Supervised Machine Learning Model for Accent Recognition in English Speech using Sequential MFCC Features

AIDE 2019

Dweepa Honnavalli
Shylaja S S

Presentation Outline

Today's Discussion

- What, and why?
- Background and related work
- How?
- The dataset
- About MFCC
- Feature vectors
- Results
- Discourse

Source: ipsoft.com

Accents- ubiquitous, but not quite.

The skewed representation of accents in technology leads to almost a billion people with accents that voice assistants cannot recognize.

What, and why?

Background and Related work

Tang et. al - 2003

Extracts features such as word-duration and word final stop disclosure duration.

Chu et. al - 2017

Comparative analysis on self produced 20 second audio clips.

Kat et. al - 1999

Fast accent classification using phenome models.

How?

Proposed methodology

Extract mfcc features

Generating feature vector

Handling under-represented data

Supervised learning

Prediction and evaluation

The Dataset

1032 / 2301

Indian Accented speech

1269 / 2301

American accented speech

About MFCC

What is it?

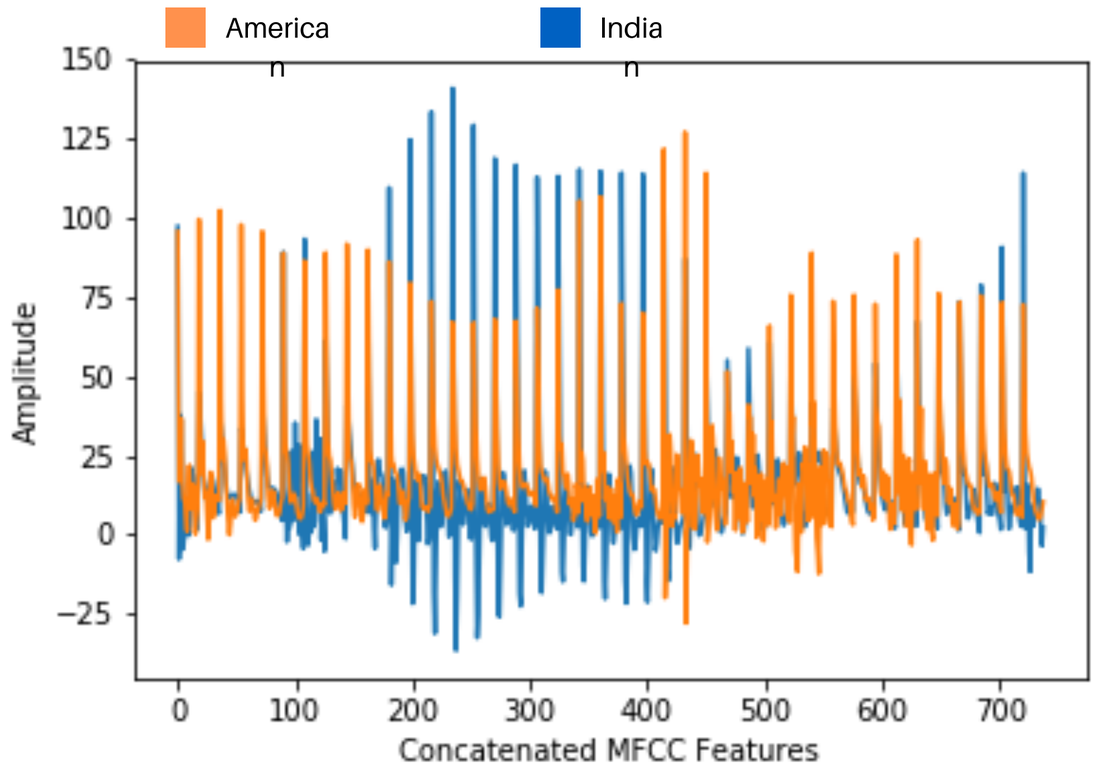
Mel frequency cepstrum is the short-term power spectrum of a sound.

Why do we use it?

The coefficients approximate the human auditory system’s response closely.

How do we use it?

A resulting set of 20 coefficients for each frame of the signal are sequentially positioned to retain context.



Feature Vectors

Extract MFCC

1

Choose Coefficient

2

Feature vector

3

Before

$$\begin{bmatrix} c_0 f_0 & c_0 f_1 & c_0 f_2 & \dots & c_0 f_m \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ c_{19} f_0 & c_{19} f_1 & c_{19} f_2 & \dots & c_{19} f_m \end{bmatrix}$$

After

$$\begin{bmatrix} c_0 f_0 & \dots & c_0 f_m & \dots & \dots & \dots & c_{19} f_0 & \dots & c_{19} f_m \end{bmatrix}$$

Supervised Machine learning

- *Neural networks*
- *Logistic regression*
- *K- Nearest Neighbors*
- *Support Vector Machines*
- *Gaussian Mixture Models*

Results and Discourse

Validation

Model	Mean Validation Score	Std Dev of Validation Score
Neural Networks	0.95	0.02
KNN	0.91	0.06
Logistic Regression	0.95	0.03
SVM	0.36	0.01
GMM	0.39	0.10

Test

Model	Precision	Recall	f-measure	Reject Rate	Accuracy
Neural Networks	0.96	0.94	0.95	0.97	0.95
KNN	0.9	0.92	0.91	0.9	0.91
Logistic Regression	0.94	0.96	0.95	0.95	0.95
SVM	1.0	0.02	0.04	1.0	0.54
GMM	0.43	1.0	0.60	0.0	0.43

The Big Three

Neural Networks
Logistic Regression
K-Nearest Neighbors

Logistic Regression
0.58 sec

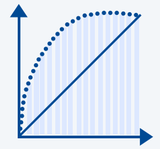
K- Nearest Neighbour
5.22 sec

Neural Networks
18.33 sec

Neural Networks
0.95

Logistic regression
0.94

K-Nearest Neighbour
0.90



Neural Networks
0.97

Logistic regression
0.96

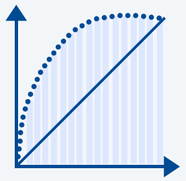
K-Nearest Neighbour
0.91

The Tiny two

SVM- *Unusually high precision*

GMM- *Unusually high recall*

1.0



Random Classifier equivalent AUC

0.51

0.52

Support Vector Machines
Gaussian Mixture Models

Source: wired.com

- Accent recognition is a preprocessing step to speech recognition.
- It fine tunes speech recognition systems- *does not* fundamentally change how they work.
- When paired with an apposite model, we can achieve required results.

Conclusion

Thank you