Project summary

Speaker Identification

Katrina Bykova

Introduction

Speaker recognition is used by digital devices and systems to label a user with an identity tag. The user identity information can then be used for multiple applications. Unlocking a mobile device is a common use of the user identity. User authentication by a personal assistant can be used as a key to unlock user's preferences or grant access to making purchases or changing system's settings. Phone customer services is another area where user authentication can be used. It can be utilized as an additional security protocol or as a way to improve user experience by speeding up the authentication process during a customer service session.

Goal

Develop a system for speaker voice recognition.

Data

LibriSpeech from OpenSLR. Corpus (1,000 hours) of read English speech from 1,000 speakers.

Each speaker was represented by 10 audio files that were 3 seconds long.

Audio processing

Audio files were converted into spectrograms with Librosa and rescaled for for the subsequent modeling step. Specifically, Short-time Fourier transform was used to transform the time domain data to the frequency domain. Specifically, the hamming window of 25 ms and the step of 10 ms were used. The data were then scaled and reshaped to be used as an input for VGG16 model (224, 224, 3). Since the data only has the amplitude for the channel dimensions, it was repeated three times over the channel dimensions.

Feature extraction

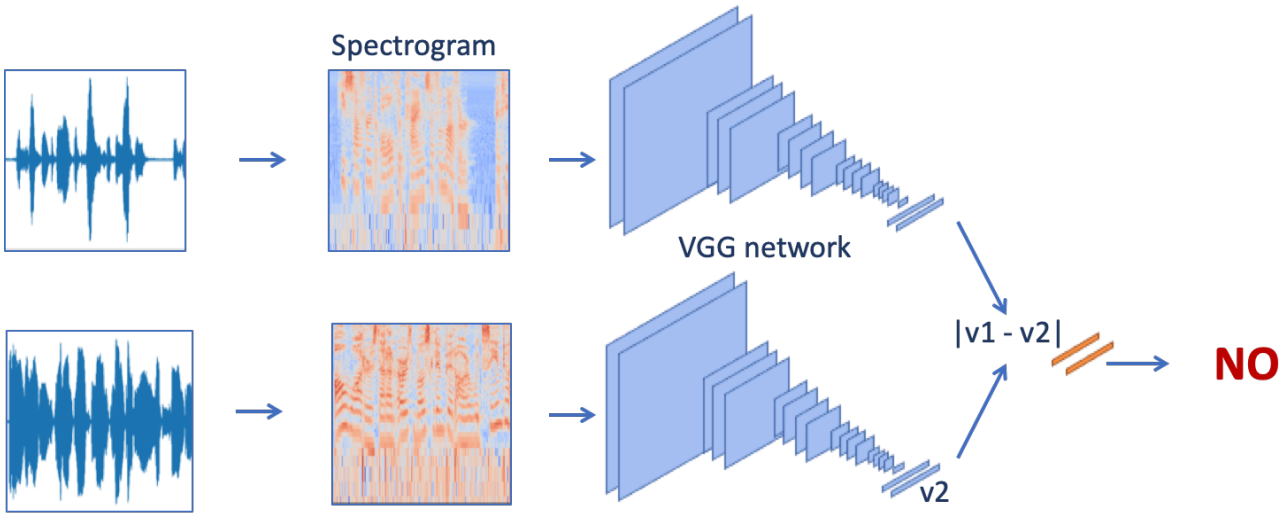
VGG16, a Convolutional Neural Network, trained as an image classifier was used for extracting features from the spectrograms (in a transfer learning manner). The top layers were removed from the networks and a new layer was trained with the speaker data.

MobileNetV2 was tested as another option for the feature generation with the ImageNet weights (transfer learning) and from scratch.

VGG16 with the ImageNet weights gave the best performance. Intuitively, the MobileNetV2 was expected to perform better. Potential reason for that is the limited number of hyperparameters that was explored.

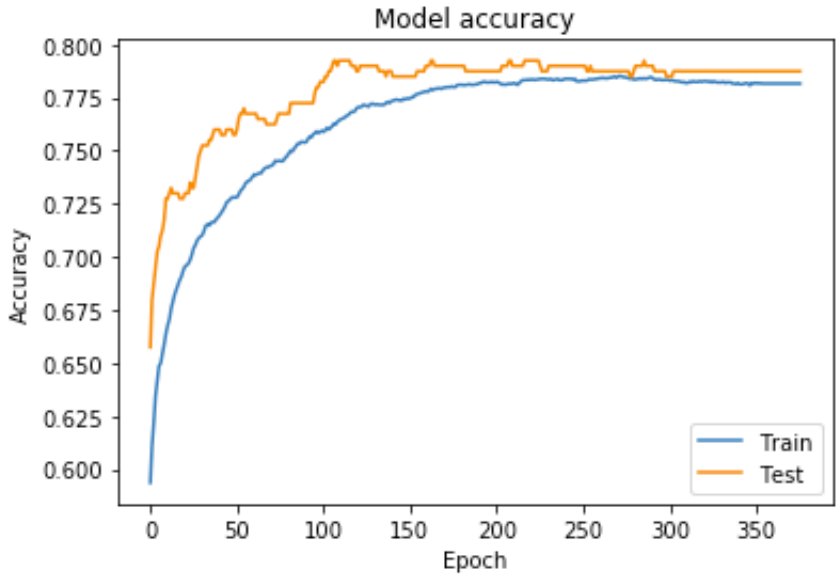
Model - Siamese Networks

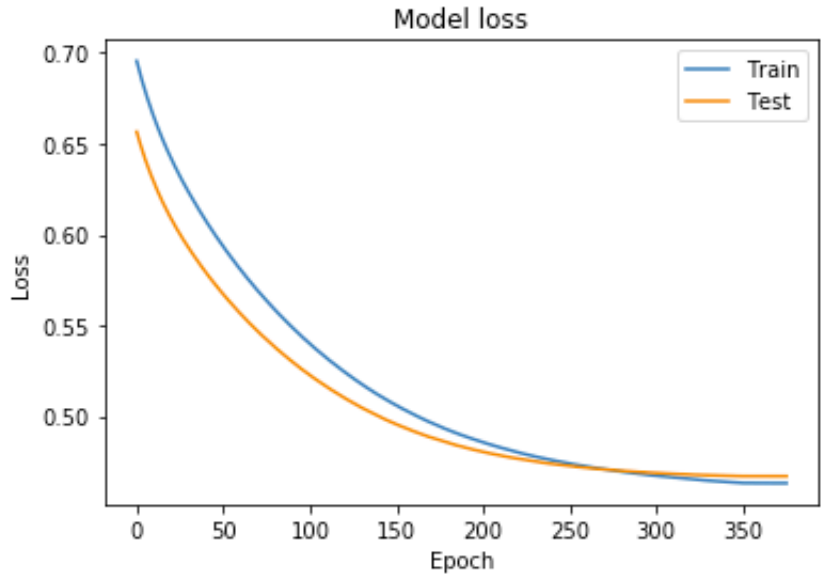
The output from the VGG16 net was used to build a Siamese network. The absolute difference between the two inputs was used for classifying the speaker pairs into either the same or different speaker categories.



Speaker recognition algorithm

For a speaker recognition algorithm, a voiceprint was built from 5 audio recordings that belong to a user speaker. A logistic regression was trained to classify the speakers. The accuracy score for the 5-audio voiceprint algorithm was 0.83.





Future work

Error analysis showed that speakers who were poorly classified shared a common feature of changing their voices to represent different book characters.

Training new models on a dataset containing audios from speaker who change their voices might improve the model performance.