Application of Convolutional Neural Networks in Accent Identification

Keven Chionh, Raymond Song & Yue Yin
Carnegie Mellon University

Contact Information:
Keven Chionh
Email: kchionh@andrew.cmu.edu
Raymond Song
Email: maoyuans@andrew.cmu.edu
Yue Yin
Email: yyin1@andrew.cmu.edu

Carnegie Mellon University

Abstract

The English language is spoken by many around the world, with different accents originating from various parts of the world. Accent identification is an important problem in technology today. Unfortunately, accent identification is suboptimal for certain accents. The purpose of this project is to improve the accuracy of accent identification in the English language for the Arabic, Italian, Japanese, and Korean accents.

The training data was taken from the Center for Spoken Language Understanding's Foreign Accented English v1.2 dataset. This data was used to train a convolutional neural network (CNN). In this project, we observed how the accuracy varied with the number of convolutional layers. We found that the 2-layer CNN performed best, with a training accuracy of 0.814 and a test accuracy of 0.779.

Introduction

In this day and age, English is widely used for global communication. However, people do not speak English the same way. English accents are characterized as a distinctive mode of pronunciation in spoken English. While native English speakers have different dialects, accents are more observable in non-native English speakers due to influences of other languages. As the population of English speakers grows, the 'language-gap' is a diminishing problem. However, the 'accent-gap' becomes a more pertinent issue.

'If you talk to a man in a language he understands, that goes to his head. If you talk to him in his language, that goes to his heart.'
- Nelson Mandela

Accents are categorized according to their geographical origin. Misclassified accents can bring frustration to users of applications that utilize voice recognition technology. The goal of this project is to improve the accuracy of English accent identification.

Main Objectives

1. Train a model to differentiate between the Arabic, Korean, Italian, and Japanese accents in spoken English.
2. Achieve an accuracy of at least 75%. This figure was chosen to be competitive with current standards found in our literature review.

Methodology

An overview of our methodology is as follows:

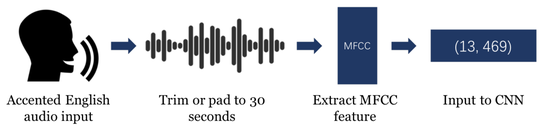

Figure 1. *Overview of methodology.*

The audio dataset was taken from the Center for Spoken Language Understanding (CSLU)'s Foreign Accented English v1.2 dataset. It consisted of continuous telephone-quality speech in English by native speakers of twenty-two different languages. We focused on accents with the worst accuracy rates: Arabic, Korean, Italian, and Japanese.

The audio files were all padded (via duplication) or trimmed to the same length of 30 seconds for consistency. They were then processed using the Librosa library into their mel-frequency cepstral coefficients (MFCCs), which is a representation of the short-term spectrum of sounds. We only took the lower 13 dimensions of the MFCC as they are the most relevant to the human voice range. Therefore we had a size 13 vector for each sound frame. Therefore each 30 second audio file was represented by a 13×469 array.

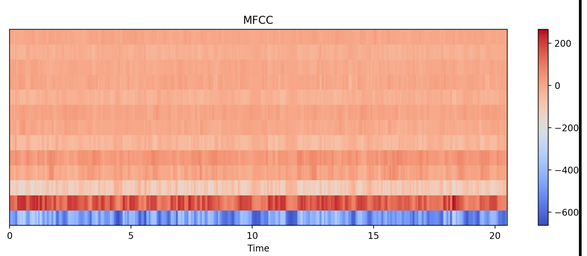

Figure 2. *Example of a MFCC array on an audio clip.*

These MFCC arrays were then used to train a convolutional neural network (CNN). We split the dataset randomly into 2 sets: 85% was put into a training set and 15% into the test set. However, we maintained the same ratio for each language within the test and training sets.

We experimented with different numbers of convolutional/pooling layers. At the same time, a dropout rate of 0.25 was used to prevent overfitting.

Results

Eventually, the best model was a CNN constructed with 2 convolutional layers and 2 pooling layers as follows:

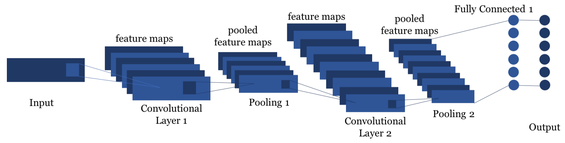

Figure 3. *Convolutional neural network trained with MFCC arrays.*

We found that in general, a 3-layer CNN did not yield much improvement over a 2-layer CNN. The accuracies are plotted in the following figure:

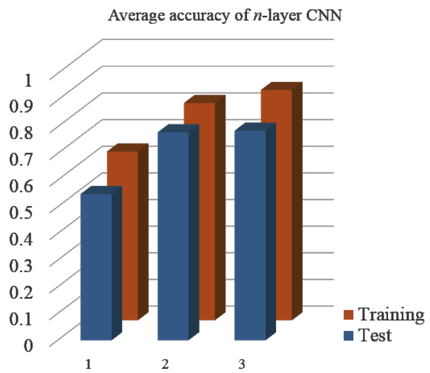

Figure 4. *The average accuracy across the 4 accents was plotted against the number of layers in the CNN.*

There was some noticeable difference in the average test accuracy between the 2 and 3-layer CNN. The 3-layer CNN had a training accuracy of 0.863 whereas that of the 2-layer CNN was 0.814. However, the difference in the average test accuracy was merely 0.785 (3-layer) and 0.779 (2-layer). In light of this insignificant difference in training accuracy, as well as increased divergence between test and training accuracy. Our findings indicated that the 2-layer CNN struck a better balance between accuracy and overfitting. We therefore opted to use the simpler model with slightly lower accuracy. However, the 1-layer CNN performed extremely poorly, with a test accuracy of only 0.548.

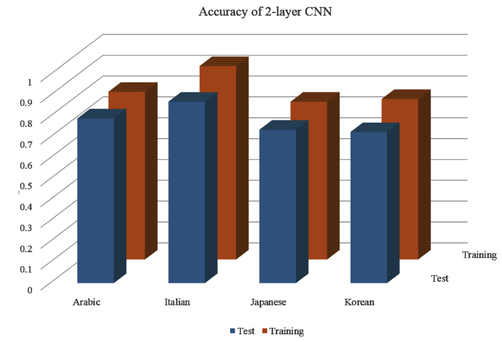

Figure 5. *Training and test accuracies categorized by accents.*

The model performed the best on the Italian accent, with a test accuracy of 0.869. On the other hand, it performed the worst on the Korean accent, with a test accuracy of 0.724. This was consistent with our literature review, which indicated that the Korean accent usually performed poorly.

Conclusions

- The 2-layer CNN struck the best balance between accuracy and model complexity.
- The Italian accent performed around 5% better on average than what we found in our literature review.
- The improvements in the Arabic and Japanese accents were less significant, and may have been due to our choice of dataset.
- The Korean accent still performed poorly, consistent with current findings.

Future extensions

- A challenging extension would be to train a model to recognize accents within closer geographical locations. For example, English spoken within Britain alone can be broken into London, Liverpool, Cockney, etc. accents.
- Given that Chinese has increased in popularity as a second language, it would be interesting to train a model to classify various accents of spoken Chinese.

References

[1] Faria A. (2006) Accent Classification for Speech Recognition. In: Renals S., Bengio S. (eds) Machine Learning for Multimodal Interaction. MLMI 2005. Lecture Notes in Computer Science, vol 3869. Springer, Berlin, Heidelberg.

[2] Teixeira, Carlos Trancoso, Isabel Serralheiro, Antnio. (1996). Accent identification. The 4th International Conference on Spoken Language Processing.