Improved Speaker Recognition for Degraded Human Voice using Modified-MFCC and LPC with CNN

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

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

***Abstract*—Economical speaker recognition solution from degraded human voice signal is still a challenge. This article is covering results of an experiment which targets to improve feature extraction method for effective speaker identification from degraded human audio signal with the help of data science. Every speaker's audio has identical characteristics. Human ears can easily identify these different audio characteristics and classify speaker from speaker's audio. Mel-Frequency Cepstral Coefficient (MFCC) supports to get same intelligence in machine also. MFCC is extensively used for human voice feature extraction. In our experiment we have effectively used MFCC and Linear Predictive Coding (LPC) for better speaker recognition accuracy. MFCC first outlines frames and then finds cepstral coefficient for each frame. MFCC use human audio signal and convert it in numerical value of audio features, which is used to recognize speaker efficiently by Artificial Intelligence (AI) based speaker recognition system. This article covers how effectively audio features can be extracted from degraded human voice signal. In our experiment we have observed improved Equal Error Rate (EER) and True Match Rate (TMR) due to high sampling rate and low frequency range for mel-scale triangular filter. This article also covers pre-emphasis effects on speaker recognition when high background noise comes with audio signal.**

***Keywords*—Data science; artificial intelligence; MFCC; LPC; CNN; mel-spectrum; speaker recognition**

I. Introduction

In the current world, voice communication is used to exchange thoughts and feelings. Most of the business are on voice communication and build trust over voice phone calls. Humans can easily identify a person from his/her voice over a phone call but in modern world voice answer machine should also need to identify person by his/her voice. Now-a-days lot of medical issues are also diagnosed through human voice. Human generates audio from throat and mouth. Fundamental frequency is original frequency, and it is modulated by vocal code structure and that makes every human voice as unique voice. Human can be identified based on his or her voice.

Human voice is generally identified in two categories, male voice, and female voice. The main difference between male and female voice is pitch and frequency. So, if it is just needed to identify gender from voice, then it's easy to classify by pitch difference. But, to recognize a human from his/her voice then all voice features are to be considered. Every person's voice is unique and it's based on different voice features like frequency, jitter, amplitude, pitch and spectral power. Male speaker has 0-900 hz as fundamental frequency and female speaker has 0-1500Hz as fundamental frequency. If we consider average frequency, then male average fundamental frequency is 110 hz and female average fundamental frequency is 211 Hz [1].

Vocal cord, teeth, jaw, and tongue are main articulators of vocal trach which modulate fundamental frequency (F0) [2][3]. To begin with the generation of sound, air pressure is produced by the human lungs. This air pressure generates sound with fundamental frequency. Fundamental frequency sound is modulated by the vocal track to create different sound variations.

In Speaker recognition process, first step to identify human voice feature. MFCC is generally use for voice feature extraction from human voice. MFCC mainly consider frequency changes in human voice to define cepstral coefficients. There are various research earlier which have used MFCC for voice feature extraction. N. V. Tahliramani and N. Bhatt [4] discussed that machine can percept human feature as per MFCC in training phase. Similarly Convolutional Neural Network (CNN) is very popular for speaker recognition process [5]. N. Gupta and S. Jain [6] discussed about Siamese network effects with CNN. A. Chowdhury and A. Ross [7] discussed about CNN with degraded human voice. CNN is one of the preferred model with MFCC and Linear Predictive Coding (LPC) when signal is degraded with high noise [7].

MFCC and LPC are used in most of speaker recognition system. In our experiment, we have modified MFCC and evaluate results. In this article our focus to explain frequency impact on feature extraction and overall improvement in speaker recognition system.

The paper is organized as follows: base model for speaker recognition as discussed in Section II which defines all necessary steps for speaker recognition system, Section III discussed about what are different types of noises and how these noises degrade human voice. Then we discussed about how we prepared dataset with different noises in Section IV, and then in Section V we have defined CNN model including detailed description of modified MFCC, LPC and loss function (cosine triplet), then we have discussed experiment results in Section VI.

II. Speaker Recognition base Model

Speaker Recognition (SR) process has three basic steps to identify speaker. These steps are stich together as per below Fig. 1.

1) Preprocessing
2) Voice Feature Extraction
3) Classification

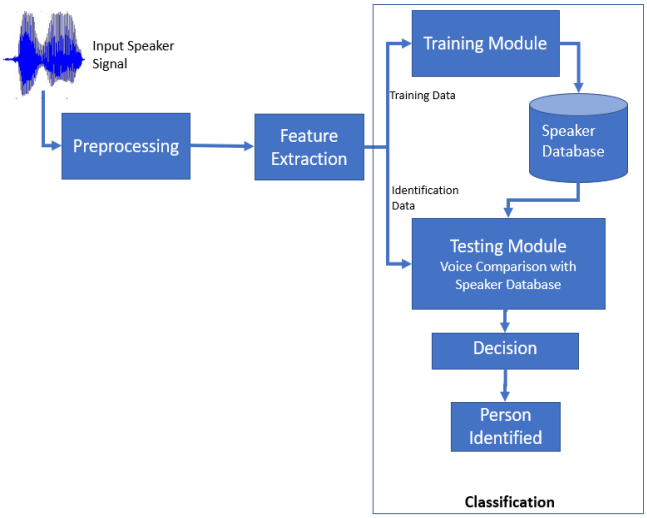

Fig. 1. Speaker recognition process blocks.

Preprocessing is mainly used for cleaning up data. In speaker recognition process preprocessing to remove possible noises from speaker recoded voice. Different types of "Low Pass Filter" (LPF), "Band Stop Filter" (BSF) and "High Pass Filter" (HPF) are used to remove noise from voice sample. Butterworth filter is also used as signal processing filter which is also known as band-stop filter.

Voice feature extraction is used to extract feature from speaker's audio file. These features are used to classify speaker in next step. Voice features can be extracted by one of the most popular techniques Mel-Frequency Cepstral Coefficient (MFCC) [4][7][8][9][10], Linear Predictive Coding (LPC) [7] is also used for feature extraction. MFCC has its own benefit for feature extraction. For feature extraction, MFCC uses following main steps [4][11][12] to identify cepstral coefficients. MFCC converts voice recording into cepstral coefficient in matrix form which is easy to feed in next classification step. Following steps may be used in combination with LPC or other feature extraction methods also, based on application and source speaker voice.

- Framing & Blocking
- Windowing
- Fast Fourier Transform (FFT)
- Triangular Bandpass Filter (Mel-scale)
- Inverse FFT

These steps are stich together as per below Fig. 2.

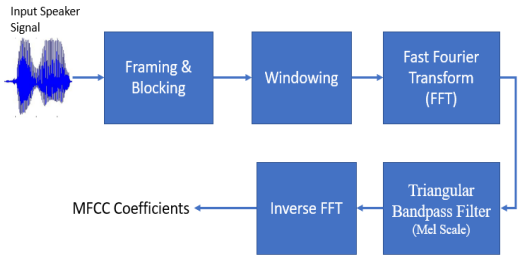

Fig. 2. MFCC processing.

A. Framing

Framing is required to make consistent voice sample. Long voice signal has lot of amplitude and frequency variation which gives lot of inconsistency during voice signal analysis for feature extraction. Short frames solve this issue and generally we create 20 to 40ms short frames for better voice signal analysis [13]. If T is duration of complete voice sample and t is sub frame sample duration, then total number of frames are F.

$$F = \frac{T}{t} * 1000$$

B. Windowing

Windowing is used to increase steadiness between first and end point within the same frame. From framing we get short duration frames, but these frames are generally discontinued because for interworld silence. Humming window makes continuity within same frame. If xi(n) is time domain signal and hi(n) is hamming window, then signal will become with hamming window is

$$xi(n)\ after\ windowing = xi(n) * hi(n)$$

Where hi(n)

$$hi(n,\alpha) = (1-\alpha) - \alpha\ cos\left(\frac{2\pi n}{N} - 1\right)$$

$$1 \leq n \leq N-1$$

C. Fast Fourier Transform

When speaker's voice is recorded by any microphone, that signal is recorded in time domain. Time domain gives limited information about human voice feature. "Fast Fourier Transform" (FFT) converts time domain signal into frequency domain signal. Computation effect of FFT and "Discrete Fourier Transform" (DFT) are equal but Fast Fourier Transform is actually a fast algorithm to conceptualize DFT. DFT is discrete version of FFT. If N represent number of samples in frame, then each sample of that frame needs to be converted in frequency domain. If xi(n) is voice time domain framed signal and Xi(k) is transformed signal as define in below in equation.

$$Xi(k) = xi(n)\ hi(n)\ e^{-\frac{j2\pi kn}{N}}$$

$$1 \leq n \leq N\ \ and\ 1 \leq k \leq K$$

Where hamming window is h(n), K is length of DFT.

D. Triangular Bandpass Filter: Mel-Scale

Humans perceive different human voice differently and similar intelligence is required in machines also. When human

time domain voice signal converted to frequency domain signal the triangular band-pass filters are used to filter human voice signal. FFT transform time-domain signal into frequency-domain signal and frequency domain signal is passed through from triangular bandpass filter.

At lower frequency range group, humans can understand two separate frequency voice, but at high frequency range group, two separate frequency voice is not recognized by humans. For example, low frequency group 'A' has F1 and F2 frequency voice signal. F1 and F2 has d difference then human can recognize F1 and F2 correctly. Similarly, high frequency group 'B' has F3 and F4 frequency voice signal. F3 and F4 also has d frequency difference then humans face difficulties to recognize it. To add similar intelligence in speaker recognition system, mel-scale is used to convert frequency into mel-scale as per given formula. If mf represents frequency in mel-scale and f in hz then

$$mf = 2595 * log_{10} \frac{f}{700} + 1$$

The purpose of mel-scale conversion is to change the signals to follow the decrease properties of the mel-scale. At lower frequency it is linear conversion but at high frequency it is more static as per below Fig. 3.

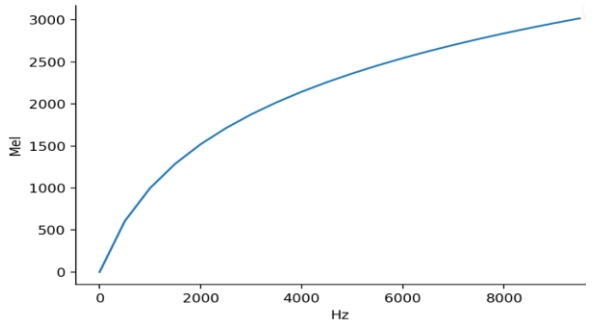

Fig. 3. Mel-scale.

Mel-scale is used to identify the space of the triangular filter bank and also to identify how much filer bank width should be maintained. At the high frequency filter band width should also get wider to maintain frequency variability. Mel-scale based on triangular filter bank makes easy calculation of energies. Once energy is calculated with the help of mel-scale based triangular filter then it's easy to calculate mel-spectrum. Mel-spectrum is used to calculate different cepstral coefficients (generally 20) with the help of Discrete Fourier Transform (DCT).

E. Inverse FFT

Inverse FFT is used to convert back mel-spectrum into to the spatial domain. Initially time domain signal is converted to frequency domain signal and then frequency domain signal is transformed in mel-scale. Then it's necessary to convert back mel-scale signal. When inverse FFT is applied then signal gets converted back to spatial domain. This transformation can be performed either by "Discrete Fourier Transform" (DFT) or Discrete Cousins Transform (DCT). Both DFT and DCT can be used to identify coefficients as DFT and DCT divides a given sequence of finite length data into discrete vector. In general, DCT is more used to identify coefficient from the specified log mel-spectrum [9]. DCT provides coefficients as output and these coefficients are known as MFCC. If m coefficient is required to calculate (generally 20 or 13 coefficients calculate) and Cn represent MFCC coefficient [9].

$$Cn = \sum_{i=1}^{i} log_{10} Xi \, Cos[m(i - \tfrac{1}{2})\, \pi / i]$$

Where Xi is FFT of the voice signal and m = 0, 1, …i-1

When an audio signal is converted in matrix data set with the help of MFCC output then this MFCC output works as a input for classifier model. There are multiple models used by different researchers for speaker recognition like "Gaussian Mixture Model" (GMM) [14][15][16][17], "Hidden Markov Model" (HMM) [18], "Support Vector Machine" (SVM) [19][20][21][22]. and deep learning based "Convolutional Neural Network" (CNN) [5][6][7][23]. In this article we have used CNN as classification model for speaker recognition.

III. Noise Effects on Speaker Recognition

Speaker Recognition efficiency is majorly dependent on the speaker's audio quality. If no background noise is included with speaker's voice, then speaker recognition results are quite good as compared to high background noise. When background noise is high then, it degrades the overall human voice signal. Degraded human voice signal is not only difficult to recognize but it also leads to a complicated speaker recognition system. There are multiple methods to eliminate noise from human voice. Some of the articles [1][24][25][26] have used frequency-based filtration logic to eliminate noise.

In speaker recognition process, speaker's voice should be noise free for better recognition. In today's world absolute silence is not possible in public places and lot of background noises are also captured in speaker voice. There are multiple use cases which use speaker recognition as authentication mechanism to identify customer. Background noise coming before and after speaker's voice can be easily eliminated in speaker recognition process but to eliminate background noise coming between two words is a major task in speaker recognition [27]. Type of noise is also important to consider in speaker recognition process.

A. Different Types of Noise

There are different types of noise classifications like colour based noise, white noise, violet noise etc. Noise can also be classified as

1) Machine based noise: This type of noise is generated by a machine. When a machine runs then that machine has its own noise like car, bus, large scale process industries machine, train, airplane have their own noises [28]. Machines also contribute to noise generation indirectly, like generation of noise from a home fan due to air friction.

2) Human generate noise: When many humans speak at same time with different words then their combined voice resembles as noise. For example, noise in restaurant, in stadium, shopping mall or in street [28].

3) Nature inspired sound: When sound is not generated by machine or human and is generated by natural resources then

the noise is considered as nature inspired sound like animal or bird sound, waterfall sound, fire sound, and wind sound. In speaker recognition process, whichever sound comes from the background making disturbance to classify speaker is considered as background noise. When speaker voice comes with waterfall sound or dog bark as a background then it disturbs with human voice signal and makes the speaker recognition process complicated.

B. Signal to Noise Ratio (SNR)

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. If Ps is voice signal power and Pn represent noise signal (background) then SNR is

$$SNR \propto \frac{1}{Pn}$$

$$SNR = \frac{Ps}{Pn}$$

Human voice signal comes from throat and modulated with vocal cord, jaw, teeth and tongue. Human voice is always in changing pattern and it's never constant. As human voice is never constant and dynamic in nature so it's not idle for voice feature analysis. Hence, to regularize this it's required to express voice signal power in logarithmic scale as:

$$Ps(dB) = 10log_{10}\, Ps$$

Similarly, noise signal power can also be described in logarithmic scale as

$$Pn(dB) = 10log_{10}\, Pn$$

SNR can also be expressed in logarithmic scale. In this article we are considering SNR in decibel(dB).

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

OR

$$SNR(dB) = Ps(dB) - Pn(dB)$$

Pn represent noise power (background noise power in SR process) and Ps represent voice signal power without noise. If two separate signals come separately then it's easy to calculate SNR as per above equation. In speaker voice if human voice signal and background noise signal come together then it is hard to separate both signal power. Below equation can be used when both signals come together. When *Px* represent human voice signal power with noise [29] and *Pn* represent noise signal power then SNR is

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

SNR can be positive in value or negative. It's based on voice signal power and noise signal power. When human voice signal is recorded in noisy environment where there is high background noise then there is possibility to have negative SNR. When human voice signal power is greater than noise signal power then SNR will be positive. When noise signal power is greater than human voice signal power then SNR will be negative [29]. For example, when human voice signal is recorded with continuous high background noise like in mechanical factory or in heavy traffic area then noise power is more than human voice signal power. In this case SNR will be negative. But when background noise is impulsive and only one or two spick comes with human voice signal then SNR will be positive but low in value.

Positive SNR (in dB) when

$$Ps > Pn$$

Negative SNR (in dB) when

$$Pn > Ps$$

IV. Dataset

In our experiment we have created following six different types of datasets with the help of TIMIT dataset and NOISEX-92 noise dataset.

A. TIMIT Dataset

TIMIT dataset provided 630 speakers' voice without any added noise. It is a mixed dataset which contains male and female human voice for ~3sec each. Out of 630 speakers, 462 speakers' voice is in training dataset and 168 speakers' voice are in testing dataset.

B. NOISEX-92 Dataset

This database contains eight different types of noises and we have considered following four noises in our experiment.

- Babble Noise: This noise is generated by many people. For example, many human continuous voices in restaurants or in public place.
- F16 Noise: This noise is generated by F16 fighter aircraft engine.
- Factory Noise: This noise is generated by heavy machines in some process/mechanical factory.
- Car Noise: This noise is generated by car engine.

C. Datasets used in Experiment

a) Data Set-1 (DS1): In this data set we used 630 speakers' voice from TIMIT dataset and babble noise which is generated by many humans in a close room like restaurant or common public place and noise generated by F16 fighter aircraft engine from NOISEX-92 dataset. Below Fig. 4 shows power spectrogram of speaker voice with babble and F16 noise.

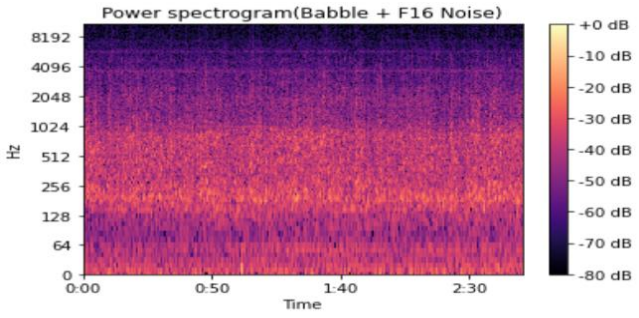

Fig. 4. Power spectrogram (babble + F16 noise).

b) Data Set-2 (DS2): In this data set we used 630 speakers' voice from TIMIT dataset, car engine noise and heavy machinery noise which is generated inside the factory from NOISEX-92 dataset. Fig. 5 shows power spectrogram of speaker voice with car and factory noise.

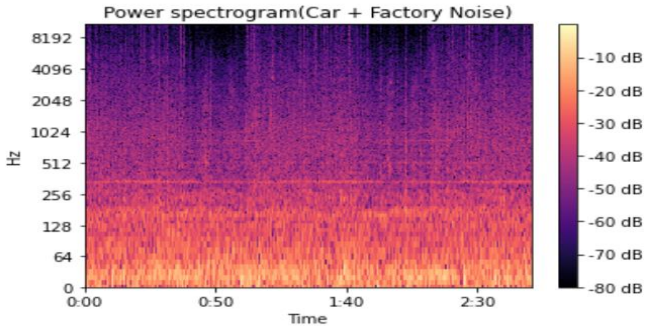

Fig. 5. Power spectrogram (car + factory noise).

c) Data Set-3 (DS3): In this data set we used 630 speakers' voice from TIMIT dataset, babble and car engine noise from NOISEX-92 dataset. Fig. 6 shows power spectrogram of speaker voice with babble and car noise.

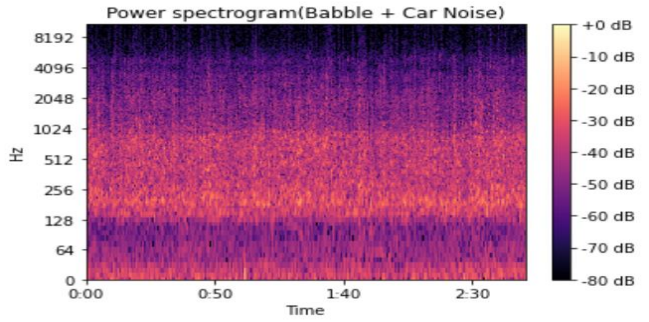

Fig. 6. Power spectrogram (babble + car noise).

d) Data Set-4 (DS4): In this data set we used 630 speakers' voice from TIMIT dataset, F16 & factory noise from NOISEX-92 dataset. Fig. 7 shows power spectrogram of speaker voice with F16 and factory noise.

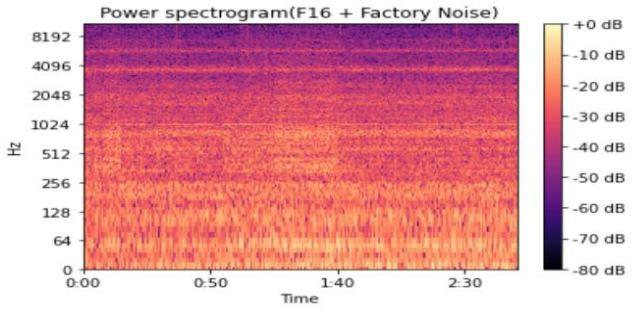

Fig. 7. Power spectrogram (F16 + factory noise).

e) Data Set-5 (DS5): In this data set we used 630 speakers' voice from TIMIT dataset, car and F16 noise from NOISEX-92 dataset. Fig. 8 shows power spectrogram of speaker voice with car and F16 noise.

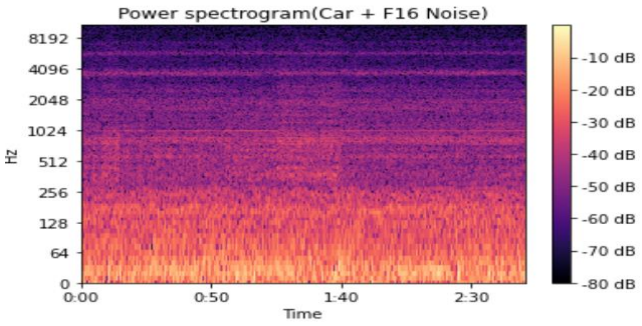

Fig. 8. Power spectrogram (Car + F16 noise).

f) Data Set-6 (DS6): In this data set we used 630 speakers' voice from TIMIT dataset with babble and factory noise from NOISEX-92 dataset. Fig. 9 shows power spectrogram of speaker voice with babble and factory noise.

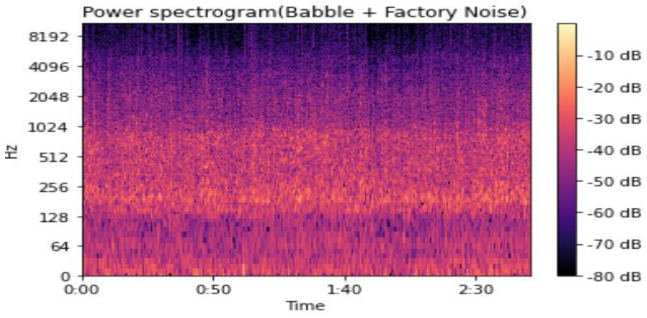

Fig. 9. Power spectrogram (Babble + factory noise).

In all datasets we have set room size as 3m and added reverberation. Below Table I represents summary of all 6-datasets.

TABLE I. DIFFERENT SIX DATASET SUMMARY

Dataset	*Base Voise*	*Added Noise 1*	*Added Noise 2*
DS1	TIMIT 630 Speakers	Babble	F16
DS2	TIMIT 630 Speakers	Car	Factory
DS3	TIMIT 630 Speakers	Babble	Car
DS4	TIMIT 630 Speakers	F16	Factory
DS5	TIMIT 630 Speakers	Car	F16
DS6	TIMIT 630 Speakers	Babble	Factory

V. PROPOSED CNN MODEL

Any speaker recognition system has two base modules, one is feature extraction module and another is classification module based on extracted feature.

A. Training Phase

In training phase we have used CNN for classification of speaker based on 1D-Triplet-CNN model defined in [7]. Convolutional layers play a key role in a CNN to determine its competency and learning capability. Each convolutional layer deeply learns different "concepts" from the data and hands it over to next convolution layer for further deeper learning in the CNN [30][31]. Speech can be represented in form of two dimensions (MFCC & LPC) but these two dimensions do not

show similarity. Generally, speech signals changes constantly but in short interval of voice it is constant in nature. This short interval of audio frame is called feature frame and it is good for speaker recognition process as it holds independent voice property. These multiple feature frames are useful in CNN for correct speaker recognition. We proposed CNN model with Modified MFCC (M-MFCC) and LPC for feature extraction. Three sets of CNN layers used in training phase as defined in Fig. 10.

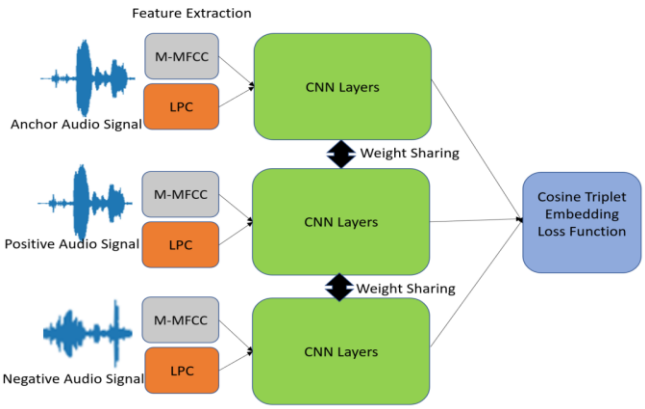

Fig. 10. CNN model training phase.

In training phase, we have used three CNN layers for positive, anchor and negative voice sample. All three CNN layers share weight to learn effectively. Triplet CNN layers are used to learn embedding function f(x).

$$f(x) \in \aleph d$$

f(x) represents data sample embedding in d-dimension Euclidean space and x represents data sample. Following are the steps of overall speaker recognition classification model. We have used modified MFCC(M-MFCC) and LPC for feature extraction and then we have combine M-MFCC and LPC output and feed to CNN layers. CNN layers detailed we have defined in sub section D.

B. M-MFCC

We have modified MFCC in three basic steps to make high similarity in anchor and positive sample and dissimilarity in anchor and negative sample.

Step-1: High Sampling Rate

Sampling rate is the 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.

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

As per Section II, MFCC uses mel-scale triangular bandpass filter. These filters use low and high frequency range to create triangular filter bank. Male speaker has 0-900 Hz as fundamental frequency and female speaker has 0-1500Hz as fundamental frequency. If we consider average frequency, then male average fundamental frequency is 110 Hz and female average fundamental frequency is 211 Hz [1][32]. We observed that most of the male and female voice power comes up to 2000Hz or less. In this work we have selected the filter bank from 0Hz to 1800/2000Hz then it mostly captures human voice relevant information for speaker recognition process and avoids noise information which in between the words. Optimization of frequency range for triangular bandpass filter gives benefit to filter out background noise and improve mel-scale for small variation.

Step-3: Pre-emphasis Effects

Humans generate sound with fundamental frequency. Vocal track modulates this voice and generates modulated voice. This modulated voice is suppressed by high frequency voice. The objective of pre-emphasis is to compensate high-frequency part that was suppressed during the sound generation by the humans [33][34]. 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 did not gain much in speaker recognition accuracy. In our experiment, when we optimize pre-emphasis value and make it at 1.00 then we received improvement in results.

C. Linear Predictive Coding

As discussed in [7], when we need to simulate human voice then human throat functioning needs to be understood. Human voice is generated from lungs and filtered at vocal track. Vocal track is nothing but a time varying digital bandpass filter which filters some frequencies. Also, for detail estimation and analysis of human vocal track we generally look all-pole model of filter design. If we consider H(z) as transfer function of vocal track then,

$$H(z) = \frac{G}{1-\sum_{k=1}^{p} \alpha_k z^{-k}}$$

Human voice signal is continuous signal so if S[n] is voice acoustics of nth sample and p is past voice sample then nth speech sample can be defined as

$$S[n] = \sum_{k=1}^{p} \alpha_k S[n-k] + G.u[n]$$

G is gain factor here and S[n-k], k=1, 2,…p are past human voice samples. Excitation of nth voice sample is represented as u[n]. α_k represents vocal track filter coefficients. As LPC also aims to identify all zero filters which come from invers vocal track model. If S'[n] represents nth speech sample with conditioned on previous speech sample, then

$$S'[n] = \sum_{k=1}^{p} \alpha'_k S[n-k]$$

For correct predication of human vocal track filter coefficients, difference between S[n] and S'[n] should be minimum. α_k represents LPC model attribute which is filter coefficient of the Inverse-Vocal Tract filter model.

D. Combine M-MFCC and LPC

MFCC and LPC bring separate feature characteristics and we combine this information to identify similarity between two voice samples [22]. MFCC brings perceptual speech feature where LPC features give information about speaker vocal track for accurate prediction.

In our experiment we combined M-MFCC and LPC coefficients which give unique voice characteristics of a speaker. For every voice sample M-MFCC is one channel and LPC is another channel.

- M-MFCC Channel: Every frame has 40 features. 20 features from MFCC and another 20 features from delta MFCC.
- LPC Channel: Every frame has 40 features. 20 features from LPC and another 20 features from delta LPC

E. CNN Layers

In this experiment we have used 6 different layers as per Fig. 11. Convolutional layer gives learning capacity from different voice features.

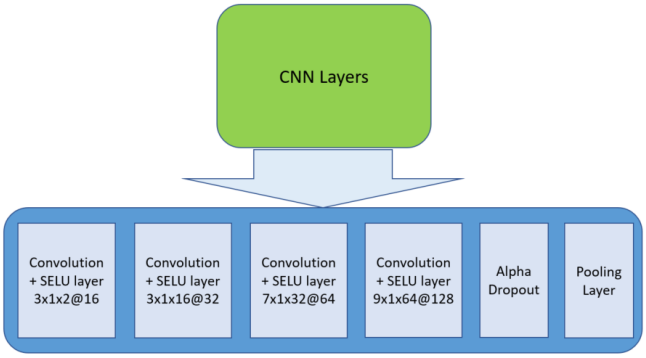

Fig. 11. Different CNN layers.

We have used 4 convolutional layers with SELU nonlinearity, one alpha drop and one pooling layer. Alpha dropout layer works effectively with SELU activation function and it goes hand-in-hand with SELU activation for self-normalization.

- First convolution layer takes 2 channels as input (M-MFCC & LPC) with 16 filters and 3x1 kernal size.
- Second convolution layer takes input from first layer and it has 32 filters with 3x1 kernal size.
- Third convolution layer takes input from 32 neurons and has 64 filters with 7x1 kernal size.
- Final convolution layer takes input from third layer and it has 128 filters with 9x1 kernal size.
- Fifth layer is alpha dropout layer to maintain variance and mean at original input level. We have used dropout rate as 0.2
- Final layer is pooling layer and we have used average pooling.

F. Cosine Triplet Embedding Loss Function

Cosine triplet embedding loss function is used during training phase to calculate loss and make correct model. Ideally Euclidean space between positive and anchor sample should be minimum but Euclidean space between anchor and negative sample should be maximum. Similar to work in [7], if we denote positive sample by Sp, anchor sample by Sa and negative sample by Sn then cosine triplet embedding loss function.

$$\mathrm{F(Sa, Sp. Sn)} = \sum_{a,p,n}^{N} cos(f(Sa, Sn) - cos(f(Sa, Sp) + \alpha_{margin}$$

Sa and Sp samples from same speaker and ideally feature distance between Sa and Sp should be zero so cosine of f(Sa,Sp) will be 1. Sa and Sn samples come from different speakers so in other way round ideal distance between Sa and Sn should be maximum so cosine of f(Sa,Sn) will be minimum. This α_{margin} is used to maintain minimum distance between negative and positive voice sample. α_{margin} is adjustable hyper-parameter. And in our experiment, we have considered it at 0.5.

G. Testing Phase

In testing phase also, we have used combined M-MFCC and LPS for feature extraction. CNN layers were trained in training phase with weight sharing together with positive, anchor and negative sample CNN layers. Same trained layers were used in testing phase to test voice samples. as per below Fig. 12.

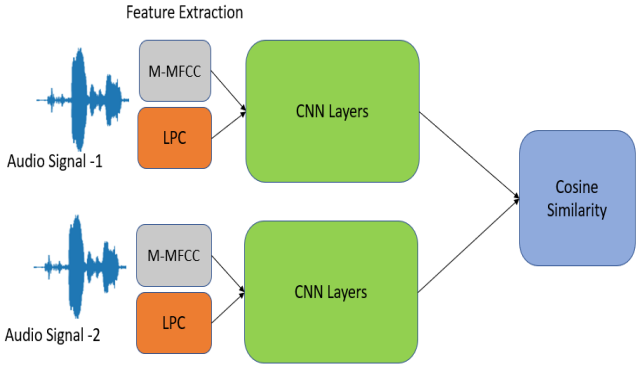

Fig. 12. CNN model testing phase.

In training we have used cosine triplet embedding loss function as we have trained model with three voice sample positive, anchor and negative audio. In testing phase, we have used cosine similarity as it is needed to validate similarity between two voice samples. Cosine similarity metric is used to compare the extracted embeddings and provide a corresponding match score as defined in Fig. 12. We have given two audio samples in testing phase Audio Sample 1 (AS1) and Audio Sample 2 (AS2). When AS1 and AS2 are from same speaker then ideally cosine similarity should 1 and when AS1 and AS2 are from two different speakers then cosine similarity should be "-1" (negative one).

VI. Experement and Result

As defined in section IV, we have created 6 different datasets from TIMIT and NOISEX-92 dataset and performed following six experiments with training and testing model defined in section V. Aim of these experiments was to train model with degraded human voice with one set of noises and

test model with different set of noises. Most of the applications need such kind of functionality to train the model in one set of noises and test in different set of noises. Below Table II sets mapping between experiment ID and corresponding training and testing dataset.

TABLE II. EXPERIMENTS WITH SIX DATASETS

Experiment ID	*Training Dataset*	*Testing Dataset*
1	DS1	DS2
2	DS2	DS1
3	DS3	DS4
4	DS4	DS3
5	DS5	DS6
6	DS6	DS5

The tables below show the comparison results for all experiments. We have compared speaker recognition system accuracy in terms of True Match Rate (TMR). TMR has calculated at 10 percent of False Match Rate (FMR). In speaker classification, increased TMR signifies reduction of false prediction (either negative or positive). It is observed that combined MFCC and LPC give significant improvement in TMR as compared to MFCC and it is similar to work in [7]. We got better result with M-MFCC and LPC as defined in Table III.

TABLE III. EXPERIMENTS RESULTS TMR

Exp ID	*TMR@FMR=10%*				
	Training Dataset	*Testing Dataset*	*MFCC*	*MFCC & LPC*	*M-MFCC & LPC*
1	DS1	DS2	65.3	80.6	89.1
2	DS2	DS1	66.9	79.7	87.9
3	DS3	DS4	63.8	77.7	83.2
4	DS4	DS3	62.9	69.4	83.4
5	DS5	DS6	60.4	69.4	72.6
6	DS6	DS5	66.6	80	92

We also compare Equal Error Rate (EER) in percentage with all three feature extraction methods. EER represents a point in DET or ROC curve where false rejection rate is equal to false acceptance rate. When both rates (false acceptance and false rejection) are equal then that common value is considered as EER. Low EER represents improved accuracy of speaker recognition system. As per Table IV results, we observed that combined MFCC and LPC gives very significant improvement in EER as compared to MFCC [7] and on top of this we got improvement with M-MFCC and LPC case.

TABLE IV. EXPERIMENTS RESULTS EER

Exp ID	*ERR in %*				
	Training Dataset	*Testing Dataset*	*MFCC*	*MFCC & LPC*	*M-MFCC & LPC*
1	DS1	DS2	21.8	12.3	10.5
2	DS2	DS1	20.2	12.5	11.3
3	DS3	DS4	20	12.8	12.1
4	DS4	DS3	22	14.4	12.0
5	DS5	DS6	23.5	14.3	13.8
6	DS6	DS5	18.7	12.5	10.8

VII. CONCLUSION

Different applications require different types of speaker recognition system. Most of the applications are looking for speaker recognition system which should work with no background noise during training, but same system should recognize speaker even with degraded human voice. In our experiment we have used same approach and used speaker voice with most silent background noise (SNR = 10-12dB) and test degraded human voice up to 1dB SNR. In our experiment (As per section V) we observed that high sampling rate, optimized triangular mel-bandpass filter frequency range and optimized pre-emphasis value gives better EER and TMR. Improved EER and TMR are impacting MFCC values for effective speaker recognition process. This can further optimize in future to get better results. In future, triangular mel-bandpass filter frequency range can optimize further to get better results.

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] 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

[5] 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

[6] 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

[7] 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

[8] S. Nakagawa, L. Wang and S. Ohtsuka, "Speaker Identification and Verification by Combining MFCC and Phase Information," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1085-1095, May 2012

[9] 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

[10] 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

[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

[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] Monalisha Barik, Susanta Kumar Sarangi and Sushanta Kumar Sahu,

[14] "Real-time speaker identification system using cepstral features", in Communication Control and Intelligent Systems (CCIS), 2016 2nd International Conference on, March 2017

[15] 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

[16] Z. Weng, L. Li and D. Guo, "Speaker recognition using weighted dynamic MFCC based on GMM," *2010 International Conference on Anti-Counterfeiting, Security and Identification*, 2010, pp. 285-288

[17] H. C. Bao and Z. C. Juan, "The research of speaker recognition based on GMM and SVM," *2012 International Conference on System Science and Engineering (ICSSE)*, 2012, pp. 373-375

[18] Z. Weng, L. Li and D. Guo, "Speaker recognition using weighted dynamic MFCC based on GMM," *2010 International Conference on Anti-Counterfeiting, Security and Identification*, 2010, pp. 285-288

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

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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

[25] S. J. Wenndt and R. L. Mitchell, "Machine recognition vs human recognition of voices," 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, 2012, pp. 4245-4248

[26] 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

[27] 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

[28] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Commun., vol. 12, pp. 247–251, Jul. 1993

[29] F. Beritelli, "Effect of background noise on the SNR estimation of biometric parameters in forensic speaker recognition," 2008 2nd International Conference on Signal Processing and Communication Systems, 2008, pp. 1-5

[30] P. Papadopoulos, A. Tsiartas, J. Gibson and S. Narayanan, "A supervised signal-to-noise ratio estimation of speech signals," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 8237-8241

[31] Z. Liu, Z. Wu, T. Li, J. Li and C. Shen, "GMM and CNN Hybrid Method for Short Utterance Speaker Recognition," in *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 3244-3252, July 2018

[32] Z. Huang, M. Dong, Q. Mao and Y. Zhan, "Speech emotion recognition using CNN", *Proc. ACM Int. Conf. Multimedia*, pp. 801-804, 2014

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

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

[35] K. I. Nordstrom, G. Tzanetakis and P. F. Driessen, "Transforming Perceived Vocal Effort and Breathiness Using Adaptive Pre-Emphasis Linear Prediction," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1087-1096, Aug. 2008

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.