John Pyjar A16123651
Aryaman Sinha A15828821
Bridget Egan A15769204
Dong Pham A15911940
Alex Tyler A16542815
Donovan Drews A15583233

accent classification
based on formant extraction
using PCA, K-Means, and GHMMs

Accent Classification based on Formant Extraction using PCA, K-Means, and HMMs

Introduction

Automatic Speech Recognition (ASR) has become a ubiquitous feature in practically all word processors that facilitate typing. As not all users have the physical ability to use a keyboard, it is important that these ASR algorithms are effective, and not just for the specific dialect used by the software developers. Unfortunately, this is where these algorithms typically fall short–even systems as international as Alexa and Siri still see issues in consistently seeing high performance across diverse groups of users, as not every accent is equally accommodated for. Hence we want to attempt to design a machine learning model for accent classification as a sort of precursive step in improving the worldwide accessibility of ASR systems. We specifically look to investigate the efficacy of unsupervised methods for this process, namely principle component analysis (PCA), K-means clustering, and a mixture of Gaussians implemented on Hidden Markov Modeling (HMM).

Related Work

In recent years, there have been multiple different approaches researchers have taken to classify accents using machine learning. In Leonmak and Edmundmk's work, they use Random Forests and Gradient Boosting as baseline, while Multilayer Perceptron (MLP) and Convolution

Neural Networks (CNN) are used to see how accurate they can get the testing. From using these different methods, CNN performed the best with 88% accuracy. In Sun's work, Ensemble Algorithms are deployed to contrast other approaches for pitch and accent classification. This approach resulted in 84% to 87% accuracy. Finally, in Demarco and Cox's work multiple methods are used to achieve accuracy of 81% on British accents. These papers provided a backdrop to understand what potential methods would work well versus which would not.

Methods

Our data comes courtesy of George Mason University's Speech Accent Archive, a database of over 2,000 international English speakers all saying the same quote:

> *"Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob…"*

Since the dataset contained a large amount of files, we set up an Apache web server to host all of our files to make downloading and accessing easy for all members on Google Colab. After downloading and organizing the 2,000 mp3 files, we used the Parselmouth library, which interfaces with Praat, an audio analysis program, to extract the first, second, and third formants of the speech across the duration of the quote reading. A formant is a representation of the concentrated energy produced by the resonance of the human vocal tract, measured in Hz. There are over nine formants, but the first three are able to accurately represent speech, divorced from acoustic features like tone and timbre, so we felt that using them would be sufficient for categorization[1].

Our goal is to classify the accents of native English speakers, and so we removed non-native speakers from our dataset, as well as the outlying files of 'english145.mp3', 'english317.mp3', 'english315.mp3', and 'english450.mp3' due to errors in extracting the

[1] This ended up being a mistake for our model.

formants, resulting in unusually short data. Upon listening, it seems like there is too much noise in the background for Praat, the program Parselmouth calls, to accurately hear the speech. Regardless, they did not provide enough data, so we threw them out. We decided to only analyze the first 15 seconds of speech, which prevents errors at the end of the files. However, to account for the formant lengths being slightly different for each file, we decided to test two separate methods of data repair. The first method, which was called the 'interpolated data,' first resampled all the formant vectors to arrive at the same length. Then, each vector was cross-correlated with one of the samples, and the argmax was taken to adjust by the optimal time shift between the two vectors. Ideally, this would maximize structural similarity between the two formants. The second approach was to only use the amount of sampled formant points equal to the shortest list of formants, which we are calling the 'clipped data'. We found that the clipped data provided the best overall results in the later clustering methods, and therefore this is the dataset we used for the remainder of the analysis.

At this point, for each speaker, for both the interpolated and clipped data, we had three time-sequence arrays for each of the formants, which are at least of length 400. To sort the data, we decided to use a mixture of countries and regions. Based on Penn State data, we grouped speakers born in America into 4 accents: West, Midwest, North, and South. The rest of the accents were grouped per world region, based on the International Dialects of English Archives – it's slightly reductive, but linguistically sound. This resulted in 13 clusters. Upon further analysis of the data, we also found that separating the audio files based on the gender of the speaker was important, as womens' formants are generally higher than mens' by around 100 Hz (Pépiot).

We decided to use PCA before implementing the K-Means and GHMM functions. Given that our dataset was rather large, and each speaker had a lot of data attributed to their elicitation, we concatenated the three formants, then performed PCA to reduce the dimension from 1200 down to 100 principal components. Using the scree plot, we found that the two largest principal components explained the vast majority of the variance. For this reason we chose to use only these first two components for the K-Means and HMM methods. Then using these components, we performed a simple K-Means clustering of the reduced data, which seemed to have good results on the gender-differentiated American dataset. To attempt another solution, we decided to use a Hidden Markov Model, specifically a Gaussian Hidden Markov Model, which is an HMM modeled with Gaussian distribution of emissions. This is a commonly-used application of GHMMs, so we figured we'd try our hand. Instead of using the standard EM functions, our current GHMM uses a Viterbi algorithm. The Viterbi algorithm obtains a maximum a posteriori (MAP) estimate of the most likely sequence of hidden states from the given data.

Results

For our results, we decided to use a Chi-Squared test on each of the assigned clusters for the speakers in every dataset generated. The null hypothesis of a chi squared test is that an observed population is the product of the random sampling of an expected population. So each cluster was treated as an observed population, and an expected population was computed for each cluster using the proportions of regions in the total dataset. The results of these tests would tell us whether the clusters just reflected the total population, or if they actually created any tangible patterns in the grouping of regions. If the null is consistently rejected, we would begin actually investigating these patterns and see how the speakers were sorted based on their formants, but if not, we could stop accuracy testing here and conclude the algorithms are inaccurate, or at least inconsistent.

Unfortunately only 3 out of the 42 total tests actually produced a test statistic exceeding the critical value with a reliable p-value, allowing us to reject the null, but even these three clusters didn't exhibit any clear groupings of one or two regions. This allows us to conclude that none of our models successfully captured any meaningful differences in the regions of speakers according to their formants.

Discussion

There are several things that we can improve on our Gaussian Hidden Markov Model. First we can try to implement a proprietary emission matrix instead of using a normal distributed matrix making the HMM non Gaussian. For predicting and training we could have used the Baum-Welch algorithm, which is a special instance of the EM model instead of using MAP estimation from the Viterbi algorithm. The BW algorithm helps us in learning model parameters (transition matrix, the emissions probability, and the starting probability) with the given dataset, while the Viberti helps us with determining the hiddenstates with a given dataset and parameters that are assumed.

It is also worth noting that we did a fair amount of preprocessing to the data, in order to get the data into a shape that was convenient to test. However, removing some of the variation of the data perhaps smoothed over the differences that would have made the accent distinctions clearer. So it appears that neither our k-means algorithm nor GHMM model was able to accurately cluster our data, given the p-values for the assignments of our clusters. We expect this could be due to several reasons. The formant structures, while relevant, are not the only deciding factors of accents. And without more data, the little variations in formants that could be predictive of accent are too drowned out in the noise of normal human speech variance. Furthermore, prosody, word choice, and even pitch of speaking are important to accent classification. Our results, then, add to the wealth of data that proves that human speech is complicated and messy, and cannot be classified by one factor alone.

Contributions

Drews - PCA, cross-correlation

Egan - Linguistic knowledge, data wrangling, and emotional labor

Sinha - K-means clustering

Tyler - Server, Related Work, Video Editing

Pham - GHMM

Pyjar - Chi-squared accuracy testing

References

DeMarco, A., & Cox, S. J. (2013). Native Accent Classification via I-Vectors and Speaker Compensation Fusion. *INTERSPEECH*.

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*, 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

Mak An Sheng, L., & Mok Wei Xiong, E. (2017). *Deep Learning Approach to Accent Classification*. http://cs229.stanford.edu/proj2017/final-reports/5244230.pdf

Meier, C. (Ed.). (2022). *IDEA - International Dialects of English Archive*. https://www.dialectsarchive.com/

Pépoit, E. (2012). Voice, speech and gender:: Male-female acoustic differences and cross-language variation in English and French speakers. *Corela*. https://doi.org/10.4000/corela.3783

Xuejing, S. (2002). Pitch Accent Prediction Using Ensemble Machine Learning. *INTERSPEECH*.