Features that distinguish languages
Insights from deep neural nets
Nicholas Monto

Background

Convolutional neural networks (CNNs) are classically used to classify images into pre-determined categories.\(^1\), \(^2\), \(^3\) CNNs have shown to be very capable at classifying speech.\(^1\), \(^2\)

The question is: What about the language does each convolutional layer use to aid in accurate classification?

Model Architecture

Convolutional Layer 2

Convolutional Layer 3

Convolutional Layer 4

Feature representation

Audio files came from the Shtooka Project (an online free-to-use multilingual audio database).

The audio files were first standardized to 16 kHZ sampling rate and 16 bit spectral density wav files.

500 audio files were randomly selected from each language and split into training and validation sets.

For each set, the audio files were converted to spectrograms that included spectral information up to 5.5 kHZ.

The final set contained 7200 training files (300 x 3 x 8) and 1600 (200 x 8) validation files.

Future Direction

Expand the dataset to include more talkers from each language.

Study consistency of class activation maps across the different languages.

Identify specific speech cues used within and across language families.

References