A Computer-Based Application for Speech Recognition in Multi-Speaker Environment to Assist Hearing Impaired People

Aishwarya Anegundi, Chalana D A, Vinuta V Pawale, Vinay G, Rudraswamy S B, Manohar N

Abstract*: *In this paper, a computer based application has been proposed that facilitates speech recognition in multi- speaker environment. It identifies the speaker, recognizes the speech and displays the text output obtained from speech to text conversion along with identified speaker's name tag on the computer screen. The proposed system provides three functionalities: i) Speech Recognition ii) Speaker Identification iii) Many to one transmission-reception of converted speech to text using client-server model and multi-threading concept. The system consists of transmission and reception capable devices like computers, computer application that connects various devices using WiFi technology and microphone to obtain speech input from the users. All the functionalities are implemented as computer application with user friendly graphical user interface(GUI). The application has two scenarios: i) Near and ii) Far. In the case when the speakers are in the range of the user's microphone, the user opts for near scenario else user opts for far scenario. The proposed system is intended mainly for people suffering from hearing loss. Although hearing aid technology is greatly improved, it still utilizes impaired person's inner, outer and middle ear including hearing nerve to support natural hearing. In situation when there is damage to hearing nerve also, hearing aid is not of use. In such cases, the proposed system assists the hearing impaired by providing speech in the form of text and by identifying the speakers involved. It thereby prevents fatigue caused by listening effort and stress. This leads to increased productivity of the impaired persons. The proposed assistive system also makes it possible for normal people who do not know sign language to speak to hearing impaired. This makes the life of hearing impaired people more contented which improves their quality of life.

Index Terms*: Client-Server Model, *Graphical User Interface(GUI),Multi-Threading, Speaker Identification, Speech Recognition.

I. INTRODUCTION

Speech plays important role in day to day communication and interaction among humans [1]. Speech processing can be extensively used to develop applications like Biometric authentication based on voice, voice controlled machines, assistive devices for hearing impaired like speech recognition systems.

Revised Manuscript Received on May 22, 2019.

AishwaryaAnegundi, BE Student, Sri Jayachamarajendra College of Engineering, Mysore
Vinuta V Pawale BE Student Sri Jayachamarajendra College of Engineering, Mysore
Chalana D A BE Student Sri Jayachamarajendra College of Engineering, Mysore
VinayG BE Student Sri Jayachamarajendra College of Engineering, Mysore
DrRudraswamy S B Sri Jayachamarajendra College of Engineering, Mysore
N Manohar Sri Jayachamarajendra College of Engineering, Mysore

Accuracy of speech processing based applications largely depend on method used to extract features from a speech signal. Speech signal varies largely within a short frame of time which makes speaker identification difficult due to dependencies on large number of parameters. These parameters are state of emotion, accent, pronunciation, speech intensity, background noise, speaker gender, speaker age and pitch.

Speaker recognition systems are widely classified as text dependent and text independent systems. Systems dependent on text can recognize properly when speaker utters the same words during testing as in training phase. Where as in text independent systems, speaker can be identified accurately irrespective of the words uttered by the speaker [2]. Based on the application, speaker recognition systems are classified as speaker identification and speaker verification. A system which verifies the speaker merely decides if the speaker is who he claims but a speaker identification system has to identify a speaker model from dataset which matches closely with feature extracted from the speech sample obtained by speaker under test [3]. A speaker identification system comprises of two steps which are feature extraction and speaker modelling. The first step is feature extraction in which features are extracted from the speech samples using techniques like Mel Frequency CepstralCoefficients(MFCC), Bark Frequency cepstral Coefficients(BFCC), Linear Predictive Cepstrum Coefficients(LPCC). The second step isspeaker modelling using models like Gaussian Mixture Model(GMM), vector Quantization(VQ) [4]. Automatic speech recognition system convert spoken words into text based on utterances [5]. Usually communication with people having complete hearing loss happens through sign language or by writing on paper. Speech recognition system facilitates easier communication between normal people who do not know sign language and special people like hearing impaired. Speech recognition and speaker identification has been implemented using different methods for feature extraction and speaker modelling in various research works. An extensive use of Speech recognition systems is made in different fields for various kind of applications such as voice based security systems, human-machine interface.

AbhinavAnand et.al [6], proposed speaker recognition method independent of text,

Retrieval Number: A10310581C19/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

suitable for identification in environment that are electronically sensitive to presence of people.

They used MFCC for feature extraction, then created information set features using fuzzy logic and used hierarchical classification for identifying user's identity. This speaker recognition system was mainly for human machine interface and it did only speaker recognition and not speech to text conversion.

ZunoWatada, Hanayuki [7], proposed speech recognition system to analyse the pitch and human voice in a meeting place or in indirect conversation and other multi-speaker environment. The system is built using MFCC for feature extraction and Hidden Markov Model(HMM) for speaker modelling and matching. Since HMM is used the system is text dependent speech recognizer. It shows better performance for speech recognition in single speaker environment rather than in multi speaker environment. Authors also mentioned speaker detection for multiple speakers to be carried in future.

Teddy Surya Gunawan et.al [8], proposed a language identification system based on MFCC and VQ techniques. By using MFCC higher accuracy was achieved for more number of languages. This system was mainly for language identification and did not include speech and speaker recognition. A. T. Rusli et.al [9], proposed a speaker verification system wherein feature extraction is carried out using MFCC and feature matching and speaker modelling using Support Vector Machine(SVM). They compared the accuracy of the system by varying the orders of the MFCC coefficients used. The results show that 20-25 MFCC coefficients provide better accuracy. It did not integrate speech to text conversion and provided verification and not identification.

Ashwin Nair Anil Kumar, SenthilArumugamMuthukumaraswamy [10], proposed voice recognition system for security purposes. It uses MFCC for feature extraction and VQ for modelling. The drawback of the system is that it is text dependent. The system performs only speaker recognition and does not include speech to text conversion. MouazBezoui et.al [11], proposed a technique to perform feature extraction for Arabic speech recognition systems. They employed MFCC to achieve it. A. Maazouzi et.al [12], proposed a speaker identification system which uses MFCC for feature extraction and similarity measurements for feature matching. The system drawback is that it is text dependent. Ivan Stefanus et.al [13], proposed a GMM based speaker verification system. The system is developed mainly for forensics department in Bahasa Indonesia. It does not identify but performs verification and there is no speech to text conversion included.

Rania Chakroun et.al [14], proposed a text-independent speaker verification system where power spectrum density was used along with GMM to improve the accuracy of the system. Wenyong Lin [15] proposed a speaker identification system using GMM-based clustering algorithm for feature matching. Here the drawbacks of k-means clustering is overcome by using initial clustering algorithm which uses T-test matrices for finding the distance. This provided improved accuracy. JiPibil et.al [16], proposed a speaker age classifying system which evaluates performance for one and two level GMM classifier used. The two level GMM was built using different number of mixtures and it was showed that two level GMM gave more accuracy than One level GMM but the computation delay was more in two level GMM based system.Text output from speech recognition system can be sent to other devices either in the one-to-many format or in many- to-one format. Alan Chern et.al [17], proposed a hearing assistive system based on smart phone, to provide speech to text conversion for multiple end users who could benefit fromenhanced listening clarity in the classroom. The Smart hear system provides application with GUI features. It suppresses the noise in speech input obtained and transmits it to many devices in classroom using multicast and frequency modulation techniques. It also provides speech to text conversion in case of poor speech received. For speech recognition it employs Google API for android application. It does not integrate speaker identification feature in this system.

In this paper a system is proposed which integrates speaker identification feature with speech recognition system. Speaker identification is implemented using Mel frequency cepstralcoefficients(MFCC) and Gaussian Mixture Model(GMM). Speech recognition is implemented using Google Cloud Speech API. The feature of sending text obtained from various speech recognition systems placed at different location connected to same WiFi network is implemented using Client-Server Model and multi threading. In this system MFCC has been been used for feature extraction as it is efficient and has been largely used in speaker identification systems [6-12]. For featuremodelling, basic GMM is used without any enhancement as in literature [14-16] since it simple to implement and to minimise computation delay so as to develop a real time system.

II. METHODOLOGY

A. Aim

The aim of this study is to implement a system that is able to perform speech and speaker recognition in a multi-speaker environment. To develop a user interactive application that will be helpful for hearing impaired.

B. Design and Implementation

The major parts of the system include speech recognition, speaker identification and implementation of client-server model using multi-threading.

1. Proposed Work

In the proposed system, a standalone application is developed using python programming language for desktop system with either windows or linux operating system. The application provides Graphical User Interface(GUI) which is built using Tkinter framework. There are two options for the user which are far and near scenario. The user opts for near scenario if the speakers are in the range of system's microphone.

Retrieval Number: A10310581C19/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The near scenario is intended to help the user in environments like group discussion in college, conversations among family members. In near scenario the system is trained for speakers with whom the user usually involves in conversation or in group discussions. Every time these speakers speak, the system identifies the speaker, converts the spoken sentences into text and displays it on the user screen along with the speaker name. The user opts for far scenario when the speakers are not in range of system's microphone but are connected to same WiFi network to which the user is connected to. The far scenario is particularly suitable for environments like conferences and workshops, where the range of the speaker and listener is about few meters. In far scenario the speakers also have the proposed applications running on their devices where they opt for "far- speaker" option. Speech input from individual speakers are converted into text on their devices respectively and are sent to listener's device. On the listener's end the application looks forward for text input from speakers and displays it on its screen with speaker names based on the Internet Protocol(IP) address of device it received the text input from. Python provides various toolkits for speech signal processing like numpy, scipy, matplot, python speech features, pyaudio, wave. It also provides packages like socket,time, thread used for client-server model implementation. Hence the system can be developed effortlessly using python platform.

2.Design of Speaker Identification System

Speaker Identification is implemented in two steps which are feature extraction and speaker modelling.

a. Feature Extraction

Feature Extraction is performed to obtain an array of feature vector coefficients from speech signal. The details needed to identify the received speeches as belonging to a particular speaker is provided by these coefficients[18]. These features are extracted using MFCC. The vocal tract of human demonstrates itself as wrapped in power spectrum over tiny time intervals. To precisely model this short time power spectrum is the task of MFCC. The block diagram of MFCC is as shown in Fig.1.

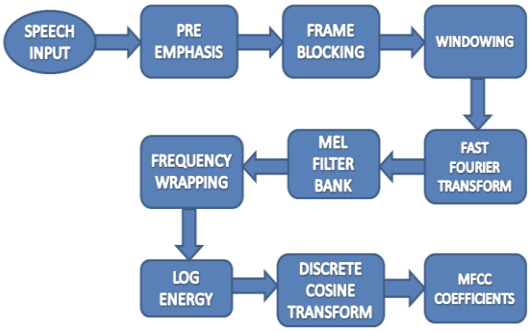

Fig. 1: Block Diagram of MFCC

The first step in MFCC is pre-emphasis. Speech signal before pre-emphasis is as shown in Fig 2. This step is realized using a first order filter as shown in equation (1).

$$a2(n) = a(n) - k * a(n-1) \quad (1)$$

where a(n) is original speech signal and a2(n) is signal after pre-emphasis and k is constant which is usually 0.97. The pre-emphasis can also be achieved using mean normalization instead of using first order filter [19].

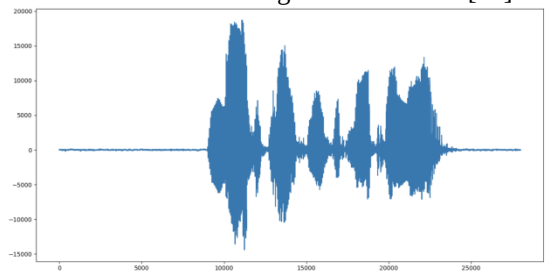

Fig. 2: Speech signal before pre-emphasis

Speech signal after pre-emphasis is as shown in Fig 3. After pre-emphasis there is balance in frequency spectrum

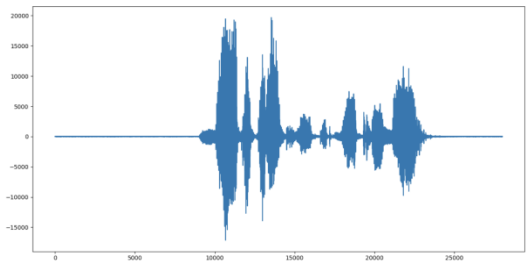

Fig. 3: Speech signal after pre-emphasis

since magnitude at higher frequencies is comparable to magnitude at lower frequencies which will minimise signal to noise ratio. The next step is frame blocking. Short intervals of speech signals of about 20 to 40 milliseconds are used for speech processing. Speech signal varies largely with time, by framing an assumption is made that speech signal remains constant in one particular frame. Pre-defined size of each frame are overlapped to ensure there is smooth transition between the frames[20]. In order to make it possible, the usage of Fast Fourier Transform(FFT), frame size is always chosen as power of 2. If this is not the case, zero padding is done to the nearest length of power of two. After this step windowing is carried out. The graph of hamming window is shown in Fig 4.

Windowing is done to maintain continuity between the first and last point of the frame [21]. This is done by multiply each frame by hamming window function given by equation (2):

$$W(n) = 0.54\text{-}0.46\cos[(2\pi n)\square(N\text{-}1)] \quad (2)$$

where, $0 \leq n \leq N\text{-}1$, N is the window length.

Once windowing is done, fast fouriertransform(FFT) is applied to the windowed signal. Each of N samples in a frame is converted from time domain to frequency domain by applying FFT. FFT gives rise to frequency spectrum with wide range of frequencies which are not linear. After FFT is applied, power spectrum is calculated for each frame. This is done to obtain the information on frequency components present in each frame. Power spectrum contains other information that are not necessary for speaker identification. The

Retrieval Number: A10310581C19/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

graph of power spectrum is as shown in Fig 5.

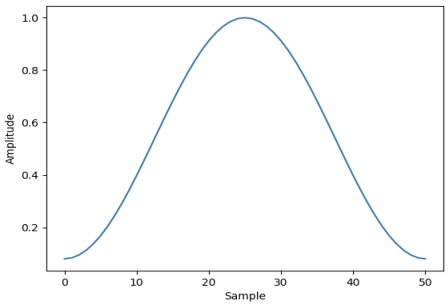

Fig. 4: Graph of Hamming Window

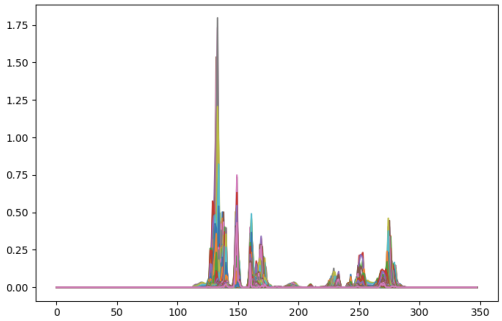

Fig. 5: Power Spectrum of entire speech signal

Power spectrum cannot differentiate between two closely spaced frequencies, and this effect increases at higher frequencies. Hence clumps of power spectrum is taken to find energy in different frequency regions. This is done by using filter banks in mel scale as given by equation (3) [22].

$$1+ f\,m = 2595 * \log_{10}[(1+f) \square(700)] \quad (3)$$

The first filterbank is narrow indicating energy close to 0 Hz anfilterbank becomes wider at higher frequency to pay more attention to amount of energy rather than variation in energy. Mel scale decides the spacing and width of filterbanks. The plot of filterbanks is as shown in Fig 6. After computing the filterbank energies, their logarithms are taken. Logarithm is applied to indicate the loudness in noise which cannot be done in linear scale. Logarithm of mel spectrum is decorrelated using discrete cosine transform (DCT)[23]. This is done to decorrelate the overlapping filter banks which are quite correlated. After these steps are carried out, only 13 of the coefficients so obtained are considered as higher coefficients represent faster changes in filterbankenergy which degrade the performance of speaker identification system.

a. Speaker Modelling

Gaussian Mixture Model is used for modelling features extracted from the dataset of individual speakers. For an utterance from a particular speaker j made up of T frames, feature vector of N-Dimensions is extracted for every frame. Further, Gaussian model for the speaker j has feature vectors that follows Gaussian distribution with mean and deviation from mean as distribution parameters for any new utterance from same speaker j[3]. The Gaussian model of speaker j, λj is weighted sum of M component densities as shown in equation (4) and (5)[3].

$$p(xt \mid \lambda j) = \sum_{i=1}^{m} \quad giN(xt; \mu i, \sum_i \quad) \quad (4)$$

wheregi are mixture weights having i=1m=gi. The individual component densities

N (x t ; μ i , ∑i) represent:

$$N(\vec{x}_t; \vec{\mu}_i, \sum_i) = \frac{1}{(2\prod)^{D/2}|\sum_i|^{1/2}} * exp\left[-\frac{1}{2}(\vec{x}_t - \vec{\mu}_i)\sum_i^{-1}(\vec{x}_t - \vec{\mu}_i)\right] \quad (5)$$

whereμi is the mean vector and ∑i is the covariance matrix. For speaker the GMM model is given by λj .λj is parameterized by mean vectors, covariance matrices and mixture weights from all M component densities as in equation (6)[3].

$$\lambda_j = \{\mu_i, \sum_i \quad , g_i\}_j \quad (6)$$

for i = 1,2,3....M

In testing phase, T feature vector X = {x1,x2,x3.....xT }for an utterance is given to the system. From N speaker models, one speaker model is chosen which gives maximum posteriori probability as in equation (7) for given input feature vector sequence in order to identify the speaker [24].

$$j\hat{} = argmax_{(1 \le j \le N)}\, p(\lambda_j | X) \quad (7)$$

where jˆ is identified speaker. The usage of logarithms influences the decision of identified speaker. The decision which also depends on allowed independence between the observations, can be shown with Maximum-Likelihood (ML) scoring of the log likelihoods as in equation (8)[3]:

$$j\hat{} = argmax_{(1 \le j \le N)} \sum_{t=1}^{T} \quad log\, p(x_t | \lambda_j) \quad (8)$$

where p(xt|λj) is as shown in equation (4). In this system basic modelling is used because of its simplicity which later can be developed to achieve more advanced features. GMM in the system is using 128 Gaussian components that represent various configuration of vocal tract. There is a compromise between accuracy and computation speed. Using more Gaussian components provides more accuracy but this decreases the computation speed. GMM system with 128 Gaussian components suffices the requirement for speaker identification system proposed in this paper[13,25-27].

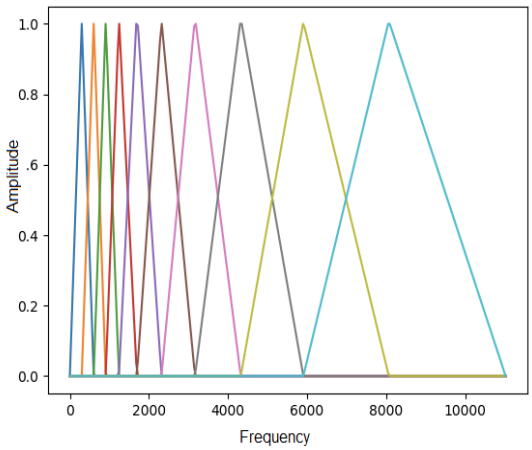

Fig. 6: Plot of Normalised Filter banks on Mel scale

Retrieval Number: A10310581C19/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

III. IMPLEMENTATION

The proposed work is implemented as two scenarios. These two scenarios are later integrated with Tkinter GUI framework.

a. Implementation of far scenario

In the far scenario a simple client-server model is employed to facilitate the converted text transmission from speakers' end to the listener's end. The client server model is developed using sockets. Sockets are nodes in network. Only the sockets connected to same port numbers on different device can communicate with each other. Every port has a port number and is associated with IP address of the host. Each port has particular protocol that it supports. Some of the ports are meant for specific protocols like HTTP, FTP and SMTP. Hence for our socket programming we have chosen port 5000 which has no pre assigned protocol.

The implementation is supported by python packages such as socket, threading, speech recognition and time. The concept is to create as many threads as the number of speakers who are trying to communicate with the listener. For a thread, before establishing the connection it is important to create a lock object. This is necessary to maintain the consistency of any shared resource that is being accessed by multiple threads. When the speaker's system is connected to WiFi to which even the listener's system is connected, the IP address of the listener's device is itself the IP address of the server.

The host address is IP address of the speaker's system. Then the connection with server is established using the function socket.connect (server, port). Then the thread is started where it obtains the speech input by using speech recognition.listen( ) . It is then converted to text using speech recognition.recognize_google ( ). This function returns the string which is stored in a string type variable. The converted text so obtained is encoded to utf-8 format and sent over connection with server using socket.send (string storing converted text). This logic keeps tracing until an exit condition is met which is when the speaker presses the stop button. When the speaker presses stop button, the system raises System Exit interrupt and then the thread execution is stopped and the lock is released. After this the socket is closed and connection is disconnected.

b. Implementation of near scenario

In near scenario, speaker identification is implemented. It is done in two stages: training and testing. In the training phase speakers' voice is pre recorded and processed. The audio samples are recorded on different days in order to account for the variation in voice produced. This is important to model the nature of vocal cord as close to real as possible. The next step is to pre-process these audio samples. In the pre-processing step the silence in speech sample is removed. This is majorly dependent on the energy during certain frame length. The speech sample is framed into milliseconds frame and energy during this frame is calculated. If this frame energy is less than 0.1 percent of the total average energy that frame is discarded. The frames of sufficient energy are all cascaded in order to obtain the sample which is free of silence frames.

After pre-processing features are extracted using MFCC and speakermodelling is done using GMM. A database consisting of speaker models of various speakers is formed. In the testing phase, features are extracted from the speech input obtained. Using these features speaker model is formed. If the logarithm of likelihood of claimed speaker is maximum when compared to other speaker models log-likelihood, claim is accepted otherwise claim will be rejected.

C. Integrating Near and Far Scenario with Tkinter GUI

The GUI is implemented using Tkinter framework and is named as "Multi speakers' speech recognition system". Multi Speakers' Speech recognition system GUI consists of one button for near scenario to start the system, two buttons for far
scenario: one for listener and another for speaker/talker to start the application and a common stop button for both the scenarios to exit the application. It also contains an entry box to input the name of the speaker/talker, before the speaker starts speaking in the far scenario. The GUI developed is shown in Fig.7.

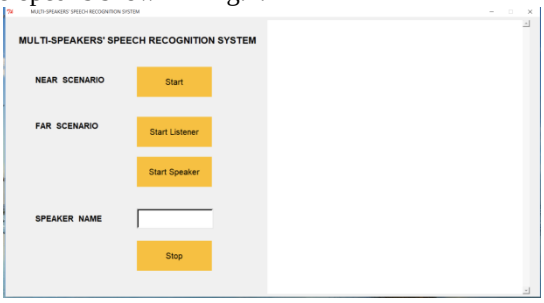

Fig. 7: GUI developed for proposed system

IV. RESULTS AND DISCUSSIONS

The application has been tested and the snapshots describing the navigation and outputs obtained is attached. When the end user taps on the start button of the near scenario, the chat window displays "YOU OPTED NEAR SCENARIO" which indicates that application is initiated. In the near scenario, the application differentiates between the speakers while all the speakers' speech input is taken by the same microphone. Fig 8 shows the conversation of two speakers on the chat window.

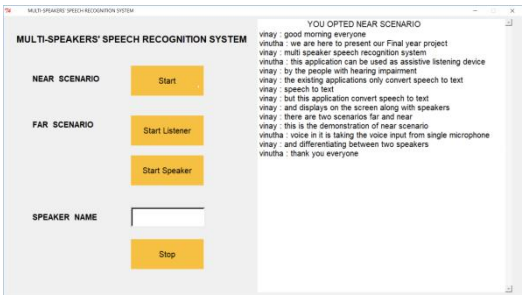

Fig. 8: Snapshot of near scenario for two speakers

Retrieval Number: A10310581C19/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

When the speakers' end the conversation, the end user taps on stop button to exit the application. When the users are far away from each other i.e not in the range of same microphone, they use different systems to run the application. And when they want to have a conversation, they all run the application on their devices which are connected to the same WiFi network. The speakers who want to speak with the listener will add their names in the entry box of the application. Fig 9 and Fig 10 are two different snapshots of the application that are running on different devices, which show the entry of the name in the entry box by speaker 1 and speaker 2 respectively. When speakers tap on the "Start speaker" button of the far scenario. On tapping this button, the chat window displays "YOU OPTED FOR FAR SCENARIO" that indicates the application's initiation. Now the Listener on the third end, runs the application and simply taps on the "Start Listener" button of the application which then displays "YOU OPTED FOR FAR SCENARIO" to let the listener know that the application is initiated. Now the listener sees the conversation happening between the two speakers on his/her chat window.

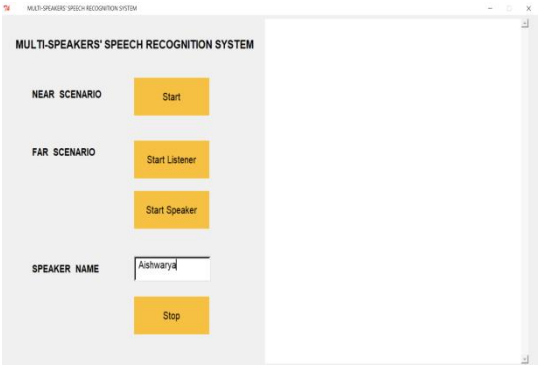

Fig. 9: Snapshot of speaker 1 in far scenario

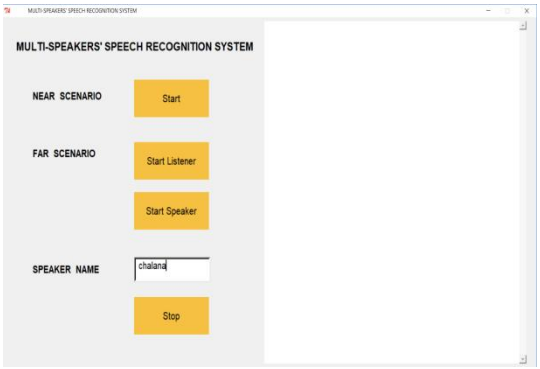

Fig. 10: Snapshot of speaker 2 in far scenario

The snapshot of the Listener's application is shown in the Fig 11. When the conversation ends, the speakers and the listener tap on the stop button to exit the application.

The proposed system performs speaker recognition based on the Gaussian mixture model. Speaker recognition employs the prediction based on test data. In order to measure learning system performance it becomes important to analyse the confusion matrix. A confusion matrix is nxnmatrix which maintains the entries as true values to the predicted values. This confusion matrix is used to calculate accuracy, error rate and precision.

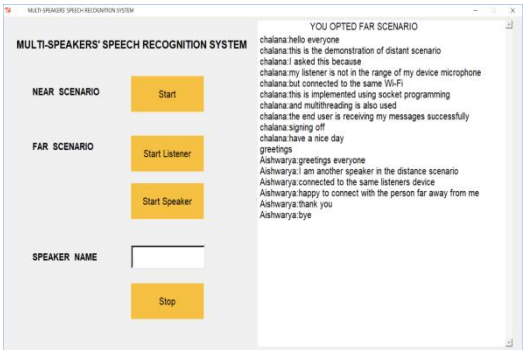

Fig. 11: Snapshot of far scenario for two speakers on listener side

1. Performance Analysis based on Number of Speakers

For two speakers confusion matrix is 2x2 which has true positive (TP), true negative (TN), false positive (FP) and false negative (FN). The false values are across diagonal and are expected to be as low as possible which signifies lesser error rate.

After training this system for two speakers a test was conducted for 100 times which equally consisted trials of each individual. The trials were not performed in any specific manner and the trials were conducted in an environment of minimum noise. The trials resulted in the confusion matrix as shown in Table I.

TABLE I: Confusion Matrix for Speaker recognition with two persons

		Predicted	
Actual		Speaker 1	Speaker 2
	Speaker 1	50(TP)	00(FN)
	Speaker 2	02(FP)	48(TN)

From the confusion matrix the error rate, accuracy and precision for speaker 1 and speaker 2 is calculated based on true positive values and true negative values. Error rate is ratio of number of wrongly recognized trials to the total number of trials calculated as in equation (9).

$$Error\ Rate(ERR) = \left(\frac{FP+FN}{FP+FN+TP+TN}\right)$$

$$ERR = 9/120 = 0.075 \quad (9)$$

Accuracy is the ratio of number of correctly recognized trials to the total number of trials as in equation (10).

$$Accuracy = \left(\frac{TP+TN}{FP+FN+TP+TN}\right) = 98/100 = 0.98 \quad (10)$$

Precision for a speaker indicates the closeness of the obtained result to the actual result which is computed using equation (11).

Retrieval Number: A10310581C19/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

$$Precision\ for\ speaker\ 1 = \left(\frac{TP}{TP+FP}\right) \quad (11)$$

Similarly precision for speaker 2 can be calculated. For three speakers confusion matrix is 3x3 as shown in Table II. The false values are non secondary diagonal elements and are expected to be as low as possible which signifies lesser error rate. After training our system for three speakers a test was conducted for 120 times which equally consisted trials of each individual. The trials were not given in any specific manner and the trials were conducted in an environment of minimum noise. The trials resulted in the confusion matrix as shown in Table II.

TABLE II: Confusion Matrix for Speaker recognition with three persons

		Predicted		
		Speaker1	Speaker2	Speaker3
Actual	Speaker1	38(TP)	01	01
	Speaker2	02	36(TP)	02
	Speaker3	01	02	37(TP)

From the confusion matrix the error rate, accuracy and precision for speaker 1, speaker 2 and speaker 3 is calculated based on true positive values and true negative values. Using equation (9) error rate for three speakers is calculated.

$$ERR = 9/120 = 0.075$$

Accuracy for three speakers is calculated using equation (12)

$$Accuracy = \left(\frac{TP+TN}{FP+FN+TP+TN}\right) = 111/120 = 0.925 \quad (12)$$

Precision for all three speakers can be calculated using equation (11)

2. Percentage Accuracy Analysis

Accuracy of speaker recognition of proposed system depends on: a) Number of speakers b) Length of speech input. When a speaker speaks, any person who cannot see the speaker can make out whether a speaker is a boy/girl/man/woman. So broadly speakers are classified based on their age and gender. Using permutation and combinations of these two parameters, four sets can be formed:

- Same age and same gender
- Same age and different gender
- Different age and same gender
- Different age and different gender

a. Number of Speakers

Here performance of each set is carried out to calculate accuracy of speaker recognition.

i. Accuracy of the speaker recognition when speakers are of same age:

Considering a set of speakers with almost same age, the test was conducted by increasing one speaker each time starting from one. Each test had 25 audio samples of each speaker. Graph of Accuracy v/s Number of speakers is plotted as shown in the Fig 12. The orange colored line represents the accuracy in percentage for the set of speakers with same age but different gender. The blue colored line represents the accuracy in percentage for the set of speakers

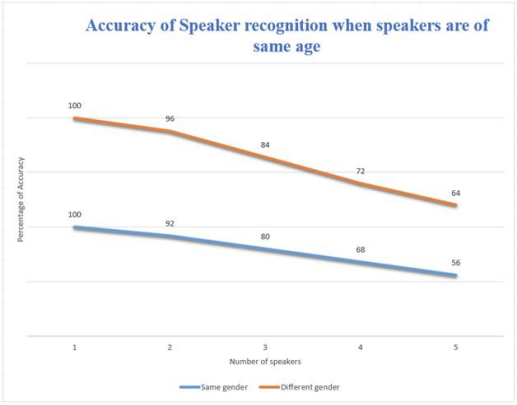

Fig. 12: Graph of Accuracy of the speaker recognition when speakers are of same age

with same age and same gender. As seen in the Fig 12. Accuracy is better for set of speakers with same age and different gender than for the set of speakers with same age and same gender. This is because the set of speakers with different gender will have significant difference in voice characteristics and hence it is easy to distinguish and recognize.

ii. Accuracy of the speaker recognition when speakers are of different age:

Considering a set of speakers with different age, the

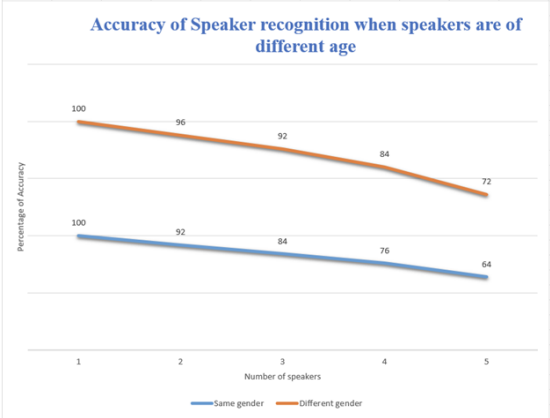

Fig. 13: Graph of Accuracy of the speaker recognition when speakers are of different age

test was conducted by increasing one speaker each time starting from one speaker. Each test had 25 audio samples of each speaker. Graph of Accuracy v/s Number of speakers is plotted as shown in the Fig 13.

The orange coloured line represents the accuracy in percentage for the set of speakers with different age and different gender. The blue coloured line represents the accuracy in percentage for the set of speakers with different age same genders.As inferred before

Retrieval Number: A10310581C19/19©BEIESP
Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

different genders will have significant variations in voice characteristics which results in improved accuracy.

Also with the speaker test set formed of different age groups has higher accuracy because the pronunciation and pitch of syllable varies largely among different age groups, hence it is easy to distinguish and recognize.

On comparing all the four sets of speakers, set of speakers with different age and different gender will give high accuracy. The set of speakers with same age and same gender gives low accuracy among all the sets. The common conclusion in all the sets of speakers is that irrespective of their age and gender, as the number of speakers increases, the accuracy in recognizing the speaker decreases. This is owing to the fact that human voice vary largely over a millisecond period of time and speaker model of the test sample during testing phase may fail to match speaker models in database accurately.

b. Length of Speech Input

The accuracy of speaker recognition also depends on the length of the speech input from a speaker. If length of the speech input is small, then there will be ambiguities in recognizing speaker and sometimes the system fails to recognize the speaker properly. This ambiguities can be minimized by increasing the length of speech input.

As seen in the Fig 14 the accuracy increases as the length of the speech of the speaker increases. This is because increased length of speech provides increased details of speaker's voice characteristics which in turn helps in computing MFCC that are close to the Speaker's model formed during the training phase.

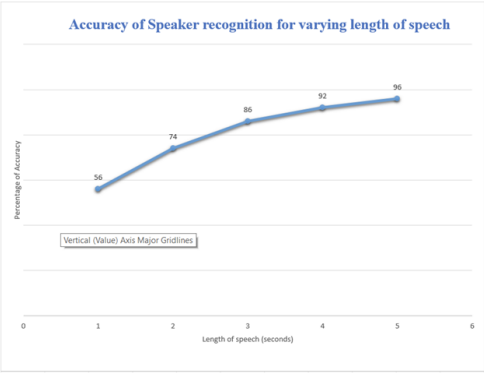

Fig. 14: Graph of Accuracy of Speaker recognition for varying length of speech

V. CONCLUSION

In this paper we proposed speaker recognition system for multi-speaker environment. The existing speech to text conversion systems are helpful for hearing impaired when they are listening to a single person who is speaking in particular with them. But when hearing impaired is in an environment where more than one speaker speaks intermittently, it becomes difficult for hearing impaired to follow up as to who is speaking what using the existing speech to text conversion system. The near scenario proposed by us combines speaker recognition with speech to text conversion which enables the hearing impaired to easily involve in multi-speaker environment by displaying the speech to text converted output along with the name of the speaker. The far scenario proposed by us keeps the hearing impaired informed about the conversations happening among other people around them which may or may not be intended to them in particular. This helps them updated, more involved in their day to day life and help them overcome the feeling of being left out. This improves their quality of life. This scenario increases the range of communication and is mostly suitable for home environment.

REFERENCES

1. FaizanurRehman, Chandar Kumar, ShubashKuma, AtifMehmood, UmairZafar,"VQ Based Comparative Analysis of MFCC and BFCC Speaker Recognition System", International Conference on Information and Communication Technologies (ICICT), pp.28-32, 2017.
2. Lu Xiao-chun, Yin Jun-xun, Hu Wei-ping, "A text-independent speaker recognition system based on Probabilistic Principle Component Analysis", 3rd International Conference on System Science, Engineering Design and Manufacturing Informatization, Volume 1, pp. 255-260, 2012.
3. Rania Chakroun, Leila BeltafaZouari, MondherFrikha, Ahmed Ben Hamida, "A hybrid system based on GMM-SVM for speaker identification", International Conference on Intelligent Systems Design and Applications (ISDA), pp.654-658, 2015.
4. Chandar Kumar, FaizanurRehman, Shubash Kumar, AtifMehmood, GhulamShabir, " Analysis of MFCC and BFCC in a speaker identification system", International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pp.1-5, 2018.
5. U. G. Patil, S. D .Shirbahadurkar, A. N. Paithane, "Automatic Speech Recognition of Isolated Words in Hindi Language using MFCC", International Conference on Computing, Analytics and Security Trends (CAST),pp.433-438,2016.
6. AbhinavAnand , RuggeroDonidaLabati , MadasuHanmandlu ,Vincenzo Piuri , Fabio Scotti," Text-independent speaker recognition for Ambient Intelligence applications by using Information Set Features", IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), pp.30 35,2017.
7. ZunoWatada, Hanayuki,Speech Recognition in a Multi-speaker Environment by Using Hidden Markov Model and Mel-frequency Approach, IEEE Third International Conference on Computing Measurement Control and Sensor Network, pp.80-83, 2016.
8. Teddy Surya Gunawan, Rashida Husain, Mira Kartiwi, "Development of language identification system using MFCC and vector quantization", IEEE 4th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA), pp.1-4, 2017.
9. A. T. Rusli, M. I. Ahmad, M. Z. Ilyas, "Improving speaker verification using MFCC order", International Conference on Robotics, Automation and Sciences (ICORAS), pp.1-4, 2016.
10. Ashwin Nair Anil Kumar, SenthilArumugamMuthukumaraswamy, "Text dependent voice recognition system using MFCC and VQ for security applications", International conference of Electronics, Communication and Aerospace Technology (ICECA), Volume 2, pp.130-136, 2017.
11. MouazBezoui,AbdelmajidElmoutaouakkil,AbderrahimBeni-hssane, "Feature extraction of some Quranic recitation using Mel-Frequency Cepstral Coefficients (MFCC)", 5th International Conference on Multimedia Computing and Systems (ICMCS), pp.127-131, 2016.
12. A. Maazouzi, N. Aqili, A. Aamoud, M. Raji, A. Hammouch, "MFCC and similarity measurements for speaker identification systems", International Conference on Electrical and Information Technologies (ICEIT), pp.1-4, 2017.
13. Ivan Stefanus, R. S. JokoSarwono, MirantiIndarMandasari, "GMM based automatic speaker verification system

Retrieval Number: A10310581C19/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

development for forensics in Bahasa Indonesia", 5th International Conference on Instrumentation, Control, and Automation (ICA), pp.56-61, 2017.
14. Rania Chakroun, Leila BeltafaZouari, MondherFrikha, Ahmed Ben Hamida, "Improving text-independent speaker recognition with GMM", 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), pp.693-696, 2016.
15. Wenyong Lin, "An improved GMM-based clustering algorithm for efficient speaker identification", 4th International Conference on Computer Science and Network Technology (ICCSNT), Volume 1, pp.1490-1493, 2015.
16. JiPibil, Anna Pibilov, JindichMatouek, "Comparison of one and two-level architecture of the GMM-based speaker age classifier", 39th International Conference on Telecommunications and Signal Processing (TSP), pp.299-302, 2016.
17. AlanChern,Ying-HuiLai,Yi-PingChang,YuTsao,RonaldY.Chang, Hsiu-Wen Chang, A Smartphone-Based Multi-Functional Hearing Assistive System to Facilitate Speech Recognition in the Classroom, IEEE ACCESS, Volume 5, pp.10339-10351, 2017.
18. MohsenSadeghi,HosseinMarvi,"OptimalMFCCFeaturesExtraction by Differential Evolution Algorithm for Speaker Recognition", 3rd Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS),pp.169-173, 2017.
19. Sonali T. Saste, Prof. S. M. Jagdale,"Emotion Recognition from Speech Using MFCC and DWT for Security System", International Conference on Electronics, Communication and Aerospace Technology ICECA, pp.701-704, 2017.
20. M.S. Likitha, Sri Raksha R. Gupta, K. Hasitha, A. UpendraRaju, "Speech Based Human Emotion Recognition Using MFCC",IEEE WISP- NET conference, pp. 2257-2260, 2017.
21. Fang-YieLeu, Guan-Liang Lin, "An MFCC-based Speaker Identification System", IEEE 31st International Conference on Advanced Information Networking and Applications, pp.1055-1062, 2017.
22. AnaghaSonawane, M.U.Inamdar ,Kishor B. Bhangale, "Sound based Human Emotion Recognition using MFCC Multiple SVM", International Conference on Information, Communication, Instrumentation and control, pp.1-4,2017.
23. Unnikrishnan V M , Rajeev Rajan, "Mimicking Voice Recognition Using MFCC-GMM Framework", International Conference on Trends in Electronics and Informatics ICEI, pp.301-304, 2017.
24. D. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", Journal Speech Communication, Volume 17, pp.91-108, 1995.
25. ZufengWengLin Li, DonghuiGuo, "Speaker Recognition Using Weighted Dynamic MFCC Based on GMM ", Anti-Counterfeiting Security and Identification in Communication (ASID), pp.285-288, 2010.
26. Diksha Sharma, IsrajAli,"A Modified MFCC Feature Extraction Technique For Robust Speaker Recognition", Advances in Computing, Communications and Informatics (ICACCI), pp.1052-1057, 2015
27. NayanaP.K,DominicMathew,AbrahamThomas,"Performance Comparison Of Speaker Recognition Systems Using GMM and i-Vector Methods with PNCC and RASTA PLP Features", International Conference on Intelligent Computing,Instrumentation and Control Technologies (ICICICT), pp.438-443, 2017.

AUTHORS PROFILE

AishwaryaAnegundiobtained her Bachelor Degree from Sri Jayachamarajendra College of Engineering, Mysore in 2018. She has done internship at Robert Bosch Engineering and Business Solutions during 2018. Her research interests are Digital Signal Processing and Artificial Intelligence.

Vinuta V Pawale obtained her Bachelor Degree from Sri Jayachamarajendra College of Engineering, Mysore in 2018. She has been an active member of IEEE. Her research interests are Data Engineering and Machine Learning.

Chalana D A obtained her Bachelor Degree from Sri Jayachamarajendra College of Engineering, Mysore in 2018. She has been an active member of IEEE. Her research interests are Data Engineering and Machine Learning.

Vinay G obtained his Bachelor Degree from Sri Jayachamarajendra College of Engineering, Mysore in 2018. He has been an active member of IEEE. His research interests are Software Designing and Development.

DrRudraswamy S B obtained his Doctorate degree from Indian Institute of Science, Bangalore in 2015. He graduated from Kuvempu University in 2002 and obtained Masters Degree from Visvesvaraya Technological University in 2006. He was a Commonwealth postdoctoral research fellow at University of Manchester, United Kingdom in 2016-17. He also worked as Visiting Professor at Department of Electrical and Computer Engineering, NJIT, New Jersey's Science and Technology University, USA, 2016.His research work is in the area of Microelectronics and Nanotechnology.

N Manohar received his B.E. degree in Electronics and Communication Engineering from Gulbarga University in 2002, the M.Tech. degree in VLSI Design and Embedded Systems from Visvesvaraya Technological University (VTU), Belgaum, Karnataka in 2007. He was a Design Engineer with AT&S E-Cad Technology, Bengaluru for three years and currently working as a Reader and Head in the Department of Electronics in All India Institute of Speech and Hearing, Mysuru, Karnataka. His area of interests includes Hearing aid Technology Biomedical Instruments, Audio and speech processing.

Retrieval Number: A10310581C19/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication