Smart Voice Recognition System For Online Class Attendance

1st Rohit Karande
Electronics and Telecommunication department
Government College of Engineering
Karad, India
rohitkarande555@gmail.com

2nd Dhanraj Chobhe
Electronics and Telecommunication department
Government College of Engineering
Karad, India
dhanrajchobhe000@gmail.com

3rd Pranav Dere
Electronics and Telecommunication department
Government College of Engineering
Karad, India
pranavdere99@gmail.com

4th Aniket Joshi
Electronics and Telecommunication department
Government College of Engineering
Karad, India
aniket115@gmail.com

***Abstract*— In the current COVID crisis, students are attending classes in an online mode. During offline classes, the teacher used to keep a few minutes after every class for attendance purposes, which was time-consuming. Another problem was when students used to mark proxy attendance. In such a situation, it is difficult to understand if the student has attended the class or not. This problem continues in online mode, which needs a solution. The Smart voice recognition system for online class attendance covers the issue. Just like every person in the world has a unique fingerprint, they also have a unique voice. To solve this problem, an application will be developed where the teacher will initiate the attendance system, the student's voice will be recognized, and their attendance will be marked. This application will make the attendance-taking process very seamless and easy for both teachers and students.**

Keywords*— *Mel frequency Cepstral coefficient (MFCC), Gaussian Mixture Modelling (GMM), Expectation Maximization(EM) algorithm, Voice based Attendance Marking

I. INTRODUCTION

Voice recognition is a technology that is used to identify an individual's voice. There are two steps to identifying an individual's voice. First is the training/verification phase, in which every student model is been created and represented by a GMM. Second is the recognition/identification phase, in which a probabilistic approach is used to predict that the given speech sample belongs to a given speaker (student).

For verification and recognition purposes, we need feature extraction and data classification. For feature extraction purpose, there are numerous algorithms like Linear Predictive Coefficient (LPC), Mel Frequency Cepstral Coefficient (MFCC), and Perceptual Linear Prediction (PLP). Here we have used MFCC as it has low computational complexity and better performance for speech recognition as compared to others. Finally, for data classification purpose, algorithms like K-means clustering, Support Vector Machine (SVM), and Gaussian Mixture Model (GMM) can be used. Here we have used GMM as it has better accuracy as compared to others [3].

A. Feature Extraction

Feature extraction is the conversion of a speech signal into a feature vector. MFCC features are based on the human auditory system. The human perception of the frequency contents of sounds for audio signals do not follow a linear scale so, for each tone with an actual frequency *f* measured in Hz, a subjective pitch is measured on a scale called the Mel Scale. The Mel frequency scale below 1000 Hz has a linear frequency spacing and above 1kHz has a logarithmic spacing [2].

First, the voice data is divided into frames. Hamming window is then applied to each frame. Second, the frames are converted into frequency domain using a short-time Fourier transform. Third, Mel filter bank is used to calculate a certain number of sub-band energies, which is a non-linear-scale filter bank that imitates the human auditory system. Fourth, the logarithm of the sub-band energies are calculated. Finally, MFCC is computed by applying the Discrete Cosine Transform [2].

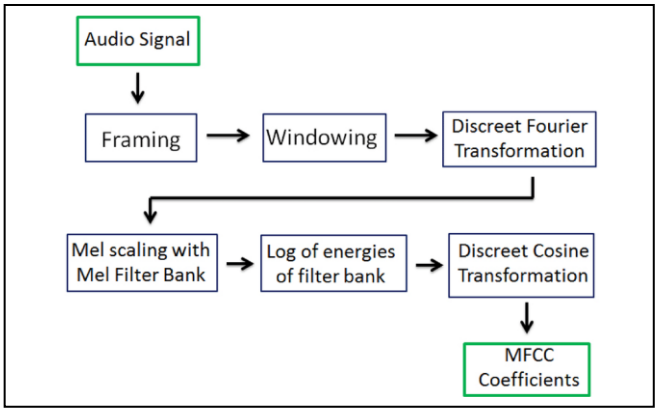

Fig. 1. Steps involved in MFCC extraction [7]

Step 1: The audio signal is a non-stationary signal. For stable acoustic characteristics we need to study the audio signal over a short period of time. So , we divide the audio signal into short frames having a length between 20-40 msec. We overlap the frames so that the frames have a correlation with each other.

Step 2: The discontinuities present at the start and end of the frame causes high frequency distortion effects in the frequency response, so windowing is used to make it continuous. Hamming window is used because it introduces least amount of distortions. Each frame is multiplied with a hamming window in order to keep the continuity of the first

and the last points in the frame. Windowing each frame assures they close to zero at ends. Hamming window defined by:

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad (1)$$

Where, 0≤ n ≤ N−1, N is the window length.

Step 3: Now the signal is in the time domain. To convert it into the frequency domain, we need the signal to be periodic and continuous, which is assured by framing the signal and by applying window on each frame. The Fast Fourier Transform (FFT) is then applied to each frame to extract the frequency components of the signal. Then compute the power spectrum (periodogram) which identifies the frequencies present in each frame using the equation:

$$P = \frac{|FFT(x_i)|^2}{N} \quad (2)$$

where, x_i is the i^{th} frame of signal x [5].

Step 4: The Mel-scale mimics the non-linear perception of sound in the human ear by being more discriminative at lower frequencies and less discriminative at higher frequencies. We achieve that by using the filter bank. Filter bank applies triangular filters, around 40 filters on a Mel-scale to the power spectrum. That will give us information about the energy in each frequency band . Below is figure of filter bank where spacing between the filters grows exponentially with frequency.

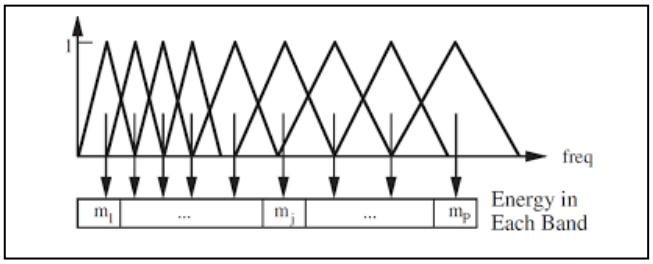

Fig. 2. Mel scale filter bank, from (toung et al,1997).

Step 5: Then we apply log to these spectrogram values to get the log filter bank energy.

Step 6: The final step in generating the MFCC is to apply the Discrete Cosine Transform. Applying DCT to Mel frequency coefficients produces a set of cepstral coefficients. Thus, speech is represented as a sequence of cepstral vector.

B. GMM

A Gaussian mixture model is a probabilistic clustering model which is used for representing normally distributed subpopulation within an overall population. A gaussian mixture is a function which consists of several gaussians each identified by k ∈ {1…, k}, where k is the number of Gaussian distributions/clusters, also called the components of the GMM. For a given data points (feature vectors) the GMM will find the probability to which cluster it belongs. The likelihood of data points for a model is given by following equation:

$$P(X|\lambda) = \sum_{k=1}^{K} \omega_k P_k(X|\mu_k, \sum_k) \quad (3)$$

where $P_k(X|\mu_k, \sum_k)$ is the Gaussian distribution.

$$P_k(X|\mu_k, \sum_k) = \frac{1}{\sqrt{2\pi|\sum_k|}} e^{\frac{1}{2}(X-\mu_k)^T \sum^{-1}(X-\mu_k)} \quad (4)$$

Each Gaussian k in the mixture comprises of the parameters mean μ , co-variance matrices ∑ and weights ω.

Now we need to find these parameters to define the Gaussian distributions. GMM finds them by using the Expectation-Maximization (EM) algorithm. The EM algorithm has two steps. The first is the Expectation step (E-step), and the second is the Maximization step (M-step). We have already decided on a k number of clusters, which means k Gaussian distributions with mean, covariance, and weight values, and then random values are assigned for these parameters. Then we apply EM. The E-step finds the probability that the data point belongs to the cluster/distribution c1, c2... ck. The second step is the maximization step, in which all of the clusters' parameters are updated in iterations until they converge. In this way, each speaker (student) is represented by such a distribution [6].

II. METHODOLOGY

A. Training Phase

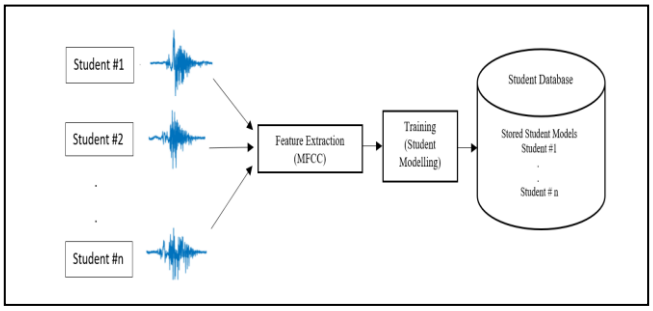

Fig. 3. Steps involved in the training phase

Working :

- Student (or speaker) audio data collection
- Feature Extraction
- Student Model Training using GMM

1) The first step in training phase is to collect audio samples from the students for model training purposes. Four audio files in wav format are collected. The path of all the audio files (4 per student) utilised for training are stored in a file.

2) The next step is feature extraction. The main aim of feature extraction is to convert the audio signal into an N-dimensional feature vector.

3) In order to build a student database from the above-extracted features, we need to model all the students independently now. We employ GMMs for this task. In this step, each student will be represented by a voiceprint, which is nothing but the unique feature of a speech command in the frequency domain. It is merely a matrix of numbers where each number represents the energy heard in a particular frequency band. This will be used as a reference in the recognition phase [4].

B. *Recognition phase*

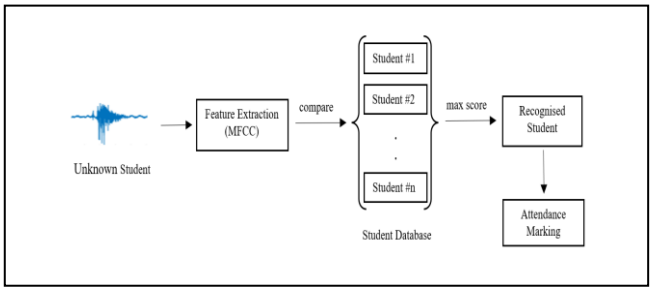

Fig. 4. Steps involved in the recognition phase

Working :

- Unknown students audio file
- Feature extraction
- Student recognition
- Attendance marking

1) First we need the audio files from all the students whose attendance has to be marked in wav format.
2) The next step is feature extraction. It is done both in training and recognition phases.
3) In this step, the maximum likelihood of x is estimated, where x is the unknown student whose voice has to be detected. We calculate the likelihood score with respect to each student by substituting the value of μ and ∑ in the likelihood equation shown previously. The student model with the highest likelihood score is considered the identified student [1].
4) Then comes the last step, attendance marking. After the student has been recognized, his/her attendance will be marked as present, while others will be marked as absent.

III. Results

Here below is a snippet of result when the web application is used. The result contains name of the student who's voice has been detected after giving the audio file as input.

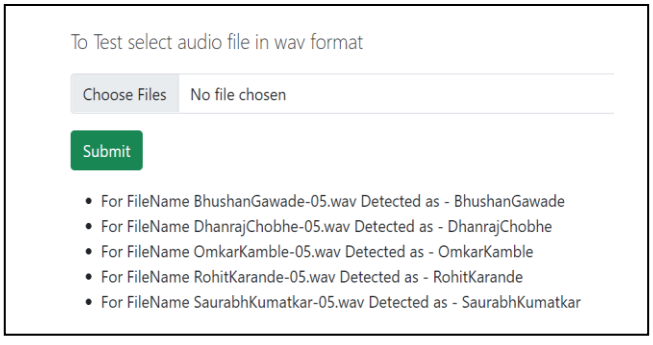

Fig. 5. Recognized students' names

After the voice recognition, the detected student attendance is marked in the excel sheet as "*present*." All the remaining students are marked as "*absent"* as shown below.

	A	B
1	**Name of Student**	2022-02-20
2	AjayKadam	Absent
3	BhushanGawade	Present
4	DhanrajChobhe	Present
5	KankshitRamteke	Absent
6	OmkarKamble	Present
7	PradnyeshChoudhari	Absent
8	PranavDere	Absent
9	PromitPanja	Absent
10	RohitKarande	Present
11	SaurabhKumatkar	Present
12	TejasBattise	Absent
13	VishalLodha	Absent
14	RahulNavgire	Absent
15	SheryaYadav	Absent
16	SwapnilPatole	Absent

Fig. 6. Attendance marking in the excel sheet

IV. Conclusions

A total of 15 students' voice data was collected and their models were trained using MFCC and GMM. We had a 90% success rate. Hence, MFCC and GMM can be used for attendance marking purposes with high accuracy. It also saves a lot of time for teachers as compared to the traditional method of calling the names of each and every student one by one to mark their attendance. This results in the process of marking attendance fast and there are very few chances of proxy attendance, which was a huge problem before.

References

[1] Abhijeet Kumar, "Spoken Speaker Identification based on Gaussian Mixture Models".

[2] K. J. Patil, P.H. Zope & S. R. Suralkar, "Emotion Detection from Speech using MFCC & GMM".

[3] Md Masudur Rahman, Debopriya Roy Dipta and Md. Mahbub Hasan, "Dynamic Time Warping Assisted SVM Classifier for Bangla Speech Recognition".

[4] Mustafa Yankavis, "Feature Extraction Mel Frequency Cepstral Coefficients (MFCC)".

[5] Haytham M. Fayek, "Speech processing for Machine Learning: Filter banks, Mel Frequency Coefficients (MFCCs) and What's in-between".

[6] Aishwarya Singh, "Build Better and Accurate Clusters with Gaussian Mixture Model".

[7] Figure 1, "A lazy learning-based language identification from speech using MFCC-2 features Scientific Figure on ResearchGate".