# Speech Recognition of Throat Microphone using MFCC Approach

**Subrata Kumer Paul*, Rakhi Rani Paul***

*Lecturer, Dept. of Computer Science and Engineering, Bangladesh Army University of Engineering and Technology (BAUET), Qadirabad Cantonment, Natore-6431.

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract -** *The Throat Microphone (TM) is a piece of non-acoustic equipment, depending on the vibrations of vocal folds rather than the audible sound generated. Correctly, apprehending vocal fold vibrations is challenging due to poor signal representation inclinations. The method identifies the TM vibrations and offers a similar speech sound. It is done by extracting features from the spectrum of the TM vibrations and analyzing the collected features with the values stored in a database. The extracted features introduce unique features of the speech waveform called Mel-Frequency Cepstral Coefficients (MFCC). The determination of the most like speech signal is accepted by the minimum mean square error estimation approach, where the signal in the database whose corresponding MFCC values present the minimum variance from the input speech MFCCs is chosen. It can be useful for natural Human-Machine communication especially for vocal tract affected people, Speech Recognition, Identification, Speech Enhancement System and so on. It can be used efficiently for patients who have lost their voices due to injury or illness. It provides excellent communications in high noise environments.*

***Key Words:* Throat Microphone, Mel Frequency Cepstral Coefficient, vocal fold vibrations, Minimum Mean Square Analysis, Discrete Cosine Transform, Fast Fourier Transform, Linear Prediction Analysis.**

## 1. INTRODUCTION

A throat microphone is a microphone that absorbs vibrations from the throat of the user through single or dual sensors on the neck. Since the sensors are worn on the neck, the sensors can pick up speech even in extremely noisy or windy environment, for e.g. traveling on a motorcycle, in a nightclub or turbulence in a fighter jet. Other types of microphones do not function well under these conditions because of high levels of background noise. The TM, being unaffected by noise and degradation, it is possible to extract the required information effectively even under different environmental conditions [1].

The project aims at using a throat microphone to capture the vocal fold vibrations of a person, recognize the vibrations in an existing database and play the corresponding speech signal. In this way, the system can aid people who have lost their voices due to some medical condition, illness or surgery like tracheotomy by giving them some voice of expression. It is essential that the vocal fold vibrations of these patients are accurate as the microphone records only these vibrations.

## 2. LITERATURE REVIEW

What makes the throat microphone distinct from other speech sensors is definitely its ability to pick up vocal fold vibrations irrespective of ambient noise. It eliminates many disadvantages posed by the normal microphone, which captures audible sound, as the sound is more susceptible to background noise. The close proximity of the transducers to the throat enables its use in noisy environments [2]. The GEMS (Glottal Electromagnetic Micro Power Sensor), developed by Aliph Corporation, transmits low-power electromagnetic waves to the glottis and analyses the reflected signal [2]. The movements of the tissue detected by the reflected signal including opening and closing phases of the glottis via a small antenna located on the throat are recorded [3].

The EGG (Electro-Glotto-Gram) sensor measures vocal fold contact area through an electrical potential (of about 1 V RMS and 2–3 MHz) across the throat at same point as the larynx. The waveform of vocal fold dynamics and relative contact patterns during continuous speech production (phonation) are provided [2]. The opening and closing of the vocal folds are measured by the EGG [4].

The P-mic is another microphone that relies on vibrations permeating a liquid-filled chamber. The microphone has a gel-filled chamber and a piezo-electric sensor behind the chamber. The liquid-filled chamber is designed such that it has poor coupling between the ambient noise and the fluid-filled pad, thus successfully attenuating vibrations of the unwanted background noise [4].

The microphone converts electric signals into mechanical vibrations and then captures the converted sound from the internal ear [1]. The Throat Microphone detects the intelligible sounds produced by the pharynx through the throat tissues. TM readings are preferred over conventional microphones as they capture pitch and some partial format structure [2].

In all of these literature about throat microphone, we understand that this system has the potential of having applications for giving voice to those with defective speech and in military communications.

## 3. FEATURE EXTRACTION

One of the first steps for speech processing is feature extraction. Just like how it is implemented in 2D images, speech waveforms also possess features that make it distinct from other waveforms. Like acoustic speech, a parametric representation of the TM vibrations is required for further analysis and processing, also called signal processing front end [6]. Since speech is a signal that slowly varies over time, examination over short periods of time gives stationary characteristics. Hence short-time spectral analysis is the most common way to characterize a speech signal [6]. Speech analysis is done by extracting a feature in such a way that the features of different speech sounds may be different. There are a number of ways that makes this possible. The methods include:

a) Linear Prediction Analysis (LPA) – extracting Linear Predictive Cepstral Coefficients in time domain.

b) Mel-Frequency Cepstrum Coefficient (MFCC) – extracting frequency domain coefficients.

The work performed in [2] makes use of Linear Prediction Analysis to perform feature extraction. However, computations in time domain are slower and more complex, and they also yield inaccurate results. In this project, the features extracted are MFCCs from the speech waveforms in the database and compare with input waveforms. MFCCs are extracted from the speech waveforms using the MFCC processor specified in [6]. The algorithm used for MFCC extraction is explained in brief in the next section.

## 4. MFCC ALGORITHM PROCESS STEPS

The final objective of the MFCC processor is to extract the MFCCs of the speech waveforms and store the MFCC values. The MFCC values are a set of integer values that vary from a large negative value to a large positive value. This is generated using a series of steps specified in the block diagram specified in Fig 1. The algorithm is specified in [7]. The blocks are explained in brief below:

1) ***Frame Blocking:*** The continuous speech signal is divided into frames containing N samples separated by M samples with the adjacent frame. Here, the first frame would consist of the first N samples and the second frame begins M samples after the first frame, overlapping it by a number of N-M samples. This procedure continues till all the speech signal samples are accounted for.

2) ***Windowing:*** This is done to minimize the signal discontinuities at the beginning and end of each frame. The spectral distortions are minimized by using the window to help the signal reduce itself to zero at the beginning and end of each frame. Here the Hamming window is used.

3) ***Fast Fourier Transform (FFT)*****:** The next processing step is computing FFT, which converts each frame of N samples from the time domain into the frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT) which is defined on the set of N samples {x(n)}, as follows:

$$\boldsymbol{y_l(n) = x_l(n) * w(n); 0 \le n \le N-1} \text{............(1)}$$

Hence the spectrum of the signal is obtained.

4) ***Mel-Frequency Wrapping*****:** A subjective pitch is measure on a 'Mel' scale for each tone with an actual frequency f in Hz. The Mel-frequency scale shows linear frequency spacing below 1000 Hz and a logarithmic spacing greater than 1000 Hz. The pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mel as a reference point. Therefore, the following approximate formula can be used to compute the Mel for a given frequency f in Hz.

$$\boldsymbol{Mel(f) = 2595 log_{10}(1 + \frac{f}{700})} \text{..................(2)}$$

One approach to simulating the subjective spectrum is to use a filter bank, one filter for each desired Mel-frequency component. The filter bank has a triangular band-pass frequency response, and the spacing and the bandwidth is determined by a constant Mel-frequency interval.

5) ***Obtain MFCC:*** After step 4, the log Mel spectrum is converted back to time. The result is called the Mel-Frequency Cepstrum coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis.

6) ***Discrete Cosine Transform (DCT):*** so their logarithm, are real numbers, they can be converted to the time domain using the Discrete Cosine Transform (DCT). Here, twenty (20) MFCC values were generated for a particular speech signal.

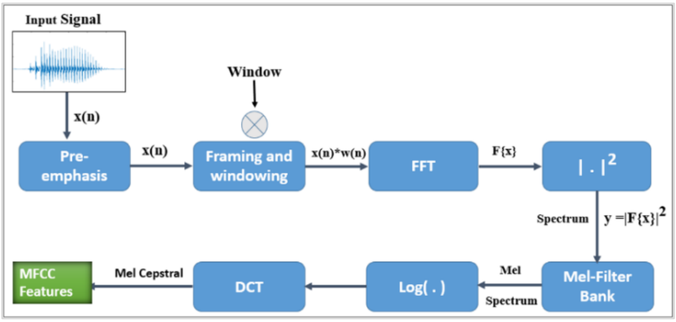

**Fig -1:** Block Diagram of MFCC process [6]

## 5. SYSREM DETAILS

The system recognizes input TM vibrations by extracting features from the vibrations and comparing the obtained features to values in a database. They are approximated to the stored values in the database and the corresponding acoustic speech is given as output. The software has been implemented using MATLAB. The system consists of two different phases:

- ***Creation of a Database:*** The database will store pre-recorded TM vibrations and the corresponding acoustic speech.
- ***Recognition and Detection:*** TM vibrations are input to the system and the values generated by the vibrations are compared to the already stored values of TM vibrations in the database. The vibrations in the database with which the input vibrations closely resemble are selected and the corresponding acoustic speech is sent as output.

### ❖ Creation of a Database:

The vocal fold vibrations are captured by the throat microphone. Simultaneously, the speech signal that is responsible for the particular vibration is recorded and feature extraction is done. Hence the database is created for various vibrations and speech signals. The block diagram is described by Fig 2.

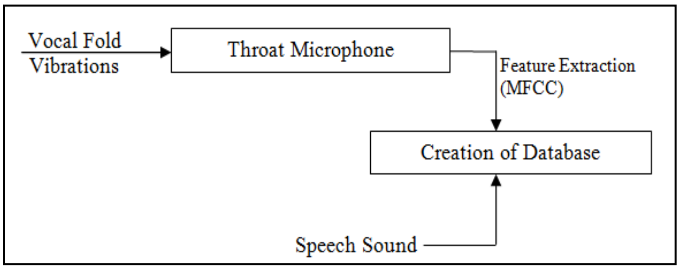

**Fig 2:** Creation of Database

### ❖ Recognition and Detection:

The input vocal fold vibrations are captured using the throat microphone and subjected to feature extraction and approximated to resemble closely to one of the sets of values that are already stored in the database. Those vibrations that resemble those in the input are selected from the database and the corresponding acoustic signal is selected as the output. The block diagram is described in Fig 3.

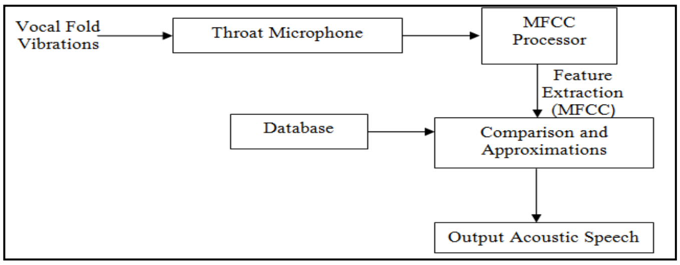

**Fig 3:** Recognition and Detection

## 6. COMPARISON AND APPROXIMATIONS

This section details how the comparison takes place between the input speech MFCC values and the values in the database. The MFCC coefficients, as mentioned earlier, are characteristic of a particular speech waveform. Under the hypothesis that identical speech waveforms would have the same MFCCs, the desired value is determined in the database by Minimum Mean Square Error (MMSE) estimation.

$$e = \sum_{n=1}^{20} \frac{(x_1(n) - x_2(n))^2}{20} \quad \ldots\ldots\ldots\ldots(3)$$

Where, e is the error matrix, $x_1(n)$ is the input speech MFCC and $x_2(n)$ is the value of database speech MFCC. The denominator denotes the taking of the mean value of the squared difference. In this case, 20 MFCC values are considered.

## 7. EXPERIMENT RESULT & DISCUSSION

In this study, a database for 600 speech values was created. The speech sounds were taken from 8 individuals (3 males and 5 females) consisting of three words, 'good', 'bad' and 'best'. The words were so chosen to give distinction between the words used and to reduce the complexity. The three words had equal representation in the database, i.e. 200 samples of each word were used.

**Table -1:** About the Database

| **Total values** | **600** |
|---|---|
| No. of male values | 200 |
| No. of female values | 400 |

Over 100 speech samples (8 speakers) of the three words were given as input to the system and it was checked if the system correctly identified the correct word from the data-base. The following table is a concise form of the results:

**Table -2:** Results of the Experiment

| | Input values | Total values |
|---|---|---|
| Male speakers | 15 | 100 |
| Female speakers | 85 | |

The case study using 100 speech samples produced an accuracy of nearly 91%. Here, the accuracy is the ratio of correctly identified words to the total number of words.

**Table -3:** Summary of the Results

| | Total Input | Correctly Identified | Incorrectly Identified | Accuracy |
|---|---|---|---|---|
| Male speakers | 15 | 11 | 4 | 73% |
| Female speaker | 85 | 80 | 12 | 94% |
| Total words | 100 | 91 | 16 | 91% |

## 8. CONCLUSION

There is a great need to increase the precision of the system. Incorporating machine learning to enable continuous learning of the reference values and updating the database accordingly will bring in more accuracy of recognition. In-creasing the precision would be beneficial to those suffering from various speech disabilities life cleft palate and cleft lips that have reduced their voice clarity.

## 9. ACKNOWLEDGEMENT

Foremost, I would like to express my sincere gratitude to my coauthor Rakhi Rani Paul for the continuous support of my research, for her patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my research study. This work was supported in part by a grant from co-author.

Last but not the least, I would like to thank my family my parents for giving birth to me at the first place and supporting me spiritually throughout my life.

## 10. REFERENCES

[1] M.A.Tuğtekin Tura and Engin Erzin, “Source and Filter Es-timation for Throat-Microphone Speech Enhancement”, IEEE/ACM Transactions On Audio, Speech And Language Processing, Vol.24, No.2, February 2016.

[2] K. Sri Rama Murty, Saurav Khurana, Yogendra Umesh Itankar, M. R. Kesheorey and B.Yegnanarayana, “Efficient Representation of Throat Microphone Speech”, 2008 ISCA

[3] K. Brady, T. Quatieri, J. Campbell, W. Campbell, M. Brand-stein, and C. Weinstein, “Multisensor MELPe using parameter substitution,” in Proc. Int. Conf. Acoust., Speech, Signal Pro-cess (ICASSP), May 2004, vol. 1, p. I–477–80, vol.1

[4] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, and K. Shikano, “Accurate hidden markov models for non-audible murmur (NAM) recognition based on iterative supervised adaptation,” in Proc. IEEE Workshop Autom. Speech Recogn. Understand. (ASRU), Nov. 2003, pp. 73–76.

[5] M. Arun Marx, G.Vinoth, A. Shahina, A. Nayeemulla Khan, “Throat Microphone Speech Corpus for Speaker Recogni-tion”, MES Journal of Technology and Management

[6] Hamdi Boukamcha, “Speaker Recognition in MATLAB”, ICEM2,Jun. 2009, pp. 33-35

[7] Kshamamayee Dash, DebanandaPadhi, Bhoomika Panda, Prof. Sanghamitra Mohanty, “Speaker Identification using Mel Frequency Cepstral Coefficient and BPNN”

## BIOGRAPHIES

**Subrata Kumer Paul** was born in Bangladesh in 1993. He completed his B.Sc. and M.Sc. Engineering from Rajshahi University, in Computer Science and Engineering 2016 and 2018 respectively. Now, he is working as a Lecturer at Bangladesh Army University of Engineering and Technology (BAUET), Natore, Bangladesh. His research field is Speech Signal Processing, Data Mining, and Machine Learning.

**Rakhi Rani Paul** was born in Bangladesh in 1996. She graduated in 2017 from Rajshahi University, in Computer Science and Engineering. Her M.Sc. Engineering (degree) is pursuing. Now, she is working as a Lecturer at Bangladesh Army University of Engineering and Technology (BAUET), Natore, Bangladesh. Her research field is Speech Signal Processing, Data Mining, and Machine Learning.