

An Evaluation of MFCC and ZCR Feature Extraction with CNN Model of Speaker Identification System

N.Kaladharan¹ and Dr.R.Arunkumar²

¹Lecturer, Dept. of Computer Engineering, Government Polytechnic College,Theni

²Associate Professor, Department of Computer Science and Engineering, FEAT, Annamalai University

Abstract—The speech is the important medium to transfer our feeling to the neighbors. Speaker recognition is the process to recognize our speech in our owned database. Speaker recognition has divided in to speaker identification and speaker verification. Speaker identification is different mode which is text dependent, independent and prompted. Identifying the speaker in our database is the major task in our process. Identifying data or speech is the main aim to handle in our database. Speaker identification skills are broadly useful in security and investigation, voice confirmation, electronic voice snooping and classify verification. Various features extraction ideas are used in identification as well as verification such as mfcc, zcr, pncc and so on. MFCC are cepstral coefficients resultant on a warped frequency scale centered on human hearing observation. In the computation of MFCC, the first thing is windowing the speech signal to split the speech signal into frames. The zero-crossing rate is the proportion at which the signal modifications from positive to zero to negative or from negative to zero to positive. Power Normalized Cepstral Coefficients is well identified for the high accuracy of automatic speech recognition systems even in high-noise situations. This paper presents the comparison of mfcc and zcr with the CNN modeling system. CNN is the top modeling idea in the recent decades.

Index Terms— MFCC, ZCR, CNN, SNR etc.

I. INTRODUCTION

Speech is the essential property of human being to convey their feelings and thoughts. Speaker recognition technique is the idea of voice recognizing in the stored database. Speaker recognition classified in to two different categories. Speaker identification and speaker verification. Speaker identification is the process to identify the voice of speakers in a pre-determined datasets. Speaker identification again classified in to text dependent, text independent and prompted text identification. Independent text identification illustrates has no model requires for speech data. Whereas text dependent speaker model identification needs the same content while test and train. Prompted text speaker identification indicates the model randomly selects some data for the speaker to say. Speaker identification process depends two most important tasks (i) feature extraction and (ii) dimensions reduction and classification.

This paper is organized as follows, Section II describes literature survey, Section III represent types of speaker recognition, Section IV approaches to speaker identification, Section V deals the existing work, VI express the proposed method, Section VII denotes Convolutional Neural Network, Section VIII refers results and discussion, Section IX deals the conclusion.

II. LITERATURE SURVEY

Many challenging factors are considered while developing a speaker recognition model such as health of the speaker, environmental noise, background factors and throat characteristics of different speaker, age, gender, emotions and so on. Whenever the application depends, speaker recognition can be widely divided into three types: speaker identification, speaker verification and speaker diarization [1]. Speaker identification mainly is used in the academic and industry fields for surveillance, diarization and speaker embedding learning purpose [2]. Speaker can be designed and modeled using these features with addition of various modeling techniques like mfcc, bfcc, zcr, lpcc, and many [3]. Speaker identification has three important tasks which are feature extraction, length discount and ordering [4]. CNN design to extract significant discriminative structures from spectrograms for performing speech emotion recognition [5]. Speaker identification system classifies a human being voice according to speech categories of men or women. [6]. Mfcc and zcr are the common techniques used to extract the feature of the speech signal. The second step of speaker identification system is speaker modeling [7].

III. SPEAKER RECOGNITION

Speaker recognition is an exercise of recognizing individual persons from their speech. Due to different voice making organs such as vocal tract forms, larynx shape, glottis type and some other parts, not all individuals have same voice. SI is categorized into two divisions. They are open set and closed set.

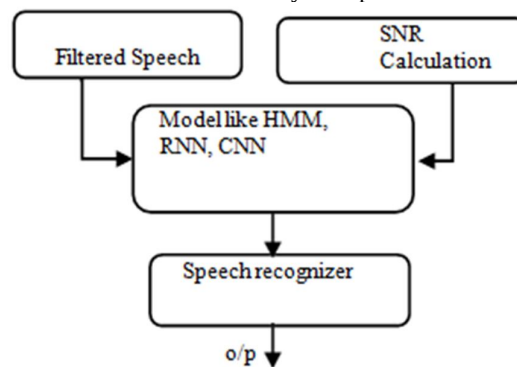


Fig.1. Speech recognizer model

In position of closed set speaker identification, the authorized person of the unknown voice sample is one of the known speakers. In exposed regular speaker identification, it is not known that the authorized speaker of the unknown voice trial is extant in the mention or not. Speaker recognition approaches can likewise be divided into text dependent and text independent methods. In case of text dependent methods a speaker is essential to utter a planned set of words or sentences. In case of text independent method, there is no determined set of words or sentences and the speaker's may not level be alert that they are being verified. Features of speech are extracted from the alike utterances.

IV. SPEAKER IDENTIFICATION

Speaker identification is the responsibility to identify an unknown speaker from a set of already known speakers. The traditional speaker identification task frequently involves three stages: Pre-processing, feature extraction and speaker modeling.

Speech pre-processing comprises preemphasis, framing, and windowing etc. Feature extraction is to extract the individual limits that efficiently illustrate the speaker's traits from the pre-processed speech signal. The extracted features are then sent to model for training or identification. Speaker identification system approves a human being according to his or her speech patterns.

When all speakers within a specified set are known, it is called closed set. Speaker identification task can be divided into two essentially processes, which are the training and the testing process. The training method can also

be divided in to four stages as.1.Entering the speech signal, 2. Pre handling of the speech, 3.regularization and 4. Feature withdrawal. Feature extraction is the greatest important step for speaker identification.

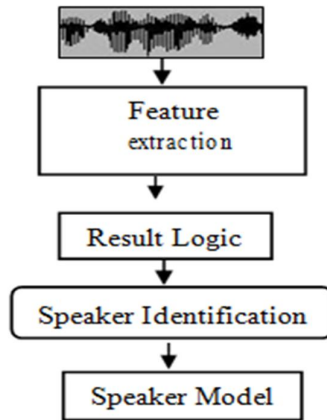


Fig.2.Block diagram of Speaker Identification model

V. EXISTING WORK

There are various tactics to solving the speaker recognition problem. The solution is defined by two constraints: The feature of the voice signal to be used like Mel Frequency Cepstral Coefficients (MFCC), ZCR and the modelling technique used to learn the voice samples like CNN. Trainings have also remained done where numerous techniques are combined, such as using different models for feature extraction and classification. Common approaches use MFCC as the feature to be used since it is most active for speaker recognition.

VI. PROPOSED METHOD

A. Feature Extraction

Feature extraction is a frequently important task in the speaker recognition. Generally, it is the process; the speech signal is converted in to a set of 12-15 feature or feature quantity for further proceeding in the design. Feature extraction system achieves the conversion of raw speech signal in to a set of feature quantity, so the speaker special characteristics are more concentrated and analyzed. Various kinds of features are available in the speaker recognition like short term spectral feature, voice source feature, spectra temporal features, high level and prosodic feature and so on. The other problem to consider while increasing the speaker, then there is a possibility of extracting likely features by two are more participants who have a same voice. The subsequent points are consider to select the feature for the good speaker is, (i) feature should not be affected by a speaker's health and mature like cold, coughed. (ii) mimicries of other's voice and (iii) free of noise, are the difficult to achieve the characters. To extract feature from speech signal, user often divide the voice signals in to units, in the form of continuous signals. A unit is a tiny time period, e.g. T, the length of which is fixed. The unit is known as frame, which will extract the feature.

B. Mel Frequency Cepstral Coefficient (Mfcc)

This feature extraction model was introduced by bridle and brown in 1947 and developed by David and Mermelstain in 1976 for speech signal. The human ear obtain frequencies less than 1 kHz at linear scale with frequencies higher than 1 kHz are being received at logarithmic scale. This things of human ear is practice in Mfcc. Mfcc is used all zero type for figuring ranges. It also treated as an algorithm used for speech parameterization that used a frequency scale called Mel scale. This scale at is linear at low frequencies and logarithmic in high as 1 kHz, it is given by the eqn. The eqn which converts frequency f to Mel scale is

$$f_{mel} = 2595X \log(1+ f / 700) \text{ --- (1)}$$

Mel is a measuring unit based on the perceived frequency of human ears. F represents the physical frequency and Mel (f) Represents perceived frequency.

A set of coefficients is calculated accurately to the number of filters and applied for a windowed speech signal. The mfcc coefficients are computed from the energies of a triangular filter band at the Mel frequency scale. Triangular Mel filter are used during the calculation of mfcc. The mfcc feature extraction process comprises framing, windowing, and discrete Fourier transform, logarithm of magnitude, warping frequency on Mel scale

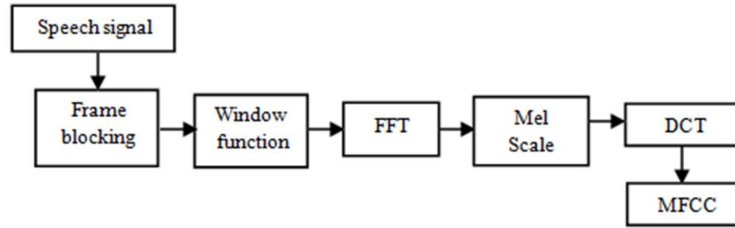


Fig.3.Block diagram of MFCC model

and applying DCTs. Each speaker utterances were divided in a frame length of 25ms. Moreover 10ms overlapping was introduced in successive frame to avoid data loss.

C. Zero Crossing rate (ZCR)

The ZCR is an audiogram is defined as the rate at which the signal changes the sign. ZCR is the powerful and simple technique to detect whether a speech length frame is a voice, unvoiced or silent. It is predictable that unvoiced segment generates higher zcr then for voices segment and ideally zcr equal to zero for silence segments. It is the basic technique to calculate the fundamental frequency, in the lowest frequency of a timing waveforms. So that the zcr, feature can therefore be used to design discriminator and classifies systems. Zcr is calculated as follows for frame X_r of length N :

$$zcr = 0.5 \sum_{n=1}^N [\text{sign}(x[n]) - \text{sign}(x[n-1])]$$

Where, r refers to the number of current frames, x_r refers to the frame in the time domain, where n is the time index and the sign function is defined by

$$\text{Sign}(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

Zcr gives a good indicates of the spectral data of a signal. The difficulty of zcr is it ensured not take noise into interpretation.

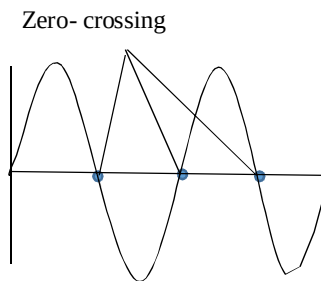


Fig .4. Wave representation of ZCR

VII. CONVOLUTIONAL NEURAL NETWORK

CNN is one of the famous neural networks in deep learning technique. It has a hierarchical neural network which consists of a variability layers in sequence. Here convolution process is the main process. So it known as convolution network. CNN network came from the following important points which are sparse interaction,

parameter sharing and equivalent representation In CNN based approach, we create a neural network compose of multiple hidden layers. Layers contains convolution layer, pooling layer, batch normalization and dropout etc. In this network, in order to learn and extract the features from input, three stages are used which are convolution stage, ReLU and pooling stage.

A Sequence of convolutional, pooling and fully connected layers form the feature extraction pipeline which

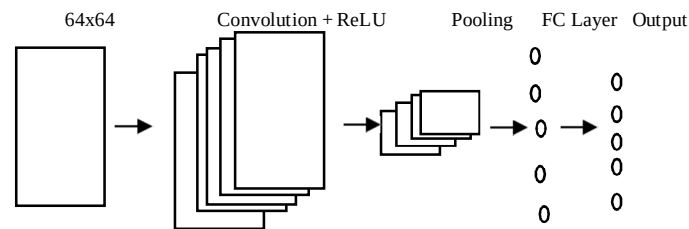


Fig .5. Stages of CNN model

models the input data in abstract form. By ending, the softmax layer executes the final classification task based on the target.64X64 of input layers are used and produces the convolution results plus Relu values.

VIII. RESULTS AND DISCUSSION

The testing was conducted to obtain results of both identification idea mentioned above. The tests are directed 20 speakers, where 10 female and 10 male voices. For each speaker, five iterations are fixed. And set mfcc coefficients of 13, 19, 21 and 25 numbers. The number of speaker is varied from 5 to 20 and to observe the changes of increasing the number of speakers on identification systemmodel.

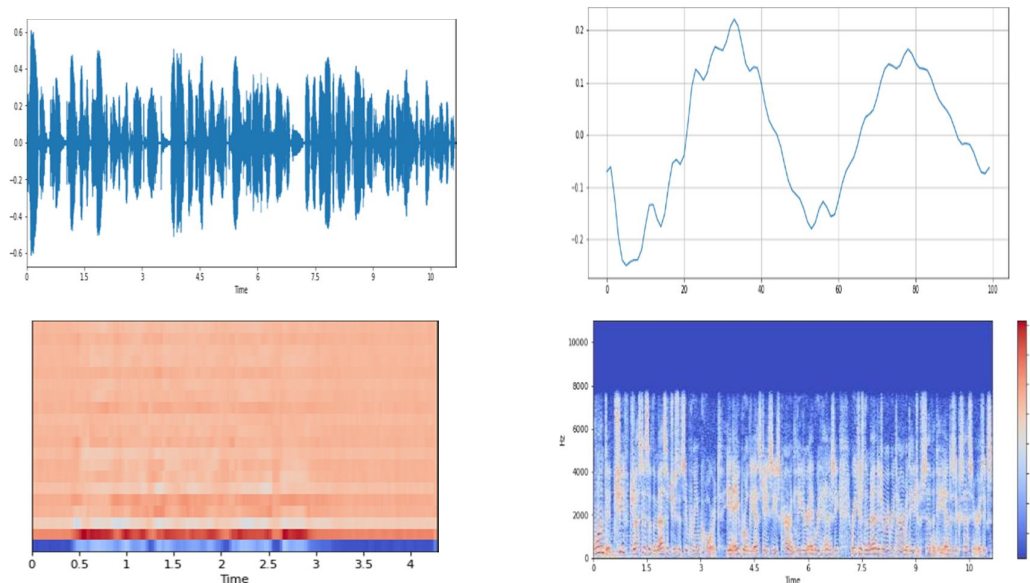


Fig.6 Mfcc for voice signal

The comparison of the mfcc with cnn is analyzed with respect of time vs frequency of this section. The numbers of speakers are taken randomly among the 20 speakers.

The number of mfcc coefficients has been varied 13, 19, 21 and 25 to process that, the recognition system gives better output in terms of identification rate. In Zcr the number of frames also be increased somewhat the identification rate is also increased.

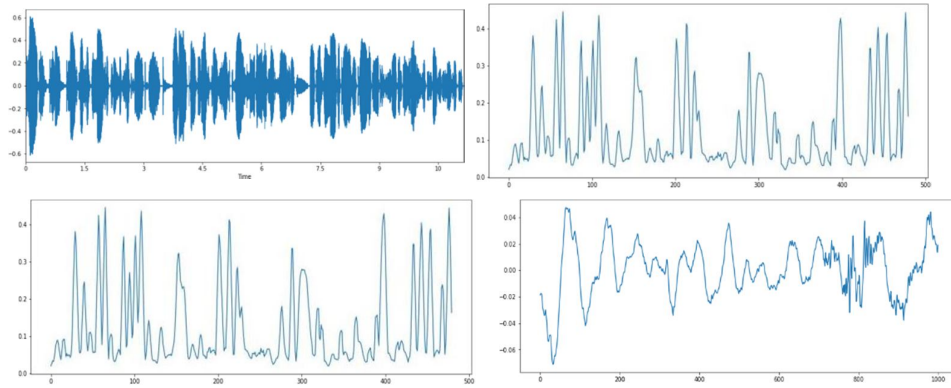


Fig.7 ZCR value for voice signal

TABLE I. IDENTIFICATION VS NO OF SPEAKERS

No of speakers	MFCC-CNN	ZCR-CNN
5	96.1	70.91
10	93.2	72.3
13	95.2	70.2
15	89.2	73.2
18	86.0	60.0
20	85.2	65.0

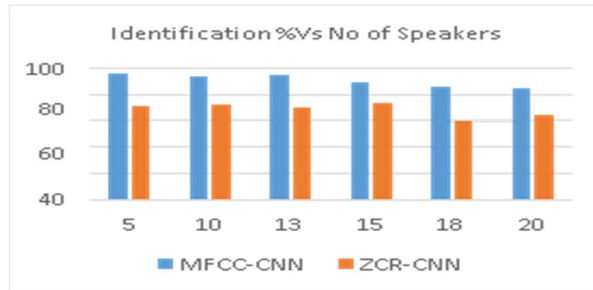


Chart I. Comparison of Cnn with mfcc and z

TABLE II. COEFFICIENTS, FRAMES VS MODEL

No of coeffi	Mfcc-cnn	No of frames	Zcr- CNN
13	8.5	20	7.0
19	9.5	40	8.9
21	10.15	50	10.0
25	10.50	60	12.5

The design also performed on male and female voices to evaluate the effect of gender of identification model. Thereare 10 female speaker were taken to execute the effect of gender onidentification rates.

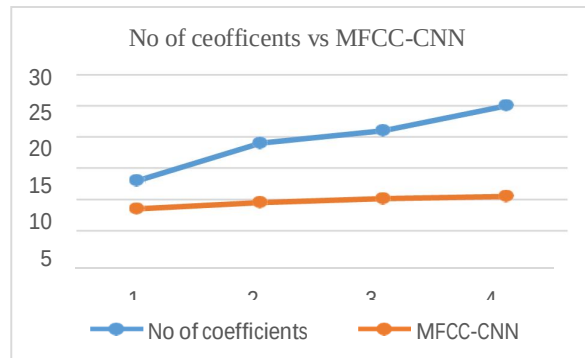


Chart II No of coefficients vs MFCC-CNN

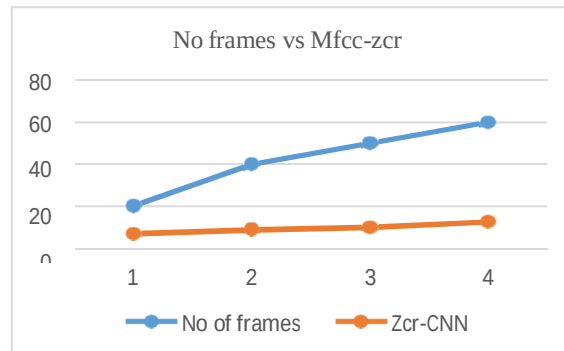


Chart III No frames vs Mfcc-zcr

TABLE III. ACCURACY CLASSIFICATION

Type of speaker	MFCC- CNN	Mfcc-Zcr
male 1	84.5	80.11
male 2	89.0	76.0
male 3	92.0	74.1
male 4	91.2	71.0
male 5	93.2	71.5
male 6	92.5	75.2
male 7	93.5	70.2
male 8	92.5	70.5
male 9	93.4	70.8
male 10	92	70.9
Average	92	70
female1	84.5	80.11
female 2	89.8	76.5
female 3	92.9	74.1
female 4	91.2	75.0
female 5	93.2	75.5
female 6	92.5	75.2
female 7	93.5	70.2
female 8	92.5	75.5
female 9	93.4	75.8
Female10	93	70.9
Average	93.4	75.5

The results shown that for the both male and female voice identification rates for the both systems female speakers are higher than male speakers. The female speakers have high frequency than male speakers, so it seems high. Hence, the feature vector of mfcc and zcr are used for particularly female speaker. SNR is calculated individually for mfcc and ZCR model of combinations. The output of SNR value is nearly same as to input voice.

Table IV. SNR Accuracy

Speakers	Mfcc-cnn	Zcr-cnn	SNR
Male voice	92	70	90%
Female voice	93.4	75.5	89%

VIII. CONCLUSION

In this paper, the two models of speaker identification systems are designed as CNN with mfcc and cnn with zcr. This two model is compared with gender and computation time and identification rate. The classification reveals the accuracy of 92 % of MFCC with CNN and 70% of ZCR with CNN for male speakers and 93.4% of MFCC with CNN and 75.5% of ZCR with CNN of female voices. The number speakers and number of coefficients increases, the computational time also high with high complexity. Female speakers easily identified when compared to male voice in both models. The voice of 20 speakers test is gives better results with better time. It is also found, in all systems combination of speaker identification model, the mfcc feature extraction having better recognition rate as well as computation time than the zcr model. So mfcc with cnn model is giving suitable identification than zcr with cnn.

REFERENCES

- [1] Amna irum etal, "Speaker verification using deep neural networks: a review", IJMLC, 2018.
- [2] Feras etal, "Speaker identification based on hybrid feature extraction techniques", IJCSA, 2019
- [3] Aws Al Qaisi, "Arabic word dependent speaker identification system using artificial neural network", IJCSS vol 14, 20
- [4] Nayana PK. etal, "Comparison of text independent speaker identification systems using gmm and i-vector methods, Elsevier, 2017
- [5] Abdul malik badsha, etal, "Speech emotion recognition from spectrograms with deep convolutional neural network", IEEE, 2017
- [6] Rania Charon, etal, "Efficient text independent speaker recognition with short utterances in both clean and uncontrolled environments", Springer, 2020
- [7] Chandra kumar et al, "Analysis of mfcc and bfcc in speaker identification system", IEEE 2018
- [8] Kydyrbekova aizat, etal, "Identification and authentication of user voice using DNN features and i-vector", Cogent engg, Gopal chaudry, etal, "Feature extraction methods for speaker recognition : a review" IJPRAI, 2017
- [9] Daniele Savati, etal, "End to end speaker identification in noisy and reverberant environments using raw waveform convolutional neural networks", Inter speech, 2019
- [10] Feng Ye, etal, "A deep neural network model for speaker identification", Applied sciences, MDPI, 2021
- [11] Supapon Bunrit, et.al, "Text independent speaker identification using deep learning model of convolution neural network", IJMLC, vol 9. 13. Anett Antony, R. Gopikakumari, "Speaker identification based on combination of MFCC and UMRT based features", ICACC-2018, Elsevier
- [12] V. M. Sardar, etal, "Speaker identification of whispering speech: an investigation on selected timbral features and KNN distance measures", International Journal of Speech Technology, 2018
- [13] Sabur Ajibola Alim, etal, "Some Commonly Used Speech Feature Extraction algorithms" Licensee IntechOpen. 2018
- [14] F. Z. Chelali, etal, "Speaker identification system using LPC-Application on Berber language", HDSKoD journal
- [15] Ankur Maurya, etal, "Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach", Elsevier, 2018
- [16] NGUYEN NANG AN, etal, "Deep CNNs With Self-Attention for Speaker Identification", IEEE, 2019
- [17] Xingmei Wang, etal, "A network model of speaker identification with new feature extraction methods and asymmetric BLSTM", Neurocomputing 403 (2020)
- [18] R. Visalakshi, etal, "Performance of speaker identification using CSM and TM" Int J Speech Technol, 2016.
- [19] Saritha Kinkiri, etal, "Speaker Identification: Variations of a Human voice" IEEE, 2020