Speaker Recognition using Deep Learning

Vikram Adhitya S, Nachiket Timmanagoudar, Shrut Makde,Prof.Christopher Clement J

School of electronics engineering, VIT University
Vellore, India

Abstract **- Speaker recognition is a task of identifying persons from their voices. Recently, deep learning has dramatically revolutionized speaker recognition. However, there is lack of comprehensive reviews on the exciting progress. In this paper, we implement a Deep learning based approach for classification of speech samples. The major advantage of deep learning over conventional methods is its representation ability, which is able to produce highly abstract embedding features from utterances, we first pay close attention to feature extraction methods applied to the speech samples and move onto classification of the samples using feeding the features into a neural network. Finally, we analyze the results produced the classifier using appropriate metrics.**

INTRODUCTION

Given the speaker's distinct pronunciation organs and speaking method, it is well recognised that a speaker's voice has personal features[1]. As a result, a computer can automatically recognise a speaker based on his or her voice. Automatic speaker recognition is the name given to this technique. Speaker recognition is a basic task in voice processing that has a wide range of applications in real-world situations. It's utilised for voice-based identification of personal smart devices like cell phones, cars, and laptops, for example. It ensures the security of bank transactions and distant payments[1]. It has been commonly used in forensics to determine whether or not a person is guilty. It's crucial for retrieving audio-based information from broadcast news, meeting recordings, and phone calls.

Speaker recognition is a useful biometric feature recognition technique. It is the task of determining someone's identify based on their voice signal. Speaker recognition is a valuable biometric recognition technology[6] that has been used in a variety of applications, including safe access to highly secure regions, voice dialling devices, banking, databases, and computers. For many years, experts in several domains of information security have been paying growing attention to speaker recognition due to the unique properties of speech signals. The application of statistical approaches to identify persons based on their distinctive auditory qualities, which are encoded in a sequence of successive samples in time, is known as speaker recognition research. Speaker recognition can be divided into two modes, depending on the application: speaker verification and speaker identification[7].

The former investigates whether the alleged speaker is the genuine article, while the latter seeks to identify the author. We primarily perform research on speaker identification in this study. Speaker identification is the process of detecting an unknown speaker's identity based on their voice signal. It is a match processing method in which the extracted feature is compared to several templates. Extracting speaker attributes from voice signals is a critical step in the speaker identification process. Human voice signals are a powerful type of communication that carry a wealth of information, including gender, emotional qualities, accent, and so on. Researchers can use voiceprint recognition to identify speakers thanks to these distinct traits[2-3]. For training, the gathered speaker utterances are fed into a deep learning network. The speaker identification system compares the retrieved speaker features to those in the model library during the recognition phase. The target speaker is then identified as the speaker with the highest probability of utterance. However, speaker identification stability is insufficient. The length of the voice, the voice collection environment, and the speaker's physical condition all influence recognition[8]. Furthermore, in a loud setting, the recognition performance is often poor.

MATERIALS AND METHODS

I. Speech Feature Extraction

Extraction of features is a very important part in analyzing and finding relations between different things. The data provided of audio cannot be understood by the models directly to convert them into an understandable format feature extraction is used. It is a process that explains most of the data but in an understandable way. Feature extraction is required for classification, prediction and recommendation algorithms. **The following features have been extracted from the speech files:**

Mel-frequency Cepstral Coefficients:

MFCC's are based on the known variation of the human ear's critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech[9-11]. This is expressed in the Mel-frequency scale, which is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

The speech input is typically recorded at a sampling rate above 10000 Hz. This sampling frequency was chosen to minimize the effects of aliasing in the analog-to digital conversion. These sampled signals can capture all frequencies up to 5 kHz, which cover most energy of sounds that are generated by humans. As been discussed previously, the main purpose of the MFCC processor is to mimic the behavior of the human ears. In addition, rather than the speech waveforms themselves, MFFC's are shown to be less susceptible to mentioned variations

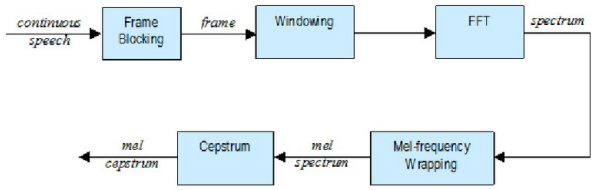

FIGURE 1: BLOCK DIAGRAM OF THE MFCC PROCESSOR

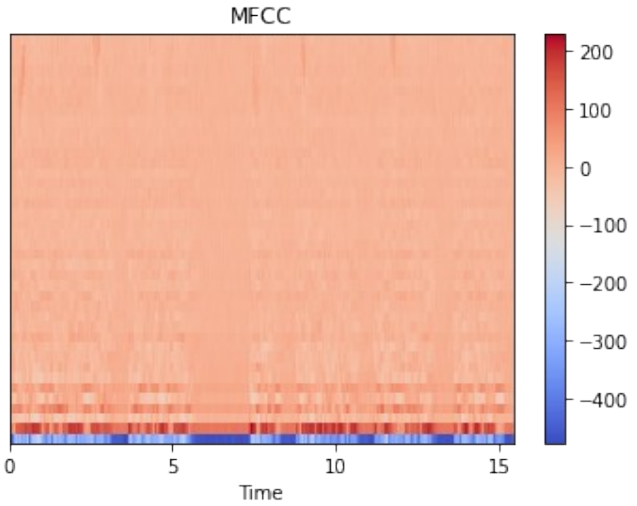

FIGURE 2: MEL FREQUENCY CEPSTRAL COEFFICIENTS

Chromagram:

Chroma features are an interesting and powerful representation for audio in which the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave[1]. Since, in music, notes exactly one octave apart are perceived as particularly similar, knowing the distribution of chroma even without the absolute frequency (i.e. the original octave) can give useful musical information about the audio -- and may even reveal perceived musical similarity that is not apparent in the original spectra.

There are many ways for converting an audio recording into a chromagram. In this project, the conversion of an audio recording into a chroma representation (or chromagram) was be performed by using short-time Fourier transforms in combination with binning strategies. Furthermore, the properties of chroma features can be significantly changed by introducing suitable pre- and post processing steps modifying spectral, temporal, and dynamical aspects.

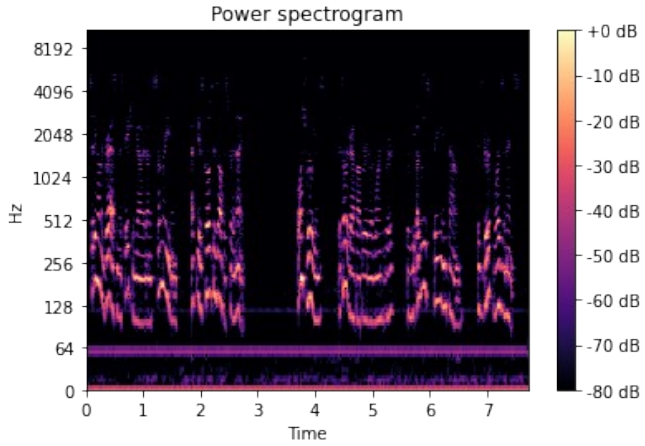

FIGURE 3: STFT IN A POWER SPECTROGRAM

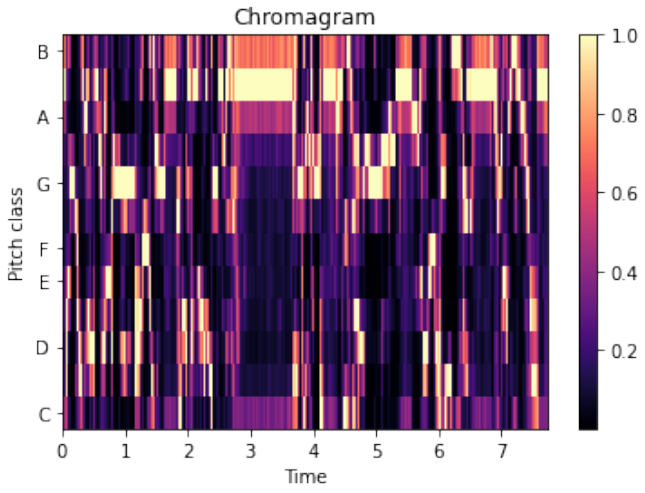

FIGURE 4: CHROMA FEATURES

Mel Spectrogram:

Mel spectrogram is a spectrogram that is converted to a Mel scale. A spectrogram is a visualization of the frequency spectrum of a signal, where the frequency spectrum of a signal is the frequency range that is contained by the signal. The Mel scale mimics how the human ear works, with research showing humans don't perceive frequencies on a linear scale. Humans are better at detecting differences at lower frequencies than at higher frequencies.

The Mel Scale, mathematically speaking, is the result of some non-linear transformation of the frequency scale[2-7]. This Mel Scale is constructed such that sounds of equal distance from each other on the Mel Scale, also "sound" to humans as they are equal in distance from one another. In contrast to Hz scale, where the difference between 500 and 1000 Hz is obvious, whereas the difference between 7500 and 8000 Hz is barely noticeable.

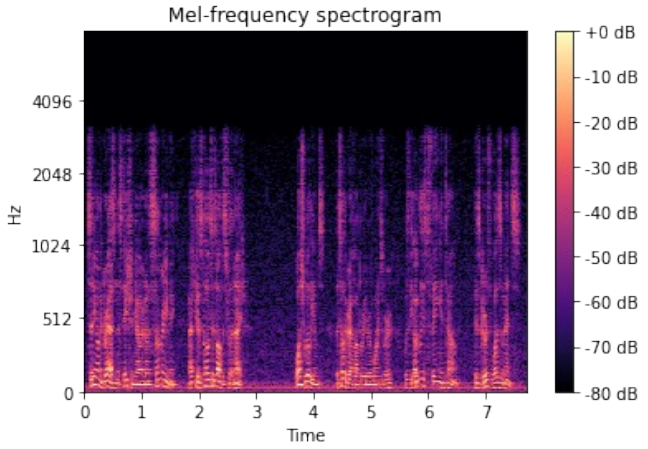

FIGURE 5: MEL SPECTROGRAM

Spectral Contrast:

Spectral contrast is defined as the decibel difference between peaks and valleys in the spectrum. There are two general motivations behind spectral contrast enhancement for hearing-impaired (HI) people. First, in a sensorineural-impaired cochlea, auditory filters[8-9] are generally broader than the normal and are in many cases abnormally asymmetrical. Processing through these abnormal filters may produce a smearing of spectral detail in the internal representation of acoustic stimuli. Differences in amplitudes between peaks and valleys in the input spectrum may be reduced, making it more difficult to locate spectral prominence (i.e.,formants) which provide crucial cues to speech intelligibility. To enhance spectral contrast may be of some help in compensating for the effects of this reduced frequency selectivity. Second, spectral analysis of speech in noise typically shows that these formants are well represented only when the input signal-to-noise ratio (SNR) is large enough but the spectral valleys between the formants are filled with noise. Therefore, spectral contrast enhancement may be beneficial for the noise reduction.

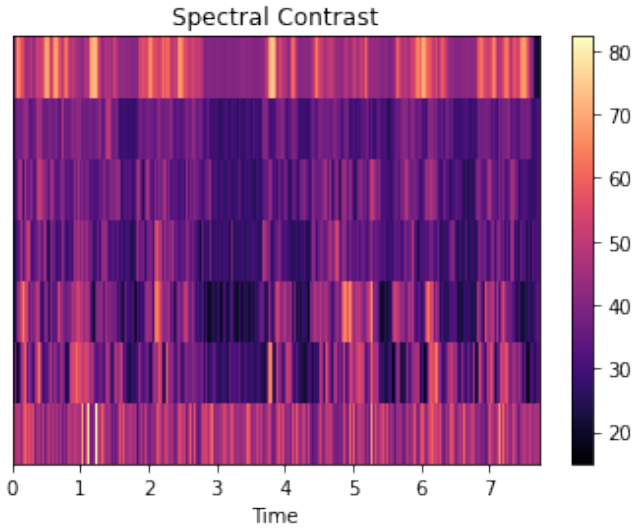

FIGURE 6: SPECTRAL CONTRAST

Tonal Centroid Features:

The Tonal Centroids (or Tonnetz) contain harmonic content of a given audio signal. The harmonic table note layout is a recently developed musical interface that uses a note layout topologically equivalent to the *Tonnetz*. A *Tonnetz* of the syntonic temparament can be derived from a given isomorphic keyboard by connecting lines of successive perfect fifths, lines of successive major thirds, and lines of successive minor thirds. Like a *Tonnetz* itself, the isomorphic keyboard is tuning invariant. The topology of the syntonic temparament's *Tonnetz* is generally cylindrical. The *Tonnetz* is the dual graphs of Schoenberg's chart of the regions, and of course *vice versa*. Research into music cognition has demonstrated that the human brain uses a "chart of the regions" to process tonal relationships.

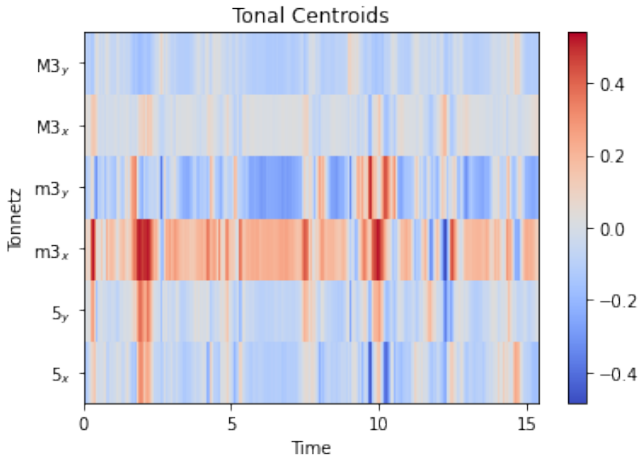

FIGURE 7: TONAL CENTROID FEATURES

II. Model Overview

In this section, we mainly describe the model that was used as our classifier. We used a simple feed forward Neural Network with 3 Dense layers[10]. The input layer has an input size of 193 which the length of out features list consisting MFCCs, chromagram, Mel Spectrogram, Spectral Contrast and Tonal Centroid Features. The input Layer and the the 2 middle layer uses ReLU (Rectified Linear Unit) activation layers. The output layer has a size of 30 because we have 30 different classes and uses softmax activation[1] and hence we also use categorical crossentropy[7] while fitting the model. The following content describes the entire architecture of the model in detail.

Figure 8 shows the different layers used in the Model and describes the output shape and number of Parameters in each Layer of the Neural Network. Figure 9 shows the architecture of the Neural Network with the same input and output shapes described in Figure 8.

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 193)	37442
dropout (Dropout)	(None, 193)	0
dense_1 (Dense)	(None, 128)	24832
dropout_1 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 128)	16512
dropout_2 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 30)	3870

Total params: 82,656
Trainable params: 82,656
Non-trainable params: 0

FIGURE 8: SUMMARY OF MODEL LAYERS

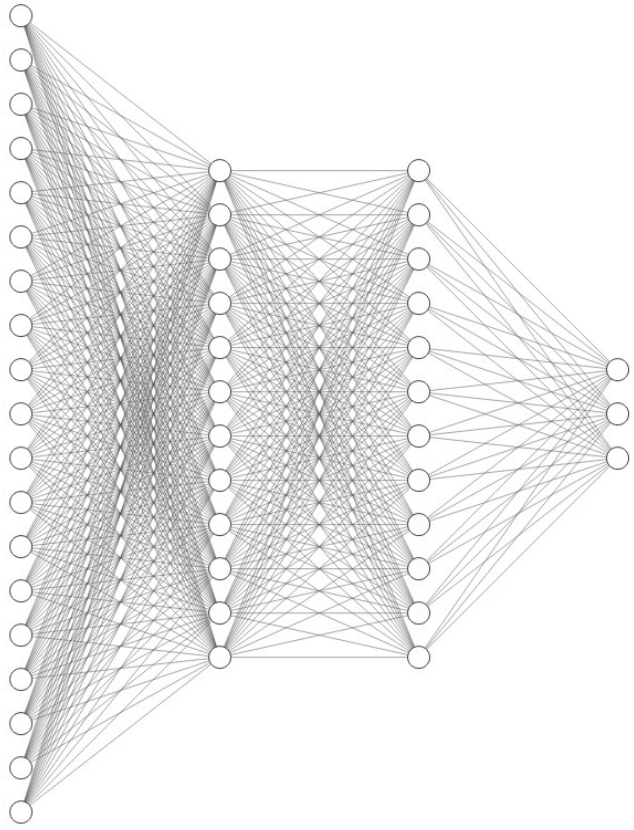

FIGURE 9: FULLY CONNECTED NEURAL NETWORK ARCHITECTURE (SCALED)
(193->128->128->30)

EXPERIMENTAL SETUP

Speaker Dataset:

The dataset for the project was made from the LibriSpeech audio files data. We have used the audio samples from 30 different speakers and 4 samples per speaker for our training data, giving us 120 samples. All the samples are around 15 seconds therefore giving approximately one minute per speaker for training. Training,Validation and Test sets were taken in the Ratio 4:3:3 resulting in 90 samples each for Validation and Testing. Therefore a total of 300 audio samples were taken in our Dataset. Table 1 shows the dataset information below.

TABLE I
USE OF LIBRISPEECH DATASET IN EXPERIMENTAL SETUP

Dataset	Number of Speakers	Ratio	Size of each Clip
Training Set	30	40%	15-16 seconds
Validation Set	30	30%	15-16 seconds
Test Set	30	30%	15-16 seconds

Data Preprocessing:

The names of the audio files and the speaker id for each audio file are put into a pandas dataframe for easy processing. Dataframes are created for training, validation and test sets.

The audio signal features listed in the speech feature extraction section above are extracted for all the elements (audio files) in the training , validation and test dataframes and the respective dataframes are updated with the extracted features as the features to be input for the classfication model, and the speaker id's are used as labels.

The feature extraction from the audio files was done using the feature extraction modules in the Librosa Signal processing package. The training set is then feature scaled using the scikit learn Standard Scaler and the labels are encoded using the scikit learn label encoder.

RESULTS

Training:

The size of features in the training set is 193 which is set as the input size for the classification model. The model is fit the with the training and Validation sets and is trained for 100 Epochs with dropouts and early stopping to prevent the model from overfitting the data.

Evaluation Metrics

Predictions are made the by the model using the test features and the predictions are added as a new parameter to the test set dataframe for ease of analysis. The performance of the model is analysed by using the metrics below.

Figure 10 is the confusion matrix indicating 4 incorrect predictions by the model in the test set for 30 speakers. The matrix was created with the sns heatmap and scikit learn metrics python packages taking the predictions test labels as parameters. Figure 11 highlights the changes in training and validation accuracy. The graph shows a good validation curve resulting in high prediction accuracy of the model. Figure 11 shows the Training and Validation losses and the

curve indicates the mode is fitting new data very well. The loss and accuracy were taken from the Neural Network model as they are parameters of the model history and records the values of validation and training accuracy and loss at each epoch during training. The graphs were generated by taking the loss and accuracy easily using the simple matplotlib python package. Prediction Accuracy of the Model was found by calculations which showed 95.5% of predictions made by the model were correct which can also be inferred from the confusion matrix in Figure 10.

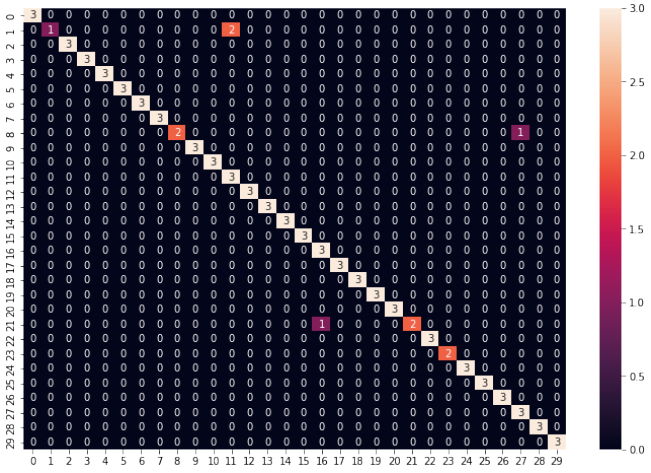

FIGURE 10 CONFUSION MATRIX

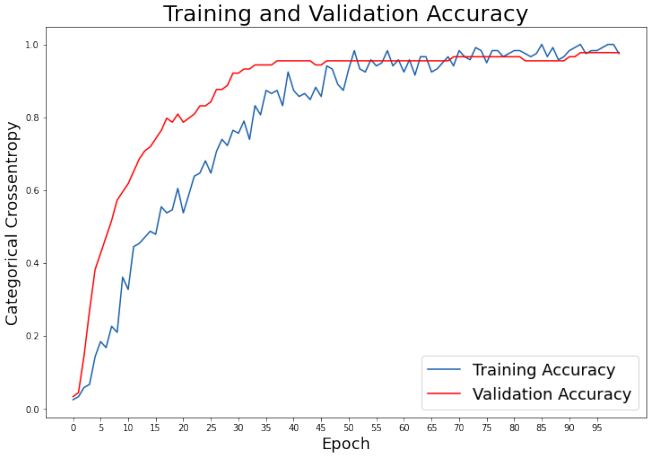

FIGURE 11 TRAINING AND VALIDATION ACCURACY VS EPOCHS

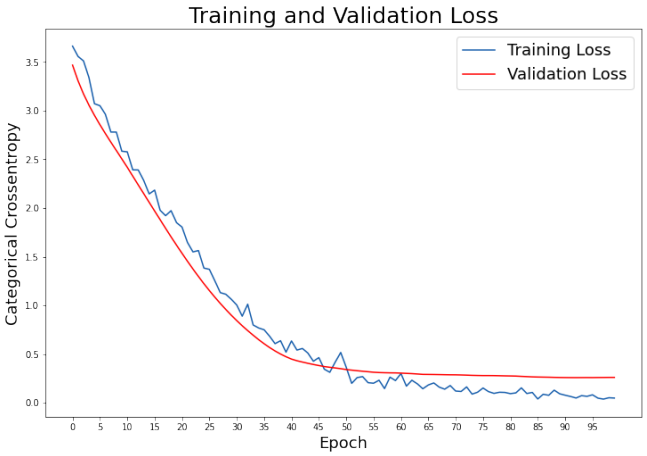

FIGURE 12 TRAINING AND VALIDATION LOSS VS EPOCHS

CONCLUSION

In this paper we discuss the working of a Neural Network in Speaker speech sample classification. Features such as MFCCs Mel Spectrogram and Tonal Centroid Features were extracted. The addition of Mel Scale features showed the best results in relation to human speech recognition. The Mel scale mimics the human ear and hence produces similar results. Confusion Matrix and Validation curves were used to evaluate the learning efficiency and prediction accuracy of the model. The accuracy of the Dense Neural Network classification model was found to be very high for a dataset consisting of samples for 30 different speakers.

REFERENCES

[1] Chen, Z., Wang, S., Qian, Y., & Yu, K. (2020). Channel invariant speaker embedding learning with joint multi-task and adversarial training. In ICASSP 2020 – 2020 IEEE international conference on acoustics, speech and signal processing (pp.6574–6578).

[2] Rami S. Alkhawaldeh, DGR: Gender Recognition of Human Speech Using One-Dimensional Conventional Neural Network (2019)

[3] Kory Becker, Identifying the Gender of a Voice using Machine Learning (2016)

[4] Jonathan Balaban, Deep Learning Tips and Tricks (2018)

[5] Faizan Shaikh, Getting Started with Audio Data Analysis using Deep Learning (with case study) (2017)

[6] Mike Smales, Sound Classification using Deep Learning, (2019)

[7] Aaqib Saeed, Urban Sound Classification, Part 1, (2016)

[8] Marc Palet Gual, Voice gender identification using deep neural networks running on FPGA, (2016)

[9] Kamil Ciemniewski, Speech Recognition from scratch using Dilated Convolutions and CTC in TensorFlow, (2019)

[10] Admond Lee, How To Build A Speech Recognition Bot With Python (2019)

[11] Adrian Yijie Xu, Urban Sound Classification using Convolutional Neural Networks with Keras: Theory and Implementation, (2019)