634-5638, doi: 10.1109/ICASSP.2018.8462028.

[9] Gnevsheva, K., & Bürkle, D, "Age Estimation in Foreign-accented Speech by Native and Non-native Speakers. Language and Speech, vol. 63, no. 1, pp. 166-183, 2020, doi:10.1177/0023830919827621

[10] T. J. Sefara and A. Modupe, "Yorùbá Gender Recognition from Speech Using Neural Networks," 2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI), Johannesburg, South Africa, 2019, pp. 50-55, doi: 10.1109/ISCMI47871.2019.9004376.

[11] M. Chen, X. He, J. Yang and H. Zhang, "3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition," in *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440-1444, Oct. 2018, doi: 10.1109/LSP.2018.2860246.

[12]Saeid Safavi, Martin Russell, Peter Jančovič, "Automatic speaker, age-group and gender identification from children's speech," Computer Speech & Language, vol. 50, pp. 141-156, July 2018, doi: 10.1016/j.csl.2018.01.001

[13] A. Jati and P. Georgiou, "Neural Predictive Coding Using Convolutional Neural Networks Toward Unsupervised Learning of Speaker Characteristics," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 10, pp. 1577-1589, Oct. 2019, doi: 10.1109/TASLP.2019.2921890.

[14] Z. Tang, L. Li, D. Wang and R. Vipperla, "Collaborative Joint Training With Multitask Recurrent Model for Speech and Speaker Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 3, pp. 493-504, March 2017, doi: 10.1109/TASLP.2016.2639323.

[15] P. J. Manamela, M. J. Manamela, T. I. Modipa, T. J. Sefara and T. B. Mokgonyane, "The Automatic Recognition of Sepedi Speech Emotions Based on Machine Learning Algorithms," 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), Durban, South Africa, 2018, pp. 1-7, doi: 10.1109/ICABCD.2018.8465403.

[16] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy and J. C. Kline, "Silent Speech Recognition as an Alternative Communication Device for Persons With Laryngectomy," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 12, pp. 2386-2398, Dec. 2017, doi: 10.1109/TASLP.2017.2740000.

[17] Đ. T. Grozdić and S. T. Jovičić, "Whispered Speech Recognition Using Deep Denoising Autoencoder and Inverse Filtering," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 12, pp. 2313-2322, Dec. 2017, doi: 10.1109/TASLP.2017.2738559.

[18] M. Kolbaek, D. Yu, Z. -H. Tan and J. Jensen, "Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 10, pp. 1901-1913, Oct. 2017, doi: 10.1109/TASLP.2017.2726762.

[19] C. Kurian, "Speech database and text corpora for Malayalam language automatic speech recognition technology," 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), Bali, Indonesia, 2016, pp. 7-11, doi: 10.1109/ICSDA.2016.7918975.

[20] P. Sharma, V. Abrol and A. K. Sao, "Deep-Sparse-Representation-Based Features for Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 11, pp. 2162-2175, Nov. 2017, doi: 10.1109/TASLP.2017.2748240.

[21] B. Wu *et al*., "An End-to-End Deep Learning Approach to Simultaneous Speech Dereverberation and Acoustic Modeling for Robust Speech Recognition," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1289-1300, Dec. 2017, doi: 10.1109/JSTSP.2017.2756439.

[22] M. Kim, Y. Kim, J. Yoo, J. Wang and H. Kim, "Regularized Speaker Adaptation of KL-HMM for Dysarthric Speech Recognition," in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 9, pp. 1581-1591, Sept. 2017, doi: 10.1109/TNSRE.2017.2681691.

[23] Y. Zong, W. Zheng, T. Zhang and X. Huang, "Cross-Corpus Speech Emotion Recognition Based on Domain-Adaptive Least-Squares Regression," in IEEE Signal Processing Letters, vol. 23, no. 5, pp. 585-589, May 2016, doi: 10.1109/LSP.2016.2537926.

[24] H. B. Sailor and H. A. Patil, "Novel Unsupervised Auditory Filterbank Learning Using Convolutional RBM for Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 12, pp. 2341-2353, Dec. 2016, doi: 10.1109/TASLP.2016.2607341.

[25] W. Zheng, M. Xin, X. Wang and B. Wang, "A Novel Speech Emotion Recognition Method via Incomplete Sparse Least Square Regression," in IEEE Signal Processing Letters, vol. 21, no. 5, pp. 569-572, May 2014, doi: 10.1109/LSP.2014.2308954.

.