Speaker Identification in a Noisy Environment
Raw Waveforms vs MFCC

David Grogan
dgrogan@stanford.edu

Phathaphol (Peter) Karnchanapimonkul
phatk@stanford.edu

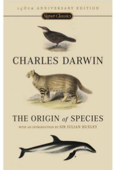
Stanford University

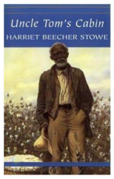
Introduction

- We built a Speaker Identification network. Input is a voice sample, output is who's talking.
- Our motivation came from Alzheimer's patients who forget or can't identify who is speaking in group calls, e.g. family calls to Grandma from one speaker phone.
- The solution extends to voice conference calls, as seen on right.
- Many existing solutions perform Speaker Identification only on clean audio. We use noisy audio to better simulate real-world conditions.
- Our network is trained on 20 speakers, as that's roughly the max size of a family or work team.

Dataset

20 audiobooks from LibriVox

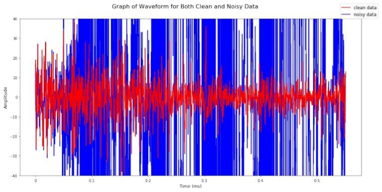

150th ANNIVERSARY EDITION
Signet Classics
CHARLES DARWIN
THE ORIGIN of SPECIES
WITH AN INTRODUCTION BY SIR JULIAN HUXLEY



- LibriVox and Internet Archive host royalty-free audiobooks read by volunteers.
- Mono-channel, 22050Hz mp3s; one file per chapter.
- We split into 0.5 second training examples.. 0.5 seconds seemed to be the shortest time a human needs to identify the speaker.
- Complication: Some audio books have multiple speakers. Some speakers read multiple audio books.

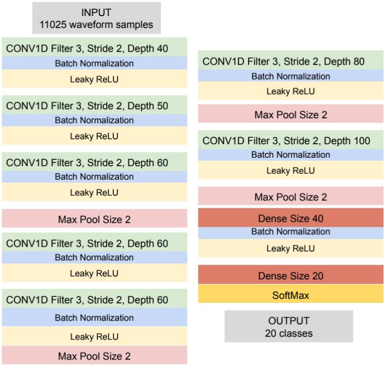
Augmenting with Noise

- Audiobooks are clean recordings. We want background noise.
- 3 Background Noises: 1) Crowd Talking, 2) Laptop Keyboard, 3) Plastic Crumple
- Overlay ENTIRE audiobook with noise in 20-second chunks. Normalize volume to match.



Algorithms and Models

Final Architecture



INPUT
11025 waveform samples

CONV1D Filter 3, Stride 2, Depth 40
Batch Normalization
Leaky ReLU

CONV1D Filter 3, Stride 2, Depth 50
Batch Normalization
Leaky ReLU

CONV1D Filter 3, Stride 2, Depth 60
Batch Normalization
Leaky ReLU

Max Pool Size 2

CONV1D Filter 3, Stride 2, Depth 60
Batch Normalization
Leaky ReLU

CONV1D Filter 3, Stride 2, Depth 60
Batch Normalization
Leaky ReLU
Max Pool Size 2

CONV1D Filter 3, Stride 2, Depth 80
Batch Normalization
Leaky ReLU

Max Pool Size 2

CONV1D Filter 3, Stride 2, Depth 100
Batch Normalization
Leaky ReLU

Max Pool Size 2

Dense Size 40
Batch Normalization
Leaky ReLU

Dense Size 20
SoftMax

OUTPUT
20 classes

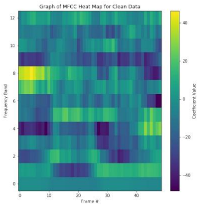
Input comprises 11025 waveform samples which represents 0.5 seconds of data.

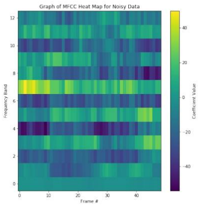
We used 7 convolutional layers and 4 max-pooling layers all of stride 2 to reduce the input from size 11025 to size 5 before flattening the channels to pass to the dense layers.

Each of the output classes corresponds to one speaker.

Experiments

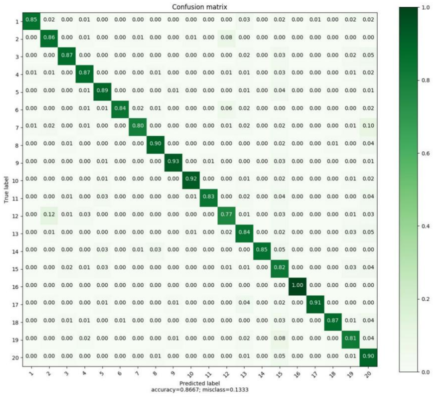
- Model trained on clean data achieves 94% test accuracy on clean data.
- Clean model achieves a paltry 14% accuracy when making predictions on noisy data.
- Significant effort training models with MFCCs, which are generated from the raw waveform via substantial signal preprocessing. Traditionally, most Speech Recognition and Speaker Identification use MFCCs as input.
- The MFCC models achieved 88% accuracy on clean data, but only 63% on noisy data.
- TA advised that the preprocessing could be discarding valuable signal, so we abandoned MFCCs and trained a bigger network using raw waveforms as input.

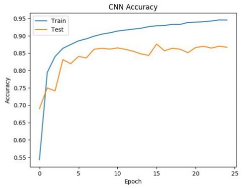


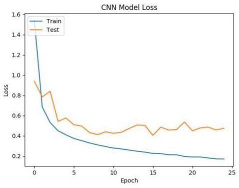


Results & Analysis

- Final model achieved 86% accuracy on the test set
- Accuracy was NOT uniform across speakers







References

Sainath, Tara N., et al. "Learning the speech front-end with raw waveform CLDNNs." Sixteenth Annual Conference of the International Speech Communication Association. 2015.

Méndez, Alfredo, "Speaker Identification with Deep Neural Networks," Stanford CS230 Past Projects – Winter 2019.

McLaren, Mitchell, and Yun Lei. "Improved speaker recognition using DCT coefficients as features." 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015.

Torfi, Amirsina, Jeremy Dawson, and Nasser M. Nasrabadi. "Text-independent speaker verification using 3d convolutional neural networks." 2018 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2018.

"Crowd Talking 8." MP3 file. http://www.soundjay.com/beep-sounds-1.html.

"Computer Keyboard 1." WAV file. https://www.soundjay.com/communication-sounds.html.

"Plastic Crumble 1." MP3 file. https://www.soundjay.com/misc-sounds-2.html.