BUET Bus Entrance System using Voice Recognition

A S Al Mahmud Sajid
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
1906050@eee.buet.ac.bd

Mohammad Ismail Chowdhury
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
1906058@eee.buet.ac.bd

Sayed Mohammad Tawsif Arefin
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
1906057@eee.buet.ac.bd

Ismam Nur Swapnil
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
1906064@eee.buet.ac.bd

***Abstract*— Our team has developed a digital system for BUET Bus Service that utilizes voice recognition technology to replace the traditional system and provide convenience for students. The software component of the system utilizes the MFCC algorithm to extract voice features. This involves calculating cepstral coefficients at the Mel frequency scale and utilizing vector quantization and Euclidean distance methods for feature matching in MATLAB. During the training phase, the voice pattern of a student is compared to patterns stored in the student database. We have conducted a thorough quality and testing analysis of the speaker recognition system and determined that the accuracy is greatly improved with the use of effective noise cancelling devices during voice input.**

Keywords*— *MFCC, FFT, DCT, Vector quantization, Euclidean distance.

I. INTRODUCTION

The BUET bus service is a crucial transportation option for attached students, but obtaining tickets presents numerous difficulties. Furthermore, the inconvenience persists as students must carry their tickets and risk losing them. To alleviate this problem, we have endeavored to digitalize the entire system. The proposed system is speaker-dependent and speech-dependent, requiring students to state their name and ID into a microphone before boarding or testing. The audio data is sampled and transformed into feature arrays, representing the unique voice patterns of each person. The MFCC algorithm is used to extract these patterns, which are then compared to previously stored patterns in the student database using the vector quantization method. This allows the software to identify the best match for the input. When a successful input is detected and the system recognizes the student, they can board the bus and the fare will be charged to their bank account.

II. LITERATURE REVIEW

Text-dependent speaker recognition systems have been designed following many algorithms over the past years. Among different methods, MFCC along with vector quantization methods are used in this project because of the simplified yet moderately high-accuracy algorithm [1][2]. To increase the precision,10 training data sets are used, and a better output is found. As a classifier, KNN is another one with the comprehensible method [3], but with the increase in data set in columns, the Euclidean distance method gives better output. Besides, there are various methods of voice recognition perfectly using more complex algorithms like Hidden Markov Models (HMM) [4]. The architecture of neural networks helps in this field [5]. DNN and end-to-end automatic speech recognition methods all are used nowadays for complex scenarios [6]. With the comparison of all the methods, MFCC with vector quantization is easy to work with high accuracy and maintenance.

III. METHODOLOGY

A. Voiced/ unvoiced/ silence detection and silence removal:

The voiced parts of speech demonstrate higher correlation among successive samples and contain more energy and useful information compared to the unvoiced and silent sections. Proper segmentation for removing this redundant data not only reduces computation but also enhances speech processing accuracy, as the silence/unvoiced portions are more vulnerable to noise. Initially, the data was sampled at 44100 samples per second and framed at Fs/100. The parameters μ and σ were obtained using the first 8820 samples, which follow a normal distribution and contain background/white noise. Any x sample that meets the criterion $(x-\mu)/\sigma >$ the adjusted threshold is considered part of the background noise distribution and is removed from the speech. Voiced and unvoiced samples were then tallied per frame, classifying the frames as voiced or unvoiced. The resultant signal with removed silence was created using only the voiced frames.

B. Framing

In MFCC feature extraction, the speech signal is divided into frames using a technique called framing. This involves splitting the signal into overlapping frames, with each frame containing a certain number of samples. A Hamming window is chosen for the frame to remove spectral leakage. The number of frames is rounded down to the nearest integer by default, and any remaining samples at the end of the signal are ignored, unless its length is a multiple of the frame increase in

samples (inc) plus the frame size (w). If the 'z' or 'r' options are used, the number of frames is rounded up, and zero padding or reflecting of the last few samples is done for the final frame.
The total number of frames can vary based on the number of samples. Initially, the frame size was set at 100ms, but to improve efficiency, hamming windowing was applied with a length of 20ms, resulting in a frame size of 20ms.

C. Calculating Discrete Fourier Transform (DFT):

After partitioning and applying a window function to the speech signal, we calculated the DFT to determine the constituent frequencies for further analysis. In order to excite the computation process, we utilized the built-in **fft function** of MATLAB to calculate the DFT. This transform is defined by

$$Y(k) = \sum_{j=1}^{n} X(j)\, W_n^{(j-1)(k-1)}$$

where,

$$W_n = e^{(-2\pi i)/n}$$

and X(j) was the time domain speech after framing and windowing. Here 1≤k≤K, where K=the size of hamming window = n= number of samples = 882.

D. Calculating Power Spectrum

After calculating the DFT for each frame of the pre-processed speech signal, the next step in MFCC feature extraction is to calculate the power spectrum. This involves taking the square of the absolute value of the DFT coefficients. The power spectrum provides information about the energy distribution across different frequency bands and is used to calculate the Mel filter bank energies. The average power of each frame is taken following the equation

$$\mathbf{P(k)= N\_^{-1} \sum_{i=1}^{N} P_i(k) \text{ where } P_i(k)=|Y(k)|^2.}$$

Here Y is the DFT calculated in the previous step.

E. Mel Filter Bank

The Mel filter Bank is used to convert the linearly spaced frequency domain obtained after DFT of a speech signal into a non-linear mel-scale frequency domain. This is because the human ear does not perceive frequency in a linear manner, and the mel-scale is a more accurate representation of the way we hear sound. The formula for mapping the actual frequency to the frequency that human beings will perceive is given below:

$$\mathbf{mel(f)=1127*ln\ (1+\frac{f}{700})}$$

The mel filter bank consists of a set of triangular bandpass filters placed uniformly along the mel-scale frequency axis. To create the Mel filter bank, the lowest and highest frequencies were converted into MEL units and divided into 32 filter banks with equal spacing. The resulting points were then converted back into Hertz using the inverse MEL equation and rounded to the nearest frequency bins to avoid spectral leakage. Then, triangular filter banks are formed by:

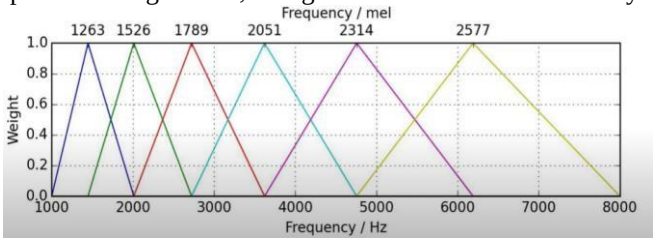


Figure1: Triangular Filter Banks

The Mel filter bank groups periodogram bins to estimate energy in different frequency ranges. The filters start narrow at 0 Hertz and widen as frequencies increase to maintain perceptual frequency distance.

F. Discrete Cosine Transform (DCT)

After calculating the power spectrum using the Mel filter bank, we use the Discrete Cosine Transform (DCT) and obtain a compact representation of the spectral envelope of the speech signal. The resulting coefficients are the Mel Frequency Cepstral Coefficients (MFCCs), which are commonly used in speech recognition tasks. In our system, the filter banks were overlapping and highly correlated with each other. To decorrelate the energies, we applied Discrete Cosine Transform (DCT) and kept only 15 of the 32 coefficients. The higher DCT coefficients represent fast changes in the filter bank energies which can degrade speech processing performance, hence dropping them slightly improved our system's performance. We prepared the function of DCT following the equation:

$$y(k,l) = w(k)\sum_{m=1}^{M} u(m,l)\cos\frac{\pi(2m-1)(k-1)}{2M}, \quad k = 1,\ldots,M$$

where

$$w(k) = \begin{cases} \frac{1}{\sqrt{M}}, & k = 1 \\ \sqrt{\frac{2}{M}}, & 2 \le k \le M \end{cases}$$

G. Vector Quantization

Vector quantization refers to the technique of mapping vectors from a huge vector space to a smaller set of discrete vectors in that space. Each area is referred to as a cluster and can be depicted by its center called a centroid [2,4]. The VQ encoder encodes a given set of k-dimensional data vectors with a much smaller subset. The subset is called a codebook and its elements Ci are called codewords, code vectors, reproducing vectors, prototypes, or design samples. Only the index i is transmitted to the decoder. The decoder has the same codebook as the encoder, and decoding is operated by table look-up procedure. The most used vector quantizers are based on nearest neighbor called Voronoi or nearest neighbor vector quantizer. Both the classical K-means algorithm and the LBG algorithm belong to the class of nearest neighbor quantizers.
A key component of pattern matching is the measurement of dissimilarity between two feature vectors. The measurement of dissimilarity satisfies three metric properties such as Positive definiteness property, Symmetry property and Triangular inequality property. Each metric has three main characteristics such as computational complexity, analytical tractability and feature evaluation reliability. The metrics used in speech processing are derived from the Minkowski metric. The Minkowski metric can be expressed as

$$D_p(X,Y) = \sqrt[p]{\sum_{i=1}^{k} \left|x^i - y^i\right|^p},$$

Where $X = \{x^1, x^2, \ldots, x^k\}$ and $Y = \{y^1, y^2, \ldots, y^k\}$ are vectors and p is the order of the metric.

The performance of the vector quantizer can be evaluated by a distortion measure D which is a non-negative cost D (X j, X j) associated with quantizing any input vector X j with a reproduction vector X j. Usually, the Euclidean distortion measure is used. The performance of a quantizer is always qualified by an average distortion Dy = E [D (X j Xj)] between the input vectors and the final reproduction vectors,

where E represents the expectation operator. Normally, the performance of the quantizer will be good if the average distortion is small.

H. Euclidean Distance Measurement

Euclidean distance is a measure of the distance between two points in Euclidean space. The voice of an unknown speaker is represented by a series of feature vectors (x1, x2,...xi) during the speaker recognition phase, and it is then compared to codebooks from the database. Based on reducing the Euclidean distance, it is possible to determine the speaker's identity by calculating the distortion distance between two vector sets [6]. The Euclidean distance is calculated using the following formula:

The Euclidean distance between two points $P = (p_1, p_2 \cdots p_n)$ and $Q = (q_1, q_2 \ldots q_n)$,

$$= \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

The speaker with the lowest distortion distance is chosen to be identified as the unknown person.

IV. RESULTS AND DUSCUSSION

Here, we have shown our result of our code.firstly we showed oud the confusion matrix. Then we can see that our snr plot was much higher than previous years plot which was our expectation. As we need to detect voice samples in noisy environment, our code must have a higher accuracy in low snr. We have been able to do this using weiner noise suppression.

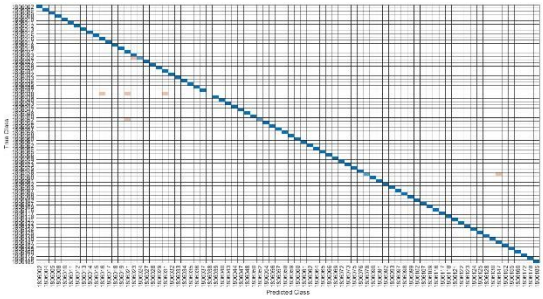


Figure2: Confusion Matrix for Name Audio Samples

There are generally two cases by which we can reduce noise from a audio sample. One of those methods are weiner noise reduction. It is a statistical approach by which we remove our noisy part of the signal estimating the noise power.

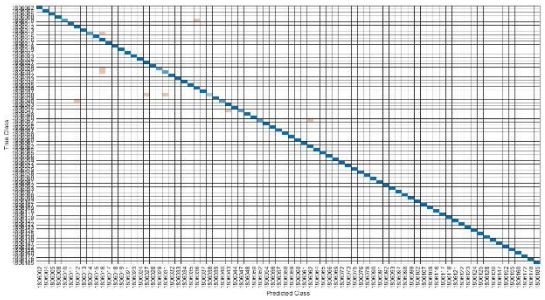


Figure3: Confusion Matrix for ID Audio Samples

"sensitivity"	1×103 double	0.92557	0.92557
"specificity"	1×103 double	0.99927	0.99927
"accuracy"	1×103 double	0.92557	0.92557

Figure4: Performance Table based on the Name Audio Samples

"sensitivity"	1×103 double	0.93851	0.93851
"specificity"	1×103 double	0.9994	0.9994
"accuracy"	1×103 double	0.93851	0.93851

Figure5: Performance Table based on the ID Audio Sample

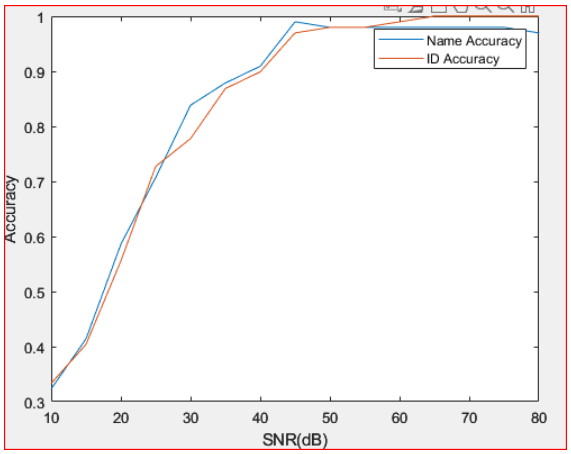


Figure 6: Accuracy vs SNR (dB)

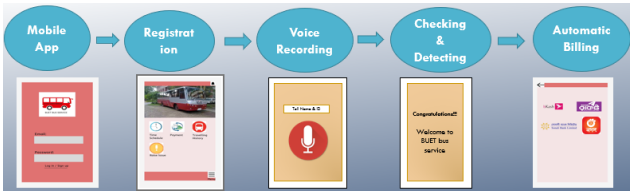
Here, we can see a significant increase in accuracy when the noise is lower. The accuracy and F1 scores are above 85% when the SNR is above 45dB.Hence, using a better noise reduction algorithm is needed to expect better performance from the designed system in noisy environment where SNR is less than 45dB.

V. IMPLEMENTATION

Our project will be implemented through a mobile app using our program. BUET students willing to use bus service have to register through the app using their name, id, email and bank account information. While boarding bus the conductor of the bus will collect the voice sample of the students through the app in the phone provided. The app will verify whether they are the registered ones or not. If verified successfully, bus fare will be deducted automatically from their account which they selected while registration.

We will have to invest a huge money for purchasing server domain for uploading our database. Moreover, we have to spend money for app design and providing smartphones to

the conductor. Our senior did propose an electric gate system for this project. But it would be very time consuming and also costly approach. Also it will need extra maintenance. In comparison, our mobile app based system would be more efficient and less time consuming. All in all, our proposed system can lessen the daily hassle of BUET students in bus of present system to a great extent.



VI. LIMITATIONS

Main limitation of our project is the data collection in proper environment. Every voice should be recorded in same settings. Otherwise, the accuracy of our system will get lower. Moreover, our present program is not capable of coping with modern situations. We need to apply machine learning for more accuracy and efficiency.

VII. MODIFICATIONS

We applied some modifications for increasing the accuracy of the project than our previous batch. We collected more data and trained them more accurately. Moreover, we used Weiner Filter for removing noise. All these increased our accuracy than our previous batch.

VIII. FURTHER IMPROVEMENT

For making our project more practical and increasing its accuracy we need to collect more realistic voice samples. As we are working with bus service we need to collect data from the surrounding of bus. So, we can set up a booth in the BUET bus stop to collect the voice samples. Furthermore, we can apply machine learning for more improved and accurate system.

IX. CONCLUSION

The goal of our project was to build up a voice-based bus entry system. For this purpose, we took 10 name samples and 10 ID samples collected from 120 students. We trained these data with our code. We used MFCC for feature extraction and KNN for feature matching. Then we took input voice to match with the samples.

Our system can accurately identify users 97% times. However, because our project is built for the roads, in high noise scenario, the percentage of success may fall. So, we need a good hardware setup to get maximum success from our project. Our budget was estimated to be 3,10,000 tk one time mainly for app design and server domain purchase and 25,000 taka monthly for maximum output. This may vary considering new hardware setup. Costing may seem higher, but it can serve all the voice based system of BUET through the database preserved in the server and interlinked. Our project's accuracy will have been more practical if we collected voice sample from bus environment.

ACKNOWLEDGMENT

We would like to thank everyone who helped us in developing our project. We are humbly grateful to Professor Dr. Celia Shahnaz mam and Md Jahin Alam sir for giving us a chance to present our project and for giving us constant guideline, valuable insights. We are also thankful to our classmates who provided the necessary training and testing data which helped us frame and test our project to the better end.

REFERENCES

[1] Shaneh, M., & Taheri, A. (2009). Voice command recognition system based on MFCC and VQ algorithms. *World Academy of Science, Engineering and Technology*, 57, 534-538.

[2] Patel, K., & Prasad, R. K. (2013). Speech recognition and verification using MFCC & VQ. *Int. J. Emerg. Sci. Eng. (IJESE), 1(7)*, 137-140.

[3] Faek, F. K., & Al-Talabani, A. K. (2013). Speaker recognition from noisy spoken sentences. *International Journal of Computer Applications*, *70*(20).

[4] L. Rabiner, "A tutorial on Hidden Marcov Model and selected applications in Speech Recognition", Proceedings of the IEEE, Vol. 77, No. 2, 1989, pp.257-286

[5] Patricia Melin, Jerica Urias, Daniel Solano, Miguel Soto, Miguel Lopez, and Oscar Castillo, "Voice Recognition with Neural Networks, Type-2 Fuzzy Logic and Genetic Algorithms"

[6] Jiang, W., Wen, F., & Liu, P. (2018). Robust beamforming for speech recognition using DNN-based time-frequency masks estimation. *IEEE Access*, *6*, 52385-52392.

[7] Kiran, U. (2021, June 13). MFCC Technique for Speech Recognition. AnalyticsVidhya. https://www.analyticsvidhya.com/blog/2021/06/mfcc-technique-for-speech-recognition

[8] Balwant A. Sonkamble (2012) Speech Recognition Using Vector Quantization through Modified K-meansLBG Algorithm (https://core.ac.uk/download/pdf/234644507.pdf)