Speaker Recognition Improvement for Degraded Human Voice using Modified-MFCC with GMM

Amit Moondra[1], Dr Poonam Chahal[2]
Researcher[1], Professor[2]
Department of Computer Science Engineering, Manav Rachna International Institute of Research and Studies, Faridabad, Haryana, India[1, 2]

***Abstract*—Speaker's audio is one of the unique identities of the speaker. Nowadays not only humans but machines can also identify humans by their audio. Machines identify different audio properties of the human voice and classify speaker from speaker's audio. Speaker recognition is still challenging with degraded human voice and limited dataset. Speaker can be identified effectively when feature extraction from voice is more accurate. Mel-Frequency Cepstral Coefficient (MFCC) is mostly used method for human voice feature extraction. We are introducing improved feature extraction method for effective speaker recognition from degraded human audio signal. This article presents experiment results of modified MFCC with Gaussian Mixture Model (GMM) on uniquely developed degraded human voice dataset. MFCC uses human audio signal and transforms it into a numerical value of audio characteristics, which is utilized to recognize speaker efficiently with the help of data science model. Experiment uses degraded human voice when high background noise comes with audio signal. Experiment also covers, Sampling Frequency (SF) impacts on human audio when "Signal to Noise Ratio" (SNR) is low (up to 1dB) in overall speaker identification process. With modified MFCC, we have observed improved speaker recognition when speaker voice SNR is upto 1dB due to high SF and low frequency range for mel-scale triangular filter.**

***Keywords*—GMM; artificial intelligence; MFCC; fundamental frequency; mel-spectrum; speaker recognition**

I. Introduction

Voice as a human identity is still a challenge for machine. Now businesses need to have effective speaker recognition though his/her voice in silent or in noise environment. Medical industries are also moving toward human voice based medical diagnosis. So human voice is not only identity of a person but also an aid in medical problem diagnosis. A human's voice is identical to himself and always different from each other. Human creates audio from mouth and throat. Fundamental frequency is prime frequency, and it is transformed due to different vocal cord shapes and that create every human's voice as identical voice.

Any speaker recognition process starts with voice feature extraction which is highly dependent on voice characteristics. Speaker recognition system is always designed similar to the mechanism that humans use to recognize two different speakers. A human generally identifies another human's voice based on few classifications. Generally, humans first identify speaker's gender, whether it's male voice or female voice. Key differentiation between female and male voice is tone/pitch and frequency. After gender identification human brain considers multiple voice features to classify speaker. Every person's voice is identical, and it's based on different voice characteristics like amplitude, pitch, frequency, jitter and spectral power. Male voice has 0-900 hz as prime or fundamental frequency and similarly fundamental frequency for female voice is 0-1500Hz. If we average out male fundamental frequency, then it is 110 hz and female average fundamental frequency is 211 Hz [1].

Human lungs generate air pressure to start sound and this sound is articulated by vocal cord. Teeth, jaw, and tongue are main articulators of vocal tract which modulate fundamental frequency (F0) [2][3]. This air pressure creates sound with prime or fundamental frequency. Fundamental frequency sound is transformed by the vocal tract to generate different sound changes. Human prime frequency or fundamental frequency is base frequency for any speaker recognition system. Following sub sections define different key voice features which are used for feature extraction in speaker recognition process.

A. Fundamental Frequency and Pitch

Human voice signal consists of different sine waves with multiple frequencies. If we segregate all these waves and identify lowest sinusoidal wave frequency, then that frequency is considered as fundamental frequency (F0) of the human voice signal. F0 is also considered to calculate the pitch of the human audio signal. The perception of F0 and equivalent harmonics is generally known as voice pitch. The fundamental frequency value should come into a certain frequency range. Else, the pitch of the human voice signal is either extremely high or extremely low. Basically, fundamental frequency is indirectly proportional to the period of the human voice signal or directly proportional to sampling frequency [4]. If N is number of overlapped segments which divide complete voice signal and Ti denotes ith-segment period, then

$$F0 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{Ti}$$

B. Jitter

Jitter is another audio characteristic of voice signal. It analyzes as periodic variation in F0 of voice signal. Frequency variation in voice signal is represented by Jitter. Jitter can be calculated as per [5]

$$Jitter = \frac{\left[\frac{1}{N-1}\sum_{i=1}^{N-1}|Ti-Ti+1|\right]}{\frac{1}{N}\sum_{i=1}^{N}Ti}$$

C. Shimmer

Shimmer also represents variation, but focuses mainly on amplitude. When a speaker generates voice then that voice amplitude may vary. It may be high or low. This amplitude variation is the shimmer of the voice signal. If Ak is amplitude of cycle k and M is number of cycles of the voice signal then shimmer defined as

$$Shimmer = \frac{1}{M-1}\sum_{k=1}^{M-1} 20\log_{10}\frac{Ak}{Ak+1}$$

Shimmer, fundamental frequency, pitch, and jitter are voice characteristics and are helpful during voice feature extraction process, which is defined in next section.

Speaker recognition is not easy when high background noise comes with speaker voice. When high noise comes with speaker voice then key human voice feature subsides and creates issue in speaker recognition. Most of previous research work uses MFCC as feature extraction method but it’s still challenging with degraded human voice to extract optimized human voice features. The purpose of this research is to identify optimized feature extraction.

The paper is organized as follows: Section II is more about related work in same research topic; base model for speaker recognition is discussed in Section III which defines all necessary steps for speaker recognition system. In this section we also discuss about what is degraded human voice, and the data set. In Section IV we present baseline model results, proposed model with modified MFCC and results comparison, then we have concluded discussion in Section VI.

II. Related Work

N. V. Tahliramani and N. Bhatt [7] study presents that speaker can be recognized by a machine from the speaker's voice. MFCC is used to extract voice features during training and can also be used for speaker identification. Silence and noise can also be present alongside the speaker's actual voice during speaker recognition. According to the proposed model, noise and silence (between words) are removed before comparing the voice for similarity with the stored voice sample of the same speaker. Framing, windowing, low-pass filtration, and transformation techniques are employed to identify voice features. GMM is used during training and testing for speaker recognition. GMM increases the probability of correctly recognizing the speaker, even when noise is present in the speaker's voice.

S. Park, Y. Park, A. Nasridinov and J. Lee [6] study presents that conference call (closed user group) over any conference application requires gender identification. Gender can be recognized based on frequency of the speaker’s voice and effectiveness of the speaker recognition process. Correct meeting notes can be created based on gender and speaker recognition methods. Female voice frequency (average 188-221Hz) is commonly high as compared to the male voice frequency (average 100-146Hz). Difference in frequency is used to recognize gender and correctly assign the person’s ID. “Text To Speech” API from Google is used to transform voice to text and create runtime Minutes of Meeting (MoM).

N. Gupta and S. Jain [21] presented that speaker recognition is possible through Convolutional Neural Network (CNN) based speaker recognition system. Siamese and CIFAR network architecture are used in speaker recognition. CNN base layers, convolutional, pooling and dense, are used in the model for pattern matching, recognizing deviation and pattern categorization. Negative and positive voice samplings are used for better speaker identification. Positive human voice sample is recorded when there is less distance between speaker and the microphone. When recording is done with some distance then it is considered as negative sample. Presented model for speaker recognition is built on “CIFAR” network architecture with shared weightage on “Siamese Network”. Main purpose of “Siamese Neural Network” (SNN) is to learn the feature vectors. This type of speaker recognition system can be used more in places like bank and telecom.

M. M. Mubarak al Balushi, R. V. Lavanya, S. Koottala and A. V. Singh [23] proposed denoising technique for better speaker recognition. Study proposed that denoising can be performed in transformation domain and it gives different results with different wavelet transformation techniques. Denoising means suppression of the noise from speaker’s voice. Denoising filter can suppress either the low frequency in the human voice when it notices it or increase in the voice where identifying that SNR is low. In addition, wavelet filter can filter the voice in the frequency domain. Proposed work is applicable for human and animal voice. Authors used Matlab programing tool to simulate results. Fejer- Korovkin wavelet filter has been compared with other available wavelet transformation filter to analyze noise elimination for better speaker recognition. Fejer- Korovkin & Dmey wavelets were verified for denoising. SNR and Mean Square Error (MSE) used as a parameter to check voice signal quality. Fejer-Korovkin results are 5% better than Dmey wavelet in terms of SNR.

R. M. Lexușan [16] presented that, which voice features are the best features for human voice. Some of the key features of human voice are MFCC coefficient, energy of the voice signal, fundamental frequency of voice signal, duration and ratio of voice and unvoiced segment. This study also performed experiment which used 172 voice samples with happiness, sadness, and neutral states. Study also using “Support Vector Machine” (SVM) and decision tree as a classification method. Study shown that decision tree gives better recognition rate as compared to liner SVM. Study also concluded that algorithm provides approximately 85% recognition rate. With help of Nao robot, it's capable to recognize speaker emotions from the recording. Similarly, R. Chakroun, L. B. Zouari, M. Frikha and A. Ben Hamida [24] presented that Support Vector Machine (SVM) is more supportive with GMM for speaker recognition. GMM is more successful for speaker recognition when speaker voice is text independent. Study used GMM supervector in SVM, it’s combining SVM results with GMM supervectors.

J. G. Liu, Y. Zhou, H. Q. Liu and L. M. Shi [25] studied multiple methods for noise elimination from speech for single channel speech betterment. It's observed that most of speech enhancement analysis performed in frequency domain. The authors first analyzed Chi priori effects with weighted Bayesian estimator on speech and then incorporated “Speech Presence Uncertainty” (SPU) into the proposed estimator to derive an efficient hybrid priori SNR (HSNR) estimator. These methods give effective result to eliminate musical noise from speech and better speaker recognition process.

Thimmaraja Yadava G et al. [26] presented a pre-processing method for noise elimination which can be used in any speech recognition system specifically for Kannada speech (One of the Indian languages). It’s based on spectral subtraction “voice activity detection” (VAD). As per spectral subtraction, noisy speech data first need to segment first, and that segment need to overlap up to 50% in subsequent frames. The authors analyzed the noisy speech using autocorrelation spectral subtraction and periodogram methods and that is in Linear Prediction Coefficient (LPC). Noise elimination observed when subtracts the periodograms of additive noisy signal from corrupted speech signal.

M. N. A. Aadit et al. [27] have proposed white and colored noise suppression method with adaptive Kalman filter approach. The authors analyzed stationary and dynamic nature noise to retrieve the desired information from noisy Bangle speech. Proposed method is using recursive filter which is based on Kalman filter approach to improve speech signals which is corrupted by both static and dynamic noises. The authors also compare speech signal before noise elimination and after noise elimination to validate performance of proposed filter and it’s based on pitch value of speech, mean square error before and after noise elimination was between -0.4 to 0.2dB where SNR vary from -35 to -20.

Most of the researchers have suggested different type of filters, which are mostly effective when noise before and after human voice or when noise have different frequency which is not in human voice range. It’s easy to eliminate noise when noise has low frequency or high frequency as compared to male and female voice frequencies. Main concern is when human voice and noise comes together, and noise signal strength is high. Our experiment is focused in this area to address this issue.

III. Research Methodology

This section details the speaker recognition baseline model, what is degraded human voice and details of dataset.

A. *Speaker Recognition Base Model*

When speaker voice come with high background noise then it is challenging to identify speaker with any speaker recognition system. Fig. 1 is showing basic process for any Speaker Recognition system.

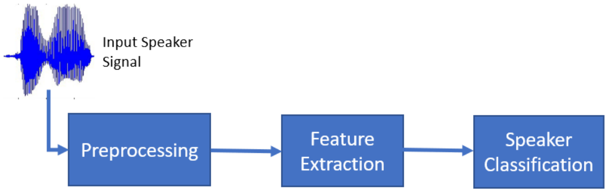

Fig. 1. Speaker recognition process blocks.

1) Preprocessing: Preprocessing step is used to clean speaker data. When speaker voice come for identification then it can have high or low frequency noise. Generally different types of filters used to remove such kind of noises. Band Stop Filter (BSF) and “Low Pass Filter” (LPF) are common to filter noise from speaker audio. Butterworth filter is commonly used as band-stop filter [1].

2) Voice feature extraction: Voice feature or characteristics extraction is utilized to extract feature from speaker’s audio file. Extracted feature is used for speaker classification. There are multiple voice feature extraction methods and MFCC [6][7][8][9] is widely used. Linear Predictive Coding (LPC) [10] is also used for feature extraction. For feature extraction, MFCC uses following key steps [7][11][12] to identify cepstral coefficients. MFCC converts human voice into cepstral coefficient in matrix form and it is easy to use in classification step. Following steps may be used in combination also, based on application and source speaker voice.

- Framing and Blocking
- Windowing
- Fast Fourier Transform (FFT)
- Triangular Bandpass Filter (Mel Scale)
- Invers FFT

In our experiment we have also used MFCC and modify it for better speaker classification results.

3) Classification: MFCC output works as an input for speaker classification. Classification process differentiates one speaker from another speaker through speaker’s voice features. Speaker classification can be performed based on different type of available classifier. Different researcher used GMM [13][14], “Hidden Markov Model” (HMM) [15], “Support Vector Machine” (SVM) [16][17][18][19] and deep learning based “Convolutional Neural Network” (CNN) [10][20][21][22] classifier. In this article we have focused on most common classifier as GMM.

4) Gaussian mixture model: GMM is a probabilistic model and it’s another type of clustering algorithm. GMM creates different types of clusters, and every cluster is modeled as per different Gaussian distribution. In another word GMM generates data points which are derived from a mixture of a limited Gaussian distributions that has no known parameters. There are two approaches which help to drive these paraments. One is maximum a posteriori estimation and another is prior trained “Expectation-Maximization” algorithm. Generally, “Expectation-Maximization” clustering aka EM clustering is mostly used for speaker recognition.

B. *Degraded Human Voice*

Nowadays we can’t expect silence in public places and generally lot of background noises are also come with in speaker voice. Current need of speaker recognition system is to identify speaker from noisy environment. Efficiency of good speaker recognition is dependent on the speaker's audio quality. If speaker's voice comes without any background noise, then it's easy to recognize by machine as compared to high background noise human voice. We consider human voice as degraded human voice when high background noise with human voice. Such degraded human voice is not easy to recognize by machine. There are multiple articles [1][26][27][28] which used frequency-based noise elimination methods. If background noise is continuous and low as compared to speaker's voice, then probability to recognize the speaker is high. But if background noise is impulsive (sudden high) and high as compared to speaker voice then probability of speaker recognition will decrease. This article mainly focuses on noise between words and sentences.

Generally, voice degradation is defined based on human voice and noise power. SNR signifies ratio of voice signal power to the noise power. SNR is indirectly proportional to the noise signal power. SNR is low when high noise signal power. Low values of SNR indicate highly degraded human voice. For example, if SNR is high like 12dB then that voice signal is good and have almost no noise in other if SNR is 1dB then this signal has lot of noise and hard to recognize speaker from this voice signal. If Ps is voice signal power and Pn represent noise signal (background) then SNR is

$$SNR(dB) = 10\log_{10}\left(\frac{Ps}{Pn}\right)$$

OR

$$SNR(dB) = \mathrm{Ps(dB)} - \mathrm{Pn(dB)}$$

C. Data Set

In major speaker recognition applications, customer (speaker) training sample recorded in the silent environment or without noise environment but testing sample comes with lot of background noise.

In our experiment we have also used same mechanism that training sample is recorded in silent environment when no background noise is there but testing sample comes with different background levels. In general, two different datasets are used (human audio and noise) and merged them to create degraded human voice dataset. In our experiment we have not mixed two signals (human voice and noise) even instead we have created dataset as per real life scenario like human speaks in noisy environment. We have recorded human voice in noisy environment. It means in same room we have noise source and human voice. It's not mixed of human voice and noise through any application. We have used 10x10 feet room and recoded human voice with air friction noise. During data creation we have recorded voice through "Microsoft Conexant ISST Audio" with driver version 10.0.18362.1

We have defined five different categories and each category has multiple male speaker voice for training and testing purpose. We have used voice sample from five categories based on SNR range as defined in Table I. SNR calculation is based on voice signal strength (with python noisereduce library) and noise signal strength as principally explained in sub section B of Section III.

TABLE I. DATASET SNR RANGES FOR TRAINING AND TESTING

Voice sample Category	*SNR (dB) Range*
A	10 to 12
B	8 to 10
C	3 to 5
D	2 to 3
E	1 to 2

Category A data has been used for training purpose and remaining four categories voice samples have been used for testing purpose. Speaker voice from Category-A has SNR in range of 10-12dB and power spectrum as define in Fig. 2.

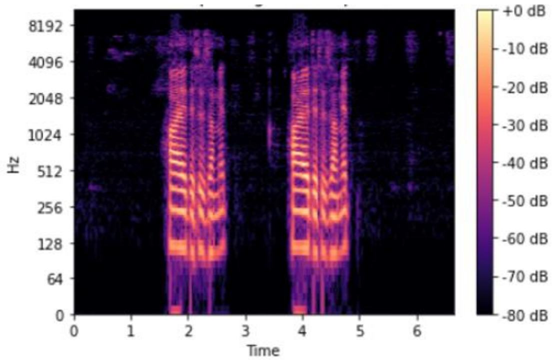

Fig. 2. Power spectrum category-a sample.

Category B data has been used for testing purpose. Speaker voice from Category-B has SNR in range of 8-10dB and power spectrum as defined in Fig. 3. Category-B voice signal has slightly high background noise as compared to Category-A voice signal.

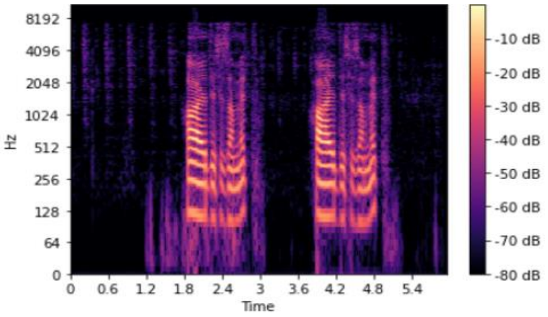

Fig. 3. Power spectrum category-b sample.

Category C data has been used for testing purpose. Speaker voice from Category-C has SNR in range of 3-5dB and power spectrum as defined in Fig. 4. Category-C voice signal has high background noise as compared to Category-A and Category-B voice signal.

Category D data also has been used for testing purpose. Speaker voice from Category-D has SNR in range of 2-3dB and power spectrum as define in Fig. 5. Category-D voice signal has high background noise as compared to Category-A and Category-B voice signal but there is not much different as compared to Category-C voice signal.

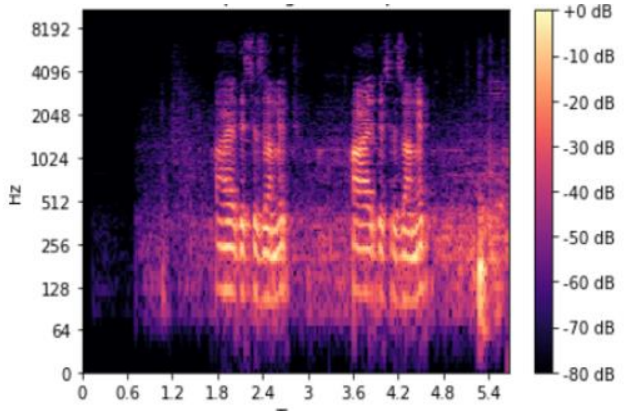

Fig. 4. Power spectrum category-c sample.

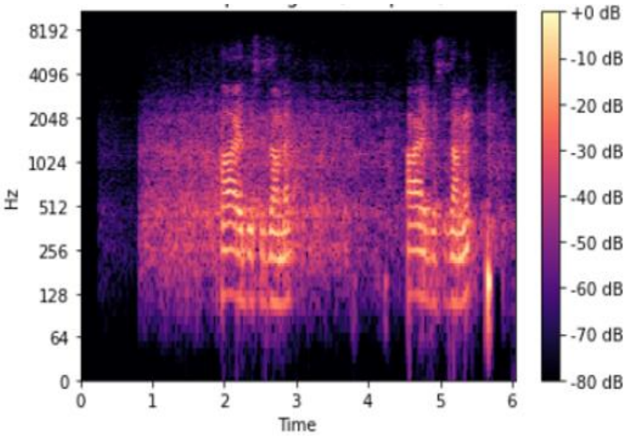

Fig. 5. Power spectrum category-d sample.

Category E data is also used for testing purpose. Speaker voice from Category-E has SNR in range of 1-2dB and power spectrum as defined in Fig. 6. Category-E voice signal has high background noise as compared to Category-A and Category-B voice signal but there is not much different as compared to Category-C and Category-D voice signal.

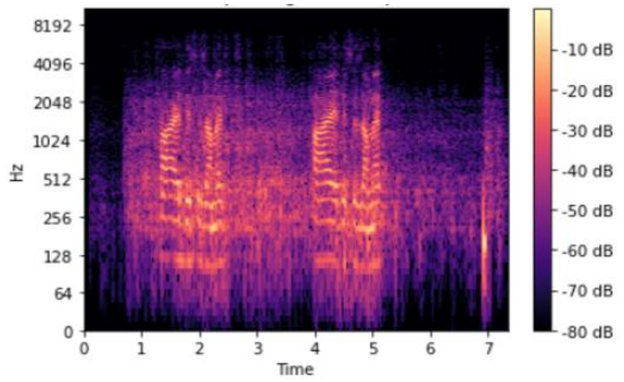

Fig. 6. Power spectrum category-e sample.

When we analyze power spectrum of Category-E sample then as per Fig. 7, yellow highlighted power is noise power before and after human voice and green highlighted is showing noise between words.

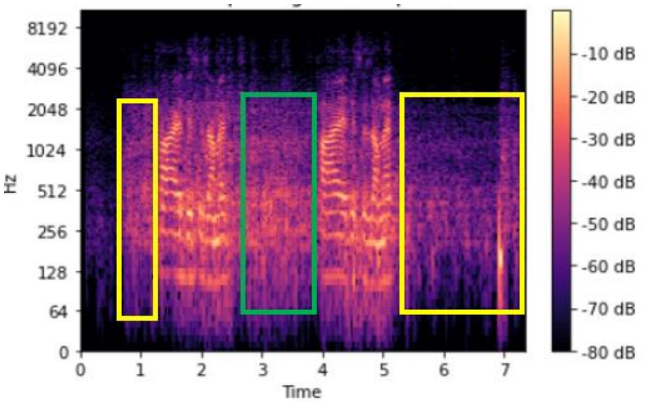

Fig. 7. Power spectrum analysis.

IV. Experiment and Results

A. Baseline Model and Results

As discussed in Section III, MFCC is used as voice feature extraction method and GMM for learning and pattern matching. We have used same device to produce different strengths of air friction noise. In all samples voice recorder remains same.

With baseline model, Samples from category A is used for training and other categories sample are used for testing purpose. As per experiment results, if SNR is reduced by 2dB then baseline model can’t identify speaker voice as per Table II.

TABLE II. Experiment Result with Baseline Model

Voice sample Category	*SNR (dB) Range*	*Identify correctly*
A	10 to 12	Training Sample
B	8 to 10	No
C	3 to 5	No
D	2 to 3	No
E	1 to 2	No

B. Proposed Model and Results

Proposed model is also based on MFCC and GMM. On top of baseline model, we have performed three steps modification at MFCC level. We proposed GMM model with modified MFCC for feature extraction. When feature extraction provides more information at MFCC coefficient level then GMM also adopts this information and improves speaker recognition performance. Similar approach is used with TIMIT dataset and CNN [28]. Environment and data set are also remaining same as in baseline model. Only modification is performed at feature extraction level.

Step-1: High Sampling Rate

Sampling rate is key factor to identify voice feature. High sampling rate captures more signal dissimilarity info as compared to low sampling rate. We have increased sampling rate from 22050 to 44100 and identified test result as per Table III.

TABLE III. EXPERIMENT RESULT WITH PROPOSED MODEL STEP-1

Voice sample Category	*SNR (dB) Range*	*Identify correctly*
A	10 to 12	Training Sample
B	**8 to 10**	**Yes**
C	3 to 5	No
D	2 to 3	No
E	1 to 2	No

Table III shows slight improvement in test result for classification of the speaker when SNR in range of 8-10dB.

Step-2: Frequency Range for mel-scale triangular filter bank

As per Section III, MFCC uses Mel-scale triangular bandpass filter. These filters use low and high frequency range to create triangular filter bank. Human male fundamental frequency comes in range of 0-900Hz. If we create mel filter bank in just twice from male fundamental frequency ~1800Hz then it mostly captures human voice relevant information for speaker recognition process and avoid noise information which occurs in between the words. Optimization of frequency range for triangular bandpass filter gives benefit to filter out background noise.

Table IV shows results when we create mel-scale filter bank from 0 to 1800 Hz (high frequency as 1800 Hz) during MFCC coefficient calculation.

TABLE IV. EXPERIMENT RESULT WITH PROPOSED MODEL STEP-2

Voice sample Category	*SNR (dB) Range*	*Identify correctly*
A	10 to 12	Training Sample
B	**8 to 10**	**Yes**
C	**3 to 5**	**Yes**
D	2 to 3	No
E	1 to 2	No

Table IV shows good improvement in test result for classification of the speaker when SNR in range of 8-10dB, 3-5dB and 2-3dB sample category. But still Category-E sample is not classified correctly because of very low SNR.

Step-3: Pre-emphasis Effects

Humans generate sound with fundamental frequency at 0-900Hz for male and up to 1500Hz for female. Vocal tract modulates this voice and generates modulated voice. This modulated voice is suppressed by high frequency voice. The objective of pre-emphasis is to compensate on the high-frequency part that was suppressed during the sound generation by the humans [29]. When we look voice power spectrum of data set in Fig. 4 to 7 then we can realize that spectrum power is more on lower frequencies and it’s reducing at high frequency. So, it is required to boost the energy levels at the high frequencies. Pre-emphasis is one of the key steps in feature extraction process and is considered during MFCC coefficient calculation. Fig. 8 is showing low power level when frequency is more than 1500Hz (approximately).

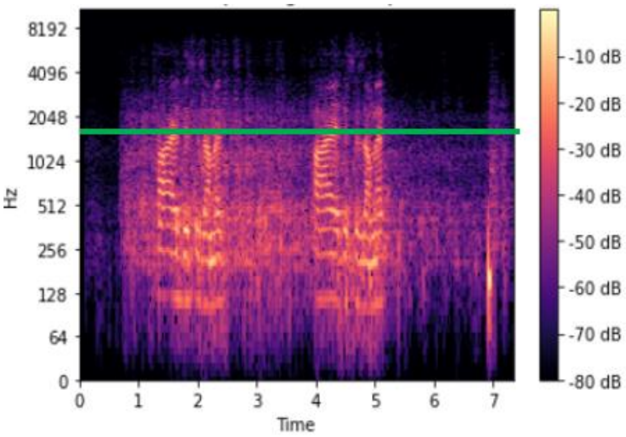

Fig. 8. Power spectrum freq > 1500Hz.

In general, when MFCC coefficient is calculated by different libraries like “*python_speech_features*” then standard value is 0.97 and when we used same pre-emphasis values then we have not received much gain in speaker recognition accuracy. In our experiment we optimized pre-emphasis value and make it at 1.00 then we have received results as per Table V. It’s also observed that if we tune pre-emphasis values from .90 till 1.10 then result values getting changed and is not much effective like at pre-emphasis = 1.00.

TABLE V. EXPERIMENT RESULT WITH PROPOSED MODEL STEP-3

Voice sample Category	*SNR (dB) Range*	*Identify correctly*
A	10 to 12	Training Sample
B	**8 to 10**	**Yes**
C	**3 to 5**	**Yes**
D	**2 to 3**	**Yes**
E	**1 to 2**	**Yes**

V. CONCLUSION

Different speaker recognition systems are required for various applications. Most applications seek a speaker recognition system that functions well without any background noise during training, but same system should recognize speaker even with degraded human voice. In our experiment, we addressed this issue and utilized speaker voices with minimal background noise (SNR = ~11dB) during training, and tested the system's performance with degraded human voices at SNRs as low as 1dB. According to Section IV of our experiment, we observed that high sampling rate, optimized frequency range for the triangular mel bandpass filter, and optimized pre-emphasis value, all contribute to the effectiveness of the feature extraction mechanism for calculating MFCCs in the speaker recognition process. In future, this experiment could be expanded to include different datasets comprising voices of various genders and languages. Modified MFCC can also tested with K-Nearest Neighbor (KNN) and Random Forest [30].

REFERENCES

[1] W. Meiniar, F. A. Afrida, A. Irmasari, A. Mukti and D. Astharini, "Human voice filtering with band-stop filter design in MATLAB," 2017

International Conference on Broadband Communication, Wireless Sensors and Powering (BCWSP), Jakarta, 2017, pp. 1-4, doi: 10.1109/BCWSP.2017.8272563.

[2] J. Wang and M. T. Johnson, "Physiologically-motivated feature extraction for speaker identification," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 1690-1694,

[3] S. Ostrogonac, M. Sečujski, D. Knezevic and S. Suzić, "Extraction of glottal features for speaker recognition," 2013 IEEE 9th International Conference on Computational Cybernetics (ICCC), 2013, pp. 369-373

[4] M. Sigmund, "Illustrative Method of Determining Voice Fundamental Frequency Using Mathcad," *2021 30th Annual Conference of the European Association for Education in Electrical and Information Engineering (EAEEIE)*, Prague, Czech Republic, 2021, pp. 1-4.

[5] S. S. Upadhya, A. N. Cheeran and J. H. Nirmal, "Statistical comparison of Jitter and Shimmer voice features for healthy and Parkinson affected persons," *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Coimbatore, India, 2017.

[6] S. Park, Y. Park, A. Nasridinov and J. Lee, "A Person Identification Method in CUG Using Voice Pitch Analysis," 2014 IEEE Fourth International Conference on Big Data and Cloud Computing, Sydney, NSW, 2014, pp. 765-766

[7] N. V. Tahliramani and N. Bhatt, "Performance Analysis of Speaker Identification System With and Without Spoofing Attack of Voice Conversion," 2018 2nd International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE), Ghaziabad, India, 2018, pp. 130-135..

[8] F. Eyben et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," in IEEE Transactions on Affective Computing, vol. 7, no. 2, pp. 190-202, 1 April-June 2016

[9] M. Sadeghi and H. Marvi, "Optimal MFCC features extraction by differential evolution algorithm for speaker recognition," 2017 3rd Iranian Conference on Intelligent Systems and Signal Processing (ICSPIS), 2017, pp. 169-173

[10] A. Chowdhury and A. Ross, "Fusing MFCC and LPC Features Using 1D Triplet CNN for Speaker Recognition in Severely Degraded Audio Signals," in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 1616-1629, 2020

[11] Gupta, Shikha & Jaafar, Jafreezal & Wan Ahmad, Wan Fatimah & Bansal, Arpit. (2013). Feature Extraction Using Mfcc. Signal & Image Processing : An International Journal. 4. 101-108. 10.5121/sipij.2013.4408.

[12] A. Winursito, R. Hidayat and A. Bejo, "Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition," 2018 International Conference on Information and Communications Technology (ICOIACT), 2018, pp. 379-383

[13] O. Büyük and L. M. Arslan, "Age identification from voice using feed-forward deep neural networks," 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1-4.

[14] H. Bounazou, N. Asbai and S. Zitouni, "GMM Evaluation for Speaker Identification," *2022 International Conference of Advanced Technology in Electronic and Electrical Engineering (ICATEEE)*, M'sila, Algeria, 2022, pp. 1-5

[15] Y. Wei, "Adaptive Speaker Recognition Based on Hidden Markov Model Parameter Optimization," in IEEE Access, vol. 8, pp. 34942-34948, 2020

[16] R. M. Lexușan, "Comparative study regarding characteristic features of the human voice," 2015 7th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Bucharest, 2015, pp. WSD-1-WSD-4, doi: 10.1109/ECAI.2015.7301206.

[17] B. K. Baniya, J. Lee and Z. Li (2014), " Audio feature reduction and analysis for automatic music genre classification", In IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 457-462.

[18] S. Cumani and P. Laface, "Large-Scale Training of Pairwise Support Vector Machines for Speaker Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 11, pp. 1590-1600, Nov. 2014

[19] R. Mardhotillah, B. Dirgantoro and C. Setianingsih, "Speaker Recognition for Digital Forensic Audio Analysis using Support Vector Machine," 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2020

[20] A. H. Meftah, H. Mathkour, S. Kerrache and Y. A. Alotaibi, "Speaker Identification in Different Emotional States in Arabic and English," in IEEE Access, vol. 8, pp. 60070-60083, 2020

[21] N. Gupta and S. Jain, "Speaker Identification Based Proxy Attendance Detection System," 2019 International Conference on Signal Processing and Communication (ICSC), NOIDA, India, 2019, pp. 175-179.

[22] D. Snyder, D. Garcia-Romero, D. Povey and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification", Proc. of Interspeech, pp. 999-1003, 2017

[23] M. M. Mubarak al Balushi, R. V. Lavanya, S. Koottala and A. V. Singh, "Wavelet based human voice identification system," 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), Dubai, 2017, pp. 188-192

[24] R. Chakroun, L. B. Zouari, M. Frikha and A. Ben Hamida, "A hybrid system based on GMM-SVM for speaker identification," *2015 15th International Conference on Intelligent Systems Design and Applications (ISDA)*, Marrakech, 2015, pp. 654-658

[25] J. G. Liu, Y. Zhou, H. Q. Liu and L. M. Shi, "An improved generalized weighted Bayesian estimator for speech enhancement," 2016 IEEE International Conference on Digital Signal Processing (DSP), Beijing, 2016, pp. 249-252

[26] Thimmaraja Yadava G, Jai Prakash T S and Jayanna H S, "Noise elimination in degraded Kannada speech signal for Speech Recognition," 2015 International Conference on Trends in Automation, Communications and Computing Technology (I-TACT-15), Bangalore, 2015, pp. 1-6

[27] M. N. A. Aadit, S. G. Kirtania and M. T. Mahin, "Suppression of white and colored noise in Bangla speech using Kalman filter," 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Dhaka, 2016, pp. 1-6\

[28] Amit Moondra and Poonam Chahal, “Improved Speaker Recognition for Degraded Human Voice using Modified-MFCC and LPC with CNN” International Journal of Advanced Computer Science and Applications(IJACSA), 14(4), 2023.

[29] Himani Chauhan et al, “Voice Recognition” in International Journal of Computer Science and Mobile Computing, Vol.4 Issue.4, April- 2015, pg. 296-301

[30] Vincentius Satria Wicaksana and Amalia Zahra S.Kom, “Spoken Language Identification on Local Language using MFCC, Random Forest, KNN, and GMM” International Journal of Advanced Computer Science and Applications(IJACSA), 12(5), 2021

AUTHORS’ PROFILE

Amit Moondra received the Master of Engineering degree in Communication Engineering from the Birla Institute of Technology and Science, Pilani, India (BITS - Pilani). He has more than 20 years of industrial and research experience with 10+ across countries. He is currently working in Ericsson Global India Limited as Senior System Manger in Product Development Unit and pursuing his Ph.D. at Manav Rachna International Institute of Research and Studies. India. His research focuses on artificial intelligence, deep learning model in speech area. He is an active member of IEEE.

Poonam Chahal received her Ph.D. in 2017 from YMCA University of Science and Technology, Faridabad India, in the field of Artificial Intelligence. Presently she is working as Professor in Department of Computer Science and Engineering at FET, Manav Rachna International Institute of Research and Studies, Faridabad. She is actively involved in research activities and is on the reviewing panel of many journals and conferences.