Speaker Identification in a Noisy Environment
Raw Waveforms vs MFCC

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 to
  group calls, e.g., family calls to Grandmas from one speaker phone.
- The solution sounds to voice conference calls, as well as on right.
- Many existing solutions perform speaker identification only on clean audios. We use noisy audios to
  better simulate real-world conditions.
- Our network is trained on 20 speakers, so that’s roughly the max size of a family or work team.

Dataset
- 20 audiobooks from LibriVox: 20-second chunks, 22050Hz, each.
- Monaural, 22050Hz, single speaker.
- We split into 5.5-second training examples, 0.5 seconds used to be
  the shortest time a human needs to identify the speaker.
- Augmentation: Some audio books have multiple speakers. Some
  speakers read multiple audio books.

Augmenting with Noise
- Audio books 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

Experiments
- Model trained on clean data achieves 94% test accuracy on clean data.
- Clean model achieves a poly 90% accuracy when making predictions on noisy data.
- Significant effort training models with MFCCs, which are generated from the raw
  waveforms 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 60% on noisy data.
- TA advised that the preprocessing could be discarding valuable signal, so we abandoned
  MFCCs and trained a larger network using raw waveforms as inputs.

Results & Analysis
- Final model achieved 88% accuracy on the test set.
- Accuracy was NOT uniform across speakers.