Speaker Identification: Text Independent

Rish Gupta, Manish Pandit and Sophia Zheng
{rishg, manish7, xszheng}@stanford.edu

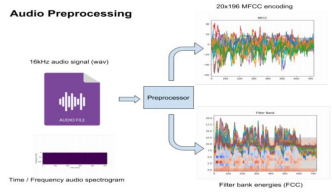
Introduction

Goal: Identify the speaker from an unknown audio signal.

Find the highest probability speaker matching an audio signal from a repository of known speakers.

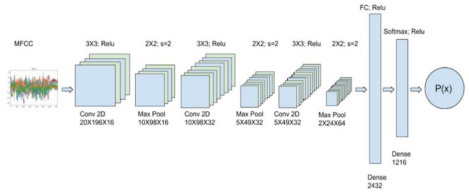
Data

- Large public dataset with **65,000+** audio samples (VoxForge).
- **1216 unique speakers**
- Preprocessed to Mel-frequency cepstral coefficients (**MFCCs**).
- **90% - 5% - 5%** Train – Dev – Test Split



Model Architecture

- Convolutional Neural Network (CNN)
- Activations: ReLu + SoftMax
- Optimizer: Adam
- L2 Regularization + Dropout
- Grid Search for optimal hyperparameters
- **10M+** Trainable parameters



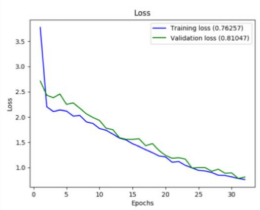
Results

Model Performance

`	Training	Dev	Test
Accuracy	**98%**	**97%**	**97%**
Loss	**0.76**	**0.81**	**0.81**

Model & Hyperparameters

`Architecture	Mini Batch	Optimizer	Activations
CNN + FC	**32**	**Adam**	**ReLu + SoftMax**



Discussions

- The model predicts the speaker with remarkable accuracy.
- The prediction is quick and done in constant time. O(1).
- The prediction errors have high correlation with high noise levels.
- L2 Regularization helped reduce variance.

Future Work and Citations

- Study the impact of additional recording sources. i.e. video, open space, conferences.
- Test model against additional datasets and analyze the impact on performance.
- Algorithm applied to streaming audio signals.

[1] voxforge: http://www.Voxforge.Org/
[2] GMM: https://en.Wikipedia.Org/wiki/mixture_model
[3] MFCC: https://en.Wikipedia.Org/wiki/mel-frequency_cepstrum
[4] speaker recognition: https://en.Wikipedia.Org/wiki/speaker_recognition
[5] Baidu, inc. Deep speaker: an end-to-end neural speaker embedding system
[6] An overview of text-independent speaker recognition: from features to supervectors