SPEAKER RECOGNITION BY NEURAL NETWORKS

ARTIFICIAL INTELLIGENCE

Manoj Kumar
Computer Science Engineering
IIT Tirupati
tcs15b013@iittp.ac.in

Pawan Kalyan
Computer Science Engineering
IIT Tirupati
tcs15b017@iittp.ac.in

Jeevan J
Electrical Engineering
IIT Tirupati
tee15b010@iittp.ac.in

ABSTRACT

We develop a Neural Network (NN) for classification or recognition of speakers. The Mel Frequency Cepstral Coefficients (MFCC) of speech signal are the input features of Neural Network. We sample each speech signal into small signals of known window size and then we compute the MFCCs for each window. Neural Network predicts one speaker for each window based on MFCCs of that particular window and Neural Network weights. Based on voting, that is, speaker who is predicted most number of times by Neural Networks; we finally decide or predict the speaker of that speech signal. We also compare Neural Network performance over Decision tree performance.

We find that our neural network model works 100% accurately for TIMIT corpus dataset with 16 speakers from eight dialects, whereas decision tree with depth twenty performs with an accuracy of 100%.

1. INTRODUCTION

Speaker Recognition is the process of automatically recognizing who is speaking by using speaker specific information included in the speech waves to verify identities being claimed by people accessing systems; that is, it enables access control of various services by voice. Speaker identity is correlated with physiological and behavioural characteristics if the speech production system of the individual speaker. These characteristics derive from both the spectral envelope (Vocal tract characteristics) and the supra-segmental features (Voice source characteristics) of speech. The most commonly used short-term spectral measurements are cepstral coefficients and their regression coefficients. As for the regression coefficients; that is derivatives of time functions of cepstral coefficients, are extracted at every frame period to represent spectral dynamics.

Speaker recognition can be classified into speaker identification and speaker verification. Speaker identification is the process of determining from which of the registered speakers a given utterance comes. Speaker verification is the process of accepting or rejecting the identity claimed by a speaker.

In the speaker identification task, a speech utterance of unknown speaker is analysed and compared with speech models of known speakers. The unknown speaker is identified as the speaker whose model best matches the input utterance. In speaker verification, an unknown speaker claims an identity, and an utterance of this unknown speaker is compared with a model for the speaker whose identity is being claimed. If the match is good enough, that is, above a threshold, the identity is accepted.

Speaker recognition methods can also be divided into text-dependent and text-independent methods. The former require the speaker to provide utterances of key words or sentences, the same text being used for both training and recognition, whereas latter do not rely on a specific text being spoken[1].

In this paper, we develop a Neural Network for text independent speaker identification. In text-independent speaker recognition, generally the words or sentences used in recognition trails cannot be predicted. It is impossible to model or match speech events

at the word or sentence level, some implanted methods are Long-Term-Statistics-Based methods, VQ-Based methods and Ergodic-HMM-Based method. Based on results for each window, we predict the speaker of speech sample by voting system.

2. SPEAKER IDENTIFICATION ANALYSIS

Speaker Identification consist of two main phases: Extraction of MFCC features of speech sample and Prediction of unknown speaker by Neural Networks. In this work, we extracted MFCC features of speech signal using normal MFCC algorithm and designed our own Neural Network for classification of speakers.

2.1 Extraction of MFCC Features:

In sound processing, the Mel-Frequency Cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on linear cosine transform of a log power spectrum on a non-linear mel scale of frequency.

Mel-Frequency Cepstral Coefficients (MFCCs) are coefficients that collectively make up an MFC. MFCCs are commonly derived as follows[5,6,7]:

1. Sample the speech signal into small frames (windows) of specific length
2. For each frame calculate the periodogram estimate of the power spectrum
3. Apply the mel filterbank to the power spectra, sum the energy in each filter
4. Take the logarithm of all filterbank energies
5. Take the Discrete Cosine Transform (DCT) of all log filterbank energies
6. Keep 2-13 DCT coefficients and discard the rest

We followed the same algorithm with a frame length of 25 milliseconds and 13 MFCC features for each frame; that is, given speech signal is framed into $\frac{Length\ of\ speech\ signal\ (in\ ms)}{25\ ms}$ frames. For each frame, we find 13 MFCC features and train them on a Neural Network[2].

2.2 Neural Network:

We develop a neural network with 13 input neurons, 26 neurons in 1st hidden layer, (four*no. of output neurons) in second hidden layer, (16*no. of output neurons) in third hidden layer and output neurons as number of speakers trained.

A sample neural network with 13 input features and 16 speakers is as follows:

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 26)	364
activation_1 (Activation)	(None, 26)	0
dense_2 (Dense)	(None, 64)	1728
activation_2 (Activation)	(None, 64)	0
dense_3 (Dense)	(None, 256)	16640
activation_3 (Activation)	(None, 256)	0
dense_4 (Dense)	(None, 16)	4112
activation_4 (Activation)	(None, 16)	0

Total params: 22,844
Trainable params: 22,844
Non-trainable params: 0

Figure 1: Typical Neural Network Architecture

We follow the above Neural Network architecture because our output neurons increases as number of speakers increases and we want our Neural Network parameters also to change because more number of speakers require more non-linear decision boundary, more non-linear decision boundary require more number of parameters.

We used TIMIT corpus dataset that is available on Kaggle. Dataset consists 16 speakers from eight different dialects, each speaker spoke a total of 10 different sentences, words, and phonetics. We divided the dataset such that every speaker's eight audio files are used for

training Neural Network and other two audio files for testing[3].

We represent 13-Dimensional MFCC features of each frame of each speaker of same speech signal; that is same sentence spoken by every speaker, to 2-Dimensions using multi-dimensional scaling[4]. We get a 2D plot of this data as follows:

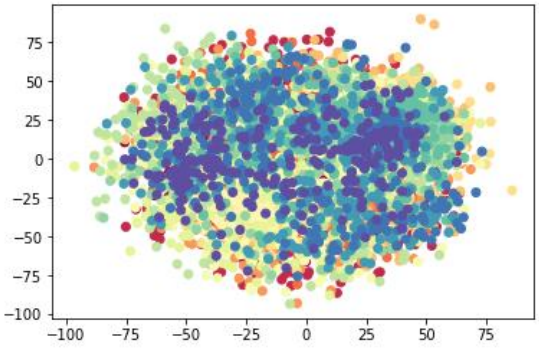

Figure 2: Plot of MFCC features on 2-Dimensions

We can see that most of the data points overlap with each other but mean and variance of data of each speaker is different. If we take only mean and variance of this data, we may not predict the speaker accurately because there is a chance that other speaker may also have same mean and variance.

In our work, we take each signal frame MFCCs and predict a speaker for each frame and finally we predict the speaker of that speech signal through voting; that is speaker with most outcomes in all frames. Here our Neural Network indirectly takes mean and variance as input but also considers deviations from every single data point.

We develop a Decision tree[8] with depth 5, 10, 15 and 20 to recognise the unknown speaker of given speech signal. Decision tree is easy to understand and it works purely based on logics developed in training, whereas Neural Network is very complex to understand.

3. RESULTS

First we look at the results produced by Neural Network. This dynamic Neural Network is working 100% accurately for TIMIT corpus dataset from Kaggle. This was tested against training data as well.

Now, if we look at Decision tree with depth 5, it predicts the speaker with an accuracy of 15.625% whereas with depth 10, it predicts with an accuracy of 56.25% and with depth 15, it predicts with an accuracy of 96.875%, finally with depth 20, it predicts with an accuracy of 100%.

Depth of the Decision Tree	Accuracy (in %)
5	15.625
10	56.25
15	96.875
20	100

Table 1: Accuracy for different depths of Decision tree

We can observe that as the depth of Decision tree increases accuracy also increases and it matched with our Neural Network accuracy.

If the number of speakers are increased, then our parameters in Neural Network also increases to achieve 100% accuracy. Similarly, Depth of Decision tree also should be increased to achieve accurate results. Now, we can say that Decision trees with larger depth acts as accurately as Neural Networks.

4. Conclusion

In our work, we showed that how Neural Networks can achieve better accuracy in Speaker Identification problem and how Decision trees produce accurate classification or prediction like Neural Networks. Decision tree are easy to understand when compared to Neural Networks which are very complex in terms of parameters and computation. Finally we can say that Decision tree with sufficient depth or dynamic Neural Networks can accurately predict the unknown speaker of given speech signal.

References

[1] Sadaoki Furui (2008) Speaker recognition. Scholarpedia, 3(4):3715., revision #64889.

[2] Wikipedia (2018) Mel-Frequency Cepstrum. Wikipedia revision #8th April, 2018.

[3] Kaggle (2017) TIMIT Corpus Sample. Kaggle-NLTK_data

[4] Scikit-Learn Multi-Dimensional Scaling. Scikit-Learn Website

[5] Davis, S. Mermelstein, P. (1980) Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, pp. 357-366

[6] X. Huang, A. Acero, and H. Hon. *Spoken Language Processing: A guide to theory, algorithm, and system development*. Prentice Hall, 2001

[7] Practical Cryptography Mel-Frequency Cepstral Coefficients Tutorial. Practical Cryptography Website

[8] Scikit-Learn Decision Tree. Scikit-Learn Website