Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

Electrolaryngeal Speech Identification using GMM

Nandana R., Nissie Mary Johnson, Jubily V. Reji,
B. S. Shreelakshmi, Ancy S. Anselam,
Lani Rachel Mathew
Department of Electronics and Communication Engineering
Mar Baselios College of Engineering and Technology
Trivandrum, India

K. Gopakumar
Department of Electronics and Communication Engineering
TKM College of Engineering, Kerala

Abstract—**People who have lost their larynx due to laryngeal cancer or due to other physical conditions cannot produce voice like other humans. As air is breathed out through the vocal folds, vocal folds are vibrated and sound is produced, heard as a speech voice. In situations where larynx is removed, air can never again go from the lungs into the mouth. The association between the windpipe and the mouth never again exists. Electrolarynx is a battery driven machine that produces sound to create a voice for laryngectomy patients. The electrolarynx could either function indirectly through contact with skin, that triggers pharyngeal vibrations or specifically by intraoral pressure, which produces vibrations of the vocal cavity. Articulation muscles are usually intact after total laryngectomy(TL) and therefore capable of transforming the stimulus noise supplied into understandable voice. Traditional electrolarynx produces a robotic voice and a mechanical humming when used. This paper focuses on increasing the quality of electrolaryngeal speech. Here, implementation of Mel Frequency Cepstral Coefficient (MFCC) Algorithm and Gaussian Mixture Model (GMM) in pair is used to achieve the target. We have considered MFCC with "tuned parameters" as the primary feature and delta-MFCC as secondary feature. And, we have implemented GMM with some tuned parameters to train our model. Speech with highest score is identified and corresponding normal speech is produced using python platform.**

Keywords—Laryngectomy; electrolarynx; DWT; gaussian; cepstral coefficients

I. INTRODUCTION

Humans produce sound by utilizing the voice box which is called larynx. The voice confinement is arranged in throat at the highest point of the windpipe. The human voice box contains two tendons known as vocal cords. Vocal cords present over the larynx stretch so that it leaves a limited space between them for the entry of air. At the point when the human talks, muscles present in the larynx get extended and the opening becomes smaller. When air is made to go through the cut, the vocal lines vibrate. As vibrations increases, higher intensity of sound is delivered.

Total laryngectomy is the elimination of larynx by surgery. People who have lost their larynx lack the ability to produce sound through the normal sound production mechanism. There are three vocal rehabilitation techniques for TL: electrolarynx, esophageal speech, and TEP. Esophageal speech is achieved with swallowed air through an esophageal insufflations process. The air released is guided from the esophagus, causing the mucosa of the upper esophagus to vibrate. Patients might feel that esophageal voicing with restricted air supply is difficult to produce. TEP with sound prosthesis helps patients guide tracheal air in the rear tracheal surface through such a cut site and push it into the esophagus. Patients with no ready access to speech-language pathologists would find it difficult to learn how to use and take care of TEP prosthesis. The electrolarynx has some benefits compared with other approaches. Electrolarynx is a small unit, powered by battery and hand held. The gadget's vibrating coupler plate is held opposite the neck. The coupler circle vibration or signal is transmitted into the vocal tract which routes the sign along those lines to generate sound. The vibrations are produced by a vibrator powered by an external battery (known as an electrolarynx or an artificial larynx) that is usually placed on the cheek or under the jaw. It creates a humming vibration that reaches the user's throat and mouth. The person then uses his / her mouth to modify the sound to articulate the sounds of speech.

EL speech identification and following conversion to normal speech makes the EL speech more intelligible. Words are distinguished using MFCC feature extraction algorithm. Identification is done based on this distinguishment, using GMM model. Using python programming platform identified word is played in normal voice. Using this system difficulty in communication with laryngectomees can be eliminated.

II. RELATED WORKS

A strategy that evaluates F0 shape for electrolaryngeal (EL) speech upgrade in Mandarin investigates the utilization of syllabic element to enhance the nature of electrolaryngeal speech has been proposed in paper[1]. Utilization of phonetic element for F0 form age as opposed to the acoustic element is made. Trial results show that the technique accomplishes outstanding improvement in regard to the understandability and the comparability with ordinary speech. In any case, the EL speech does not seem like a human delivered voice in a few different ways: one is the sound quality debases because of the noise created by the nonstop vibration of the EL and second one is EL speech sounds unnatural on the grounds that it is created by the mechanical excitation and third one is the intelligibility is constrained since the EL imperfect speech. Kasuya and Kikuchi [2] developed an EL model that alters F0 according to down or side wise finger developments recognized by a diode (Drove) and photograph sensor. The gadget which got greater speech coherence in Japanese and English, and it created with a F0 strategy for Mandarin utilizes development of a trackball. Users could modify the F0 of the speech by simply moving his thumb on the T-EL contact board. The speaker could create Mandarin tones to look like ordinary speech after some basic preparing. Two sorts of arrangements can be recognized[3]; one is an open-loop strategy in which EL speech is recorded, prepared and exhibited utilizing amplifiers or by means of media transmission applications like loudspeaker and the other one is a closed loop technique in which EL speech is recorded, handled and afterward a counterfeit excitation signal is produced and encouraged back to the EL gadget. These two methodologies have various potential outcomes, can utilize various strategies and will, positively get various outcomes. Changing fundamental frequency pattern can be introduced by an automatic process based on numerical models. GMMs are probabilistic models that serve to model arbitrary distributions of probabilities. In paper[4], Martin Hangmuller and his co-authors suggested a new transducer that depends on electromagnetic mechanisms. The updated transducer's technological properties exhibit major benefits over the traditional electro-dynamic transducer. The traditional method of voice conversion transforms frame by frame spectral parameters based on the square error of the minimum mean. Conversion method estimating the highest likelihood of a spectral parameter path is implemented in paper[5].

III. METHODS

A. *Using MFCC Algorithm*

Most present speech acknowledgment frameworks utilize concealed Markov models (HMMs) to manage the transitory fluctuation of speech and Gaussian blend models to decide how well each condition of each HMM fits an edge or a short window of edges of coefficients that speaks to the acoustic information. An elective method to assess the fit is to utilize a feed forward neural system that accepts a few coefficients as information and produces back probabilities over HMM states as yield. Deep neural systems with many concealed layers that are prepared utilizing new strategies have been appeared to outflank Gaussian blend models on an assortment of speech acknowledgment benchmarks, now and then by a huge edge. The exactness can likewise be improved by expanding (or linking) the info highlights (e.g., MFCCs) with "couple" or bottleneck highlights produced utilizing neural systems. GMMs are fruitful to such an extent that it is hard for any new technique to outflank them for acoustic displaying.

1) Generation of coefficients: MFCC ends up taking human speech to frequencies into account and is therefore the safest way to recognize speech / speaker. It is best to analyse a speech signal with respect to frequency due to its cyclic behaviour.

2) Pre-Emphasis: Pre-emphasis aims to compensate the high frequency component silenced during human sound processing process. Signal-to-noise ratio is maximized using preemphasis.

3) Frame Blocking: The discourse signal is sectioned into casing of 15~20 ms. Normally the the size of the frame equals the power of two inorder to enable the FFT usage. If this is not the situation, zero padding is done to the closest length of intensity of two.

4) Hamming Window : To preserve the consistency of the first and last closures in the frame, each frame must be increassed with hamming window. Let x(n) be the signal in a casing, then y(n) the signal obtained after increasing with Hamming windowing is given by,

$$y(n) = x(n) * w(n) \qquad (1)$$

where w(n) is the Hamming window defined by,

$$w(n) = 0.54 – 0.46 * \cos (2n /(N-1)) \qquad (2)$$

where 0 <n< N-1

5) Fast Fourier Transform (FFT): FFT is employed to acquire the size frequency reaction of each casing. When FFT is employed on a casing, it is accepted that the sign inside an edge is occasional, and persistent when folding over. Otherwise when this is not the situation, FFT can in any case be employed yet the irregularity at the edge's first and last directs probably toward present unfortunate impacts in the frequency reaction.

B. *Using GMM*

This method includes two phases; training and testing. Training part includes feature extraction and creation of Gaussian Mixture model whereas testing comprises of computation of log likelihood ration of each input and comparison with models created. Methodology used can be summarized in 5 basic phases.

1) Data Acquisition: Data set of EL speech was made with the help of a laryngectomee patient. Model’s working and accuracy is tested on self-made data set.

2) Data preprocessing: The data must be pre-processed in order to achieve better outputs and prediction results. This is to ensure that the model is trained with minimum errors.

a) Noise Reduction + Silence Removal : Noise leads to a loss of device efficiency. The de-noise method is made using the technique of wavelet decomposition. The de-noising method includes breaking down the original signal, thresholding the information parameters and restoration of the signal. The portion of de-noising decomposition is accomplished via the DWT. The Discrete Wavelet Transform (DWT) is mostly used with multi-rate filter banks, that are filter collections which divide a band of signal frequencies into sub channels. Filter banks include high-pass, band pass or low-pass filters. If the banks are wavelet filter banks including high pass and low pass wavelet filters, then low pass filter output will be the quantization coefficients. The detail coefficients will be the outputs of the high-pass filter, too. The method of having coefficients for description and estimation is called decomposition. If a threshold function is implemented to the DWT output, and wavelet coefficients which are less than the threshold are discarded, then a "de-noising" function is performed by the system. This task could be completed by using a music/ audio editing software; named Audacity.

b) Conversion from .mp3 to .wav format: Further converted all of the mp3 files into .wav file format. Since *Scipy.iofile.wav* module for reading only *.wav* files is used.

3) Feature Extraction: Two main features MFCCs and its derivatives, Delta-MFCC are focused. 20 coefficients each were calculated. So, totally 40 features in hand. Delta MFCC was calculated by a custom defined function under featureextraction.py module.

4) Model Training: We implemented the GMM approach for model training. The system module can be separated into four modules:

a) Front- end processing: The part of "signal processing," that transforms the sampled speech signal into a set of feature coefficients, characterizing the characteristics of speech which can differentiate different words. The objective of this step is to modify the speech signal, so that It would be more relevant for an overview of the feature extraction.

b) Modeling: Modeling technique aims to develop models to every speech using extracted unique feature vector. It performs a function for data reduction by modeling feature vector distributions. sklearn.mixture is a package which allows one to learn, sample and estimate Gauussian Mixture Models from data.

c) Database: The speech models are stored here. These models are obtained for each word by using feature vector extracted. These models are used for identification of words during the testing phase.

d) Decision logic: Final decision regarding the word spoken is made, through comparing unknown word to every models in the self-made database and maximum suiting one is chosen.

5) Perform Identification: In the model training process, the log-likelihood for each .gmm model of each speech sample was determined. This had been placed in a different folder as a database. This data dictionary is used for matching 1: N sample's gmm file. The sample with the highest score is chosen and identified.

i. Block Diagram

Proposed block diagram is as indicated in Fig. 1.. Two phases of the project is differentiated. Training phase includes the steps from data set creation to model training. Whereas testing phase comprises of the steps from comparison of input EL sample to output in normal intelligible voice.

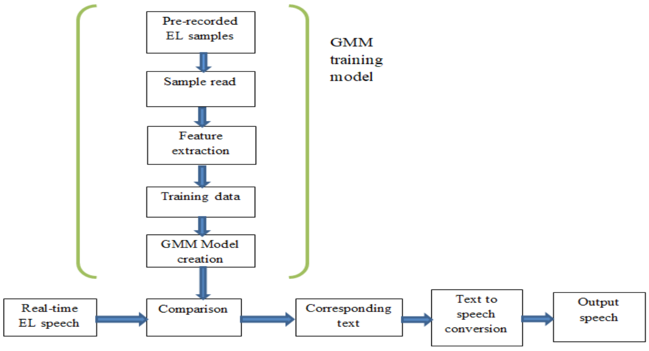


Fig. 1. Block diagram of the system

- Pre recorded EL &Sample read: Data will be read from the pre-recorded sample produced by EL.
- Feature extraction: It is based on frequency domain the usage of Mel scale. A unique record from voice information that may later be used to identify the speech is extracted. The foremost intention of feature extraction is to reduce the four lengths of the speech sign before the popularity of the signal. Mel frequency cepstral coefficients are obtained using MFCC algorithm. To calculate delta feature from MFCCs, the following equation is applied,

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \qquad (3)$$

where 'N' is number of deltas summed over. Typically taken as 2.

- Training data: The concept of GMM training include to estimate the probability distribution in a class which uses a sequential mixture of 'k' Gaussian density mostly known as GMM additives. The probability for a model with data sets is calculated by the following expression,

$$P(X|\lambda) = \sum_{k=1}^{K} \omega_k P_k(X|\mu_k, \sum k) \qquad (4)$$

where $P_k(X|\mu_K, \sum K)$ is the Gaussian distribution

$$P_k(X|\mu_K, \sum K) = \frac{1}{\sqrt{2\Pi \sum K}} e^{\frac{1}{2}(X-\mu_K)^T \sum -1 (X-\mu_K)} \qquad (5)$$

The training data X$_i$ referring to class λ are being used inorder to approximate the matrices of co-variance Σ, mean μ, weight ω, of components and similar parameters. k clusters within data are identified initially, by K-means algorithm then every cluster is valued equally with weight ω=1/k. Such k-clusters are thus equipped with 'k' gaussian distributions. Up until it converges, all clusters' parameters ω, μ, and σ are modified in iterations. For such estimations Expectation Maximization (EM) algorithm has been the most popular method.

- GMM creation: A Gaussian Mixture Model (GMM) is a stochastic opportunity distribution function defined as a biased sum of the densities of Gaussian components. It is used as a stochastic version of oppurtunity density of measuring functions in biometric systems. GMM is represented through its Gaussian distribution and every Gaussian distribution is calculated by means of its mean, variance and weight of the Gaussian distribution.

- Comparison of real time EL speech with GMM training model: Incoming speech signal is sampled at a frequency of 44100 Hz, is in comparison with the GMM model and the corresponding speech is recognized.
- Text to speech conversion: Normal speech which is already saved as data set is chosen corresponding to identified text.

ii. Working

Gaussian Mixture Model is used as a classifier to evaluate the function extracted from the MFCC Algorithm with the saved templates. Firstly, the data will be read from the pre-recorded sample produced by means of EL. These samples get feature extracted and it extract 40 dimensional capabilities from speech frames. Extracting the features aims to lessen the dimensions of the speech before recognizing it. MFCC cepstral coefficients are available from the output of MFCC block. These samples are converted to training data and the concept of GMM training include to estimate the probability distribution in a class which uses a sequential mixture of 'k' Gaussian density mostly known as GMM additives. The speech signal is diagnosed by way of the use of Gaussian Mixture Model. At the same time speech signal is recorded via using electrolaryngeal device with a sampling frequency of 44100 Hz and its miles saved as a wave file through the use of sound recorder software in Python. Wav files are transformed into corresponding text samples the use of Python software program's command and these text samples again transformed to speech samples. The silence part of the speech as well as background noise is eliminated at the initial stage of processing. For our purpose we converted speech signal to normal human voice, by the use of python library file pyAudioAnalysis.

Similarly it can be done in real time where the input from the microphone is read and the corresponding word will be identified for which the data sheet is created.

IV. Results

A. De-noising in Audacity

Noise removal and Silence removal are done manually one by one through a GUI interface of the software. Audacity, version 2.3.1 published by Audacity Development Team is used in Ubuntu operating system.

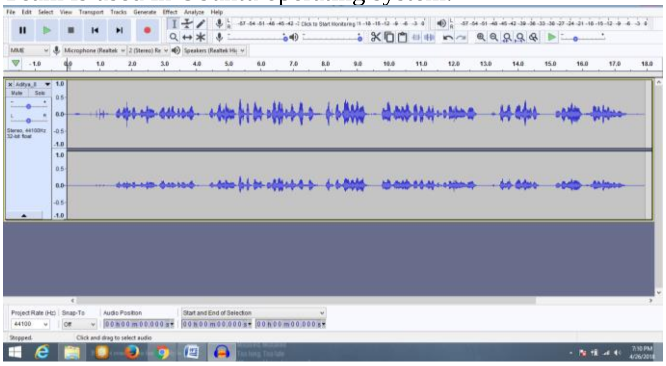
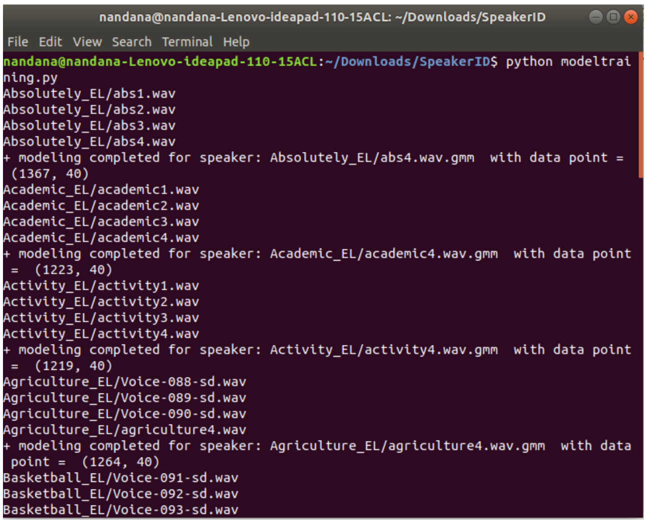

File Edit Select View Transport Tracks Generate Effect Analyze Help
Click to Start Monitoring
Project Rate (Hz)
Snap-To
Audio Position
Start and End of Selection
44100
Off
Stopped.
Click and drag to select audio

Fig. 2. Output from Audacity

B. GMM Training

Gaussian Mixture Model is trained using 4 samples per word which are of different modulation. This process is repeated for every set of EL speech in the data set. Path of saved EL speech files and storage of model is provided. Model corresponding to each EL speech is prepared based on the feature extracted.

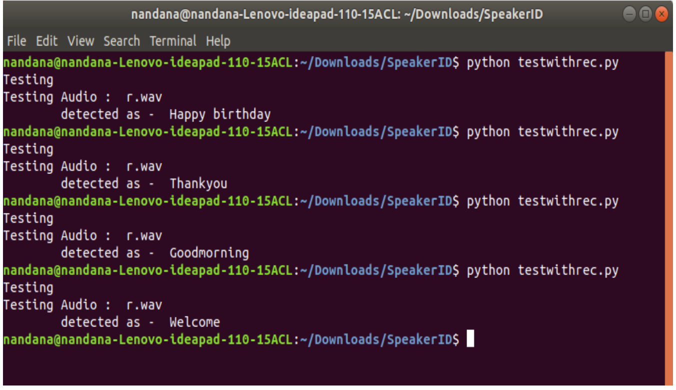


```
nandana@nandana-Lenovo-ideapad-110-15ACL: ~/Downloads/SpeakerID
nandana@nandana-Lenovo-ideapad-110-15ACL:~/Downloads/SpeakerID$ python modeltrai
ning.py
Absolutely_EL/abs1.wav
Absolutely_EL/abs2.wav
Absolutely_EL/abs3.wav
Absolutely_EL/abs4.wav
+ modeling completed for speaker: Absolutely_EL/abs4.wav.gmm  with data point =
 (1367, 40)
Academic_EL/academic1.wav
Academic_EL/academic2.wav
Academic_EL/academic3.wav
Academic_EL/academic4.wav
+ modeling completed for speaker: Academic_EL/academic4.wav.gmm  with data point
 =  (1223, 40)
Activity_EL/activity1.wav
Activity_EL/activity2.wav
Activity_EL/activity3.wav
Activity_EL/activity4.wav
+ modeling completed for speaker: Activity_EL/activity4.wav.gmm  with data point
 =  (1219, 40)
Agriculture_EL/Voice-088-sd.wav
Agriculture_EL/Voice-089-sd.wav
Agriculture_EL/Voice-090-sd.wav
Agriculture_EL/agriculture4.wav
+ modeling completed for speaker: Agriculture_EL/agriculture4.wav.gmm  with data
 point =  (1264, 40)
Basketball_EL/Voice-091-sd.wav
Basketball_EL/Voice-092-sd.wav
Basketball_EL/Voice-093-sd.wav
```

Fig. 3. Model Training

C. EL speech Identification using pre-recorded audio

Audio file which consist of the word that should be identified, is stored in a pre defined location. As the code is executed features from the audio is extracted and comparison with trained models is done. Out of which the one with highest score is chosen as the result. Resultant word is played in normal intelligible voice.

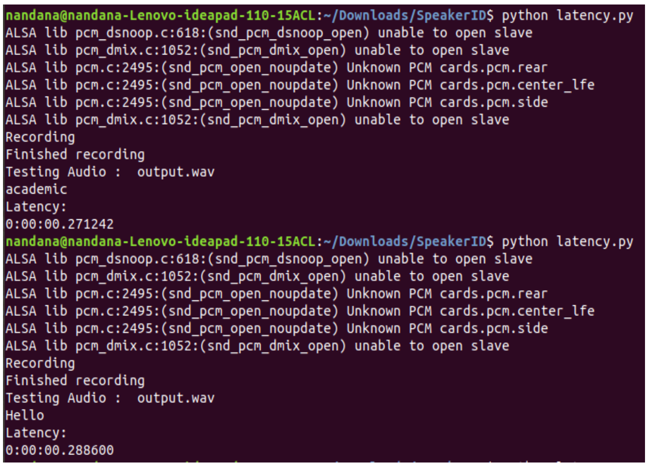


```
nandana@nandana-Lenovo-ideapad-110-15ACL: ~/Downloads/SpeakerID
nandana@nandana-Lenovo-ideapad-110-15ACL:~/Downloads/SpeakerID$ python testwithrec.py
Testing
Testing Audio :  r.wav
        detected as -  Happy birthday
nandana@nandana-Lenovo-ideapad-110-15ACL:~/Downloads/SpeakerID$ python testwithrec.py
Testing
Testing Audio :  r.wav
        detected as -  Thankyou
nandana@nandana-Lenovo-ideapad-110-15ACL:~/Downloads/SpeakerID$ python testwithrec.py
Testing
Testing Audio :  r.wav
        detected as -  Goodmorning
nandana@nandana-Lenovo-ideapad-110-15ACL:~/Downloads/SpeakerID$ python testwithrec.py
Testing
Testing Audio :  r.wav
        detected as -  Welcome
nandana@nandana-Lenovo-ideapad-110-15ACL:~/Downloads/SpeakerID$
```

Fig. 4. Detection of pre-recorded EL

D. EL Identification in real time

EL speech to be identified could be given as input using microphone in real time. As the code is executed recording occurs for first 3 seconds. Features of this recorded audio is extracted and compared with trained models. Out of which the one with highest score is chosen as the result. Resultant word is played in normal intelligible voice.



```
nandana@nandana-Lenovo-ideapad-110-15ACL:~/Downloads/SpeakerID$ python latency.py
ALSA lib pcm_dsnoop.c:618:(snd_pcm_dsnoop_open) unable to open slave
ALSA lib pcm_dmix.c:1052:(snd_pcm_dmix_open) unable to open slave
ALSA lib pcm.c:2495:(snd_pcm_open_noupdate) Unknown PCM cards.pcm.rear
ALSA lib pcm.c:2495:(snd_pcm_open_noupdate) Unknown PCM cards.pcm.center_lfe
ALSA lib pcm.c:2495:(snd_pcm_open_noupdate) Unknown PCM cards.pcm.side
ALSA lib pcm_dmix.c:1052:(snd_pcm_dmix_open) unable to open slave
Recording
Finished recording
Testing Audio :  output.wav
academic
Latency:
0:00:00.271242
nandana@nandana-Lenovo-ideapad-110-15ACL:~/Downloads/SpeakerID$ python latency.py
ALSA lib pcm_dsnoop.c:618:(snd_pcm_dsnoop_open) unable to open slave
ALSA lib pcm_dmix.c:1052:(snd_pcm_dmix_open) unable to open slave
ALSA lib pcm.c:2495:(snd_pcm_open_noupdate) Unknown PCM cards.pcm.rear
ALSA lib pcm.c:2495:(snd_pcm_open_noupdate) Unknown PCM cards.pcm.center_lfe
ALSA lib pcm.c:2495:(snd_pcm_open_noupdate) Unknown PCM cards.pcm.side
ALSA lib pcm_dmix.c:1052:(snd_pcm_dmix_open) unable to open slave
Recording
Finished recording
Testing Audio :  output.wav
Hello
Latency:
0:00:00.288600
```

Fig. 5. Detection in real time

E. Latency Test

Total internal time taken by the system to recognize an EL word can be obtained by latency test. Latency is obtained by excluding time of recording and time of playing normal sound out of the total test. Hence latency of the test is same as the time taken to obtain likelihood ratio and selecting the model with maximum value. Let total time of the test be T_{tot}, time taken for obtaining likelihood ratio and selection be T_{lat}, time for recording be T_{rec} and time for playing normal sound be T_{normal}. Equation for total test time is given by,

$$T_{tot}= T_{rec}+T_{lat}+T_{normal} \tag{6}$$

From the above equation latency can be obtained as

$$T_{lat}= T_{tot}-(T_{rec}+T_{normal}) \tag{7}$$

Latency obtained ranges from 0.25 to 0.3 approximately, out of which average is obtained as 0.28seconds.

F. Result Analysis

Self made data set of EL speech comprises of 41 set of words each word set have 4 samples each which differ in voice modulation. Hence in total data set consist of 164 EL words. In word identification using pre-recorded samples, out of 164 words 161 words are identified correctly. That is 98.17% of accuracy. In real time testing, 152 words are identified correctly. That is 92.68% of accuracy has been obtained.

TABLE I. RESULT ANALYSIS

Total number of EL samples		164
Number of samples detected correctly	Pre-record	161
	Live	152
Success rate	Pre-record	98.17%
	Live	92.68%
Average latency		0.28 seconds

CONCLUSION

The study of present technologies of speech enhancement shows that there are certain limitations for it. Even electrolarynx which is a suitable and helpful technique for voice recovery for patients who have had laryngectomees has various disadvantages like humming sound etc. The proposed model will be able to overcome these limitations so that the listener does not feel discomfort when they listen to a patient. It consist of GMM trained using MFCC feature extracted from EL speech of different tones, so that same speech produced in different manner could be recognized efficiently. New mechanical advancements will permit the electrolarynx to be used without any difficulty in patients and improve the capacity to talk with increasingly differed modulations.

REFERENCES

[1] Z. Cai, Z. Xu and M. Li, "F0 Contour Estimation Using Phonetic Feature in Electrolaryngeal Speech Enhancement," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 6490-6494.

[2] W. Li, Q. Zhaopeng, F. Yijun and N. Haijun, "Design and Preliminary Evaluation of Electrolarynx With F0 Control Based on Capacitive Touch Technology," in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 3, pp. 629-636, March 2018.

[3] Malathi, P. & G R, Suresh and Moorthi, M., "Enhancement of electrolaryngeal speech using Frequency Auditory Masking and GMM based voice conversion" in *Research Gate*, pp. 1-4, 2018.

[4] M. Hashiba, Y. Sugai, T. Izumi, S. Ino and T. Ifukube, "Development of a wearable electro- larynx for laryngectomees and its evaluation," *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Lyon, 2007, pp. 5267-5270.

[5] A. K. Fuchs, M. Hagmüller and G. Kubin, "The New Bionic Electro-Larynx Speech System," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 5, pp. 952-961, Aug. 2016.

[6] Akila. S, Karpagameena. U, Kasthuri V. K, Ms. L. Padmini, "Neural Network based New Bionic Electro Larynx Speech System", in *International Journal of Engineering Research & Technology (IJERT) ICONNECT* , vol. 5, no. 13, pp. 1-5, 2017

[7] S. Bhattacharyya, T. Srikanthan and P. Krishnamurthy, "Ideal GMM parameters & posterior log likelihood for speaker verification," *Neural Networks for Signal Processing XI: Proceedings of the 2001 IEEE Signal Processing Society Workshop (IEEE Cat. No.01TH8584)*, North Falmouth, MA, USA, 2001, pp. 471-480, doi: 10.1109/NNSP.2001.943151.

[8] Xin-xing Jing, Ling Zhan, Hong Zhao and Ping Zhou, "Speaker recognition system using the improved GMM-based clustering algorithm," *2010 International Conference on Intelligent Computing and Integrated Systems*, Guilin, 2010, pp. 482-485, doi: 10.1109/ICISS.2010.5655122.

[9] Ming Li, Jangwon Kim, Adam Lammert, Prasanta Kumar Ghosh, Vikram Ramanarayanan, and Shrikanth Narayanan, "Speaker verification based on the fusion of speech acoustics and inverted articulatory signals," Computer speech & language, vol. 36, pp. 196–211, 2016.

[10] Tomoki Toda, Alan W Black, and Keiichi Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2222–2235, 2007.