ST311 ARTIFICIAL INTELLIGENCE

Accent Recognition

Candidate Numbers: 31948, 24692, 24155

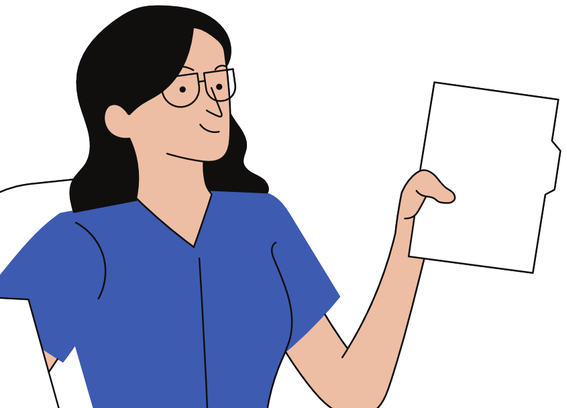
London School of Economics

Why Accent Recognition Important?

Taking Amazon Alexa and Google Home as example

A recent study found that voice recognition systems from virtual assistant technologies like Amazon's Alexa and Google Home have a harder time understanding people who speak English with a foreign accent or with a Southern accent.

To solve this problem, these systems need to get better at recognizing different accents. By improving how they detect and understand these accents, they can work better for everyone, no matter how they speak.



This would make such technology more useful and accurate for people all around the world!

Why Accent Recognition Important?

A Broader area of Application

Help on Speech recognition

- Better understand diverse speech patterns
- Reducing errors/misunderstandings in commands
- Can be added to other recognition: Language Sentiment Anaysis

Help on Personalized Content

- Content AI Recommendations based on accents
- Personalised Chatbot services
- And any voice personalized services!

Help on Language Learning

- Good for learning certain accents if this is the goal
- Help on Language translation (One example from Meta)
- And even save some endangered languages

Example: Language Translation

Meta's Universal Speech Translater model, developed by Pen-Jen Chen, is used to translate unwritten languages.

This model can translate Hokkien, dialect of Chinese which is an unwritten lamguage, to English.

Overview of our Research

What is the dataset we are using and the structure of this analysis?

1

- Look at what others are doing through the academic research.
- What can be the potential improvements from our research?
- What can we learn from the similar research?

Literature Review

2

- Delete information we don't need: age and gender of audio source
- Deal with the problem of class imbalance - introduce later
- Format the data that the model accepted

Data-Precessing

3

- Train the three type of models tailored to the three type of inputs
- Compare the result: How well the model can classify the data?
- What are we learning from the result and the research process

Model Training

The Dataset we used in this research

7,000 audios with corresponding scripts

The model need to identify 8 types of accents

We transfer the audio to spectrograms through existing packages

Sample Audio of Each Accent

The three types of inputs: Audio, Spectrograms and Text

The assumptions - Why are we using and comparing?

Audio

This is the most direct way, raw audio is the perfect ingredient to capture the essence of an accent, like using authentic spices in a recipe.

Speech Transcript

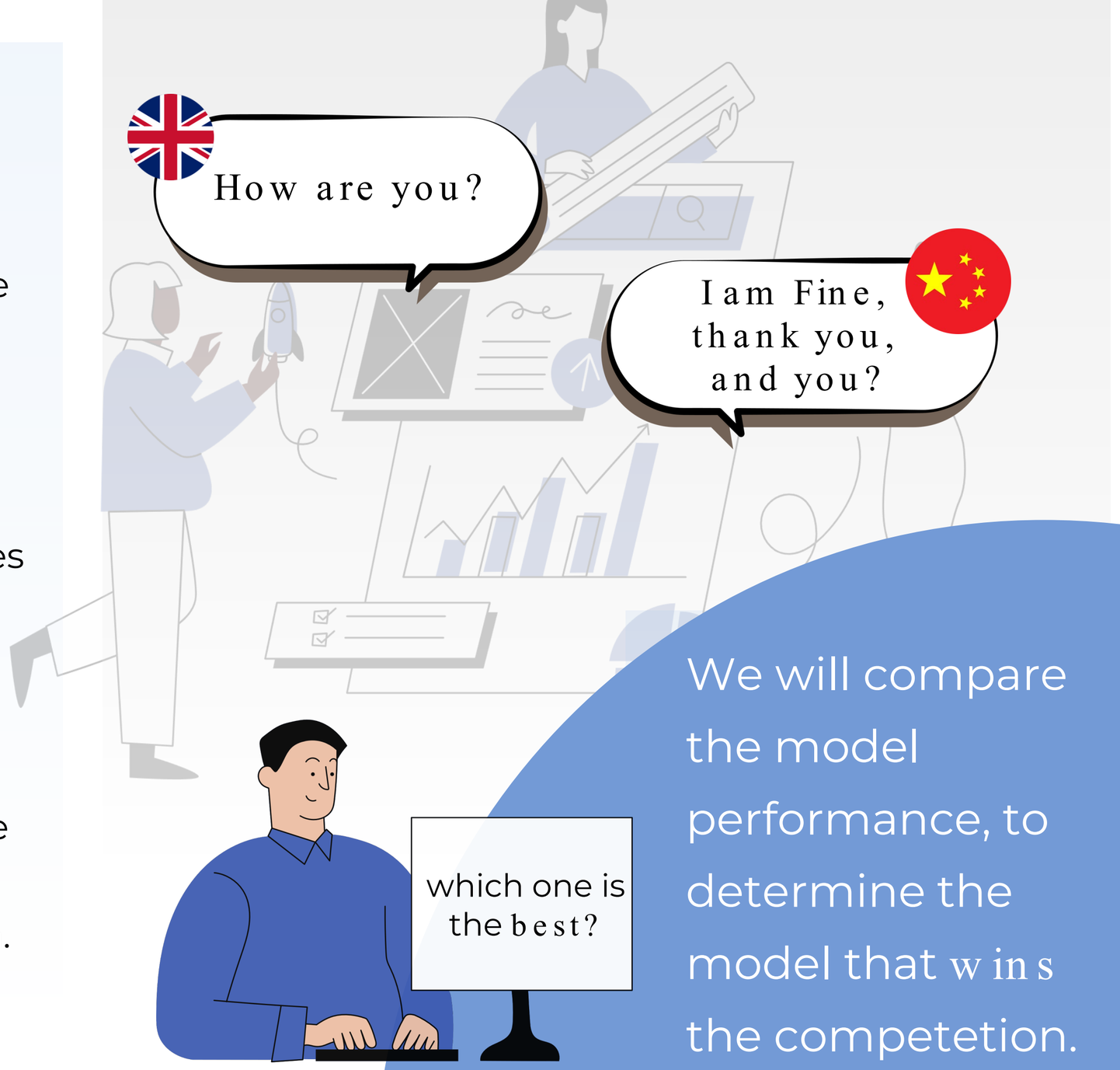
We assume that people with different English accents may have different native languages or languages environment. This type of difference may influence their way of organizing the sentence or using certain words. Therefore text as audio scripts is part of our intertest.

Spectrogram

In physics, the audio can be transferred to spectrograms based on the features of the audios such as frequency, speed, and pitch.

For the accent recognition, we assumed that different accents may have differences in terms of phoneme, stress and strength. Therefore these differences can be captured by the spectrograms.

A image-based transformer model has the capability to classify image, therefore it should be able to classify the spectrogram.

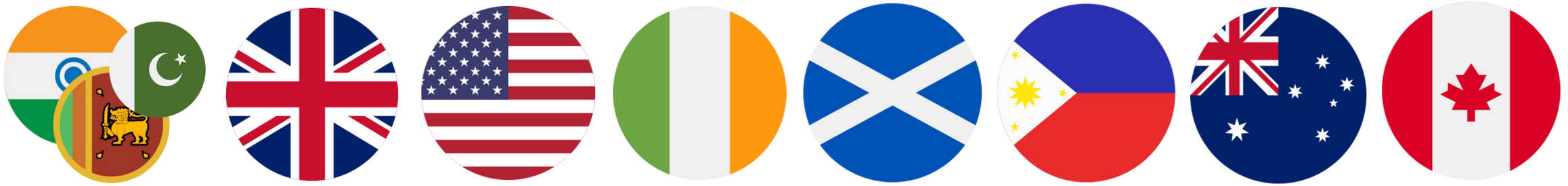


We will compare the model performance, to determine the model that wins the competetion.

Class Imbalance

Down-sample classes Up-sample classes

Some accents will have less samples than other countries accent. For example, the US, UK and India have more samples in the dataset. We have to solve this problem or the algorithm might not predict the result accurately in those less common classes. It's a bit like if a model is only trained on images of elephants and lions - it might struggle to recognize a rare tiger when it sees one!



963 891 2577 226 91 471 528 471

Data Augmentation

Creating new audios by changing the pitch, speed and adding noise into the current audios

Customized Loss

Giving more importance to those less common classes. The algorithm will assign more weight for less common classes then.

Subsampling and Duplicate

For text classification method, we choose to duplicate the minor data and take subsamples for major data to balance all classes.

Transformers - 3 Methods

- Previously trained with large datasets
- Using weights and biases similar to the coefficients and bias in linear regression.
- Unlike linear regression, which models simple linear relationships, transformers dynamically adjust these weights to capture complex dependencies and contexts within data.
- Transformers are extremely versatile for different classification problems like images, audios and texts.

Audio Model - wav2vec 2.0

Trained on raw, unlabelled audio and then fine-tuning on a small amount of transcribed data. It extracts features that outperform traditional methods, especially when labeled data is scarce.

Image Model - ViT

Chops the image into puzzle pieces and figures out how they fit together to understand what the full picture shows. It's like looking at all the pieces at once to identify the key parts.

Image Model - BERT

Like a smart assistant that’s really good at understanding context. It does this by learning from a vast amount of text about how words can have different meanings depending on other words around them.

What is the performance of Text?

The result of the model trained based on text scripts of audio

20%
prediction
accuracy

- 20% of the data can be correctly predicted by this text classifier

Understanding the result

Good or Bad

VERY BAD. The text-based model cannot identify the accents.

Limitations

SAMPLE SIZE LOW. No practice, no perfect. The data we used is not enough for the model predict accurately.

Can we do better

MORE DATA. AND COMBINE TEXT WITH OTHER TYPES OF INPUTS. For example, classifying the accents with the model of text and spectrograms together.

What is the performance of Spectrogram?

The result of the model trained based on Spectrogram image of audio

99.0%
prediction
accuracy

- 99% of the data can be correctly predicted by this text classifier

Understanding the result

Good or Bad

EXTREMELY GOOD. The image-based model cannot identify the accents.

Limitations

The US Accent Prediction is worse
When we test the model in a extra dataset of the US accent, the accuracy is only around 70%.

Can we do better

Collect Audio with More Difference
The Accuracy of our prediction is super high but it might not predict that good at a new dataset. We should test on data from different sources.

What is the performance of Raw Audio?

The result of the model trained based on Spectrogram of audio

99.8% prediction accuracy

- Only two speeches misclassified among 1536 test data
- These two speeches are very short, three and five words long.

Understanding the result

Good or Bad

ALMOST PERFECT.
The audio-based model correctly identifies the accents with close to zero errors

Limitations

POTENTIALLY OVERFITTING.
May not generalise well on out-of-sample dataset if the entries in dataset is too similar

Can we do better

LIMIT WEIGHTS. ADD MORE ACCENTS
Limit the model's inclination to overfit by limiting the model's weights and more accent allows the model to learn from less common English accent.

What else we learned from this study?

A non-techinal, interesting finding

- An interesting finding is that most misclassifications are other accents misclassified as American English accent and American accent audios misclasified as other English accent.
- We can say that American accents have more variants than other English accents from non-English countries.
- This may due to American accent is an accent with a diverse range of dialects with generations of immigrants.

That's a wrap!

April 30, 2024