Speaker Recognition

Abijith Pradeep, Anirudh Bhaskar,Dasari Karthik,Manuru Sai Suhas

August 26, 2015 ***Abstract*—In this report we investigate the ability of Mel Spectrograms combined with DNN models for identification in modern speaker recognition systems. We first show the spectrogram and the mel spectrogram in comparison to conventional CNN/DNN models and how these are better and later we provide proof as to why we need to use the spectrogram as separate layers in the neural network rather than use a librosa function to generate the spectrogram. In the end we extract the features from the Deep model and show a comparison between DNN direct classification and featured extracted classification.**

I. Introduction

Speaker Recognition is the problem of identifying a speaker from a recording of their speech. It is an important topic in Speech Signal Processing and has a variety of applications. Voice is one metric which apart from being natural to the users, provides comparable and sometimes even higher levels of security when compared to some traditional biometric approaches. The main principle behind speaker recognition is extraction of features from speech which are characteristic to a speaker, followed by training on a data set and testing. This study aims to evaluate different pre-processing, feature extraction, and machine learning techniques on audios recorded works well for speaker recognition and classification. Now our question is can we improve the speaker recognition system by using:

- A Mel Spectrogram
- A Deep Learning Network
- The MFCC from the speech directly
- or the features from the Deep Learning Network which are further fed into the Machine Learning Models.

In our project we are going to experiment and play with all those above elements for the best comparison of results and analyse which method can best improve modern speaker recognition systems.

II. Literature Survey

Joseph P Cambell designed a tutorial for automatic speaker recognition via LSP frequency features and achieved an accuracy of 98.9% for speaker identification on high quality telephone bandwidth speech collected in real world office environments. Joon Son Chung et al. implemented a deep CNN based speaker embedding system to map voice spectrograms to a Euclidean Space. This particular implementation provided a 4.42% EER (Equal Error Rate) on the ResNet50 Model on the VoxCeleb1 dataset for speaker verification. Douglas A Reynolds implemented a GMM (Gaussian Mixture Model) which achieved a 99.5% on a 630 speaker database which included TIMIT, NTIMIT, Switchboard, and YOHO. Although the accuracy dropped to 60.7% in NTIMIT database after the addition of telephone line degradation the verification task gave only 0.24% EER.. An overall EER of 0.51% and a falserejection rate of 0.65% at a 0.1% false-acceptance rate were obtained. Jose A Lopez et al. implemented a speaker recognition system for anomaly detection for different machine sounds. The model achieved close to perfect results from 95% to 99% for five out of six machine sounds.

III. Dataset

The dataset is called 16000pcm-speechs which contains 5 different international leaders' speech utterances. Each speaker has about 1500 samples.

Speaker	sample size
Benjamin Netanyau	1500
Jens Stoltenberg	1500
Julia Gillard	1500
Magaret Thatcher	1500
Nelson Mandela	1500

TABLE I
Speakers and Sample size

IV. Methodology

After we load our dataset we convert each of our audio files into a spectrogram. We compute the spectrogram by windowing the audio file and taking overlapping FFT's of those windowed frames to create a matrix which create an image. Using a mel scale we convert the created spectrogram's to mel-spectrogram. Now we extract the features from both the sets of spectrograms using a Neural Network. The first CNN model consists of the spectrogram being the first input layer and has a single convolution layer assisted with a maxpooling layer and we flatten it and use two dense layers to get our final class predictions. For the second CNN model we use the VGG architecture and put the input layer as the spectrogram and modify the last dense layer as 5 for class prediction. Now we try to extract MFCC coefficients from the mel-spectrogram and use those coefficeints as features and see how well the model performs using the same benchmarks. In order to compare the results of the basic CNN model with the VGG model we use six machine learning classifiers:

- Logistic Regression
- Random Forest
- Decision Tree
- KNN
- Gaussian Naive Bayes
- SVM

V. Results

Model	Accuracy	Precision	Recall	F1
Basic CNN	0.92	0.93	0.92	0.93
LR	0.92	0.92	0.92	0.92
RF	0.93	0.93	0.93	0.93
DT	0.91	0.91	0.91	0.91
KNN	0.95	0.95	0.95	0.95
GNB	0.68	0.68	0.68	0.68
SVM	0.94	0.94	0.94	0.94

TABLE II
Results For Spectogram

Model	Accuracy	Precision	Recall	F1
Basic CNN	0.89	0.9	0.89	0.89
LR	0.92	0.92	0.92	0.92
RF	0.9	0.9	0.9	0.9
DT	0.87	0.87	0.87	0.87
KNN	0.92	0.92	0.92	0.92
GNB	0.83	0.84	0.83	0.83
SVM	0.91	0.91	0.91	0.91

TABLE III
Results For Mel Spectogram

Model	Accuracy	Precision	Recall	F1
VGG	0.92	0.93	0.92	0.92
LR	0.95	0.95	0.95	0.95
RF	0.95	0.95	0.95	0.95
DT	0.93	0.93	0.93	0.93
KNN	0.95	0.95	0.95	0.95
GNB	0.93	0.93	0.93	0.92
SVM	0.95	0.95	0.95	0.95

TABLE IV
Results For Spectogram

Model	Accuracy	Precision	Recall	F1
VGG	0.95	0.95	0.95	0.95
LR	0.97	0.97	0.97	0.97
RF	0.97	0.97	0.97	0.97
DT	0.95	0.95	0.95	0.95
KNN	0.97	0.97	0.97	0.97
GNB	0.96	0.96	0.96	0.96
SVM	0.97	0.97	0.97	0.97

TABLE V
Results For Mel Spectogram

Model	Accuracy	Precision	Recall	F1
LR	0.198	0.198	0.198	0.197
RF	0.204	0.208	0.204	0.203
DT	0.192	0.193	0.192	0.192
KNN	0.195	0.196	0.195	0.192
GNB	0.197	0.159	0.197	0.173

TABLE VI
Results For MFCC from Mel Spectogram

VI. Conclusion

Upon observing the performance of how well prediction happens for speaker recognition for the 5 audio datasets we can see that the mel-spectrogram performs overall we compared to normal spectrogram but the performance difference is not large. The performance using mel-coefficients is very bad because the number of coefficients were too high. This resulted in accuracy being only 19-20

REFERENCES

[1] Joseph P Campbell. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, 1997.

[2] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[3] Douglas Alan Reynolds. *A Gaussian mixture modeling approach to text-independent speaker identification*. PhD thesis, Georgia Institute of Technology, 1992.

[4] Jose A Lopez, Hong Lu, Paulo Lopez-Meyer, Lama Nachman, Georg Stemmer, and Jonathan Huang. A speaker recognition approach to anomaly detection. In *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pages 96–99, 2020.

[1] [2] [3] [4]