

Deep Learning-based Models for English Accent Classification

Jiacheng Hou

300125708



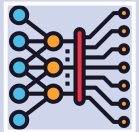
Outline

- Project Summary
- Motivation
- Dataset
- Methodology:
 - Data Pre-processing
 - A Baseline Model: Multilayer Perceptron (MLP)
 - A Refined Model: Transformer Encoder [\[1\]](#)
- Experimentation and Results
- Activity Table
- Challenges and Takeaways
- Conclusions

Project Summary



An interface for calling the Google Speech-to-Text API [\[2\]](#) with audio input (.wav) and English accents.



Two deep learning-based models to classify English speech accents:

- A baseline model: MLP
- A refined model: Transformer Encoder

Motivation

Google Speech-to-Text API [\[2\]](#)

Google Cloud

Why Google

Solutions

Products

Pricing

Getting Started



Docs Support

English

Cloud Speech-to-Text

Speech-to-Text

Benefits

Demo

Key features

Customers

What's new

Documentation

Use cases

Improve customer service

Enable voice control

Transcribe multimedia content

All features

Pricing

Take the next step

Speech-to-Text

Accurately convert speech into text using an API powered by Google's AI technologies.

[Go to console](#)

[Contact sales](#)

- ✓ Transcribe your content with accurate captions
- ✓ Deliver better user experience in products through voice commands
- ✓ Gain insights from customer interactions to improve your service

BENEFITS

State-of-the-art accuracy

Apply Google's most advanced deep learning neural network algorithms for automatic speech recognition (ASR).

Easy model customization

Speech-to-Text UI enables experimentation, creation, and management of custom resources.

Flexible deployment


Deploy speech recognition wherever you need, whether in the cloud with the API or on-premises with [Speech-to-Text On-Prem](#).

Gartner

Gartner names Google Cloud a Leader in the 2021 Magic Quadrant for Cloud AI Developer Services.


[Register to download the report](#)

Motivation

- Speech Audio [\[3\]](#) 
- Ground Truth Accent: English (United States)
- Ground Truth languageCode: en-US

INPUT ACCENT	LANGUAGECODE	WORD ERROR RATE (WER) [4]
English (United States)	en-US	0.072
English (United Kingdom)	en-GB	0.202
English (India)	en-IN	0.652

Motivation

- Speech Audio [\[3\]](#) 
- Ground Truth Accent: English (United States), languageCode: en-US
- Input Accent: English (United Kingdom) , languageCode: en-GB

Ground truth text	Ground truth text after pre-processing	Predicted text	Predicted text after pre-processing	WER
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. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.	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 we also need a small plastic snake and a big toy frog for the kids she can scoop these things into three red bags and we will go meet her wednesday at the train station	please call still here to bring these things with her from the store 6 pounds of fresh no peas 56 loads of blue cheese and maybe a snake for her brother tab we also need a small plastic snake in the big toy flag for the kids she can sort these things into 3 red bags and we will go meet her Wednesday at the train station	please call still here to bring these things with her from the store six pounds of fresh no peas 56 loads of blue cheese and maybe a snake for her brother tab we also need a small plastic snake in the big toy flag for the kids she can sort these things into three red bags and we will go meet her wednesday at the train station	0.202

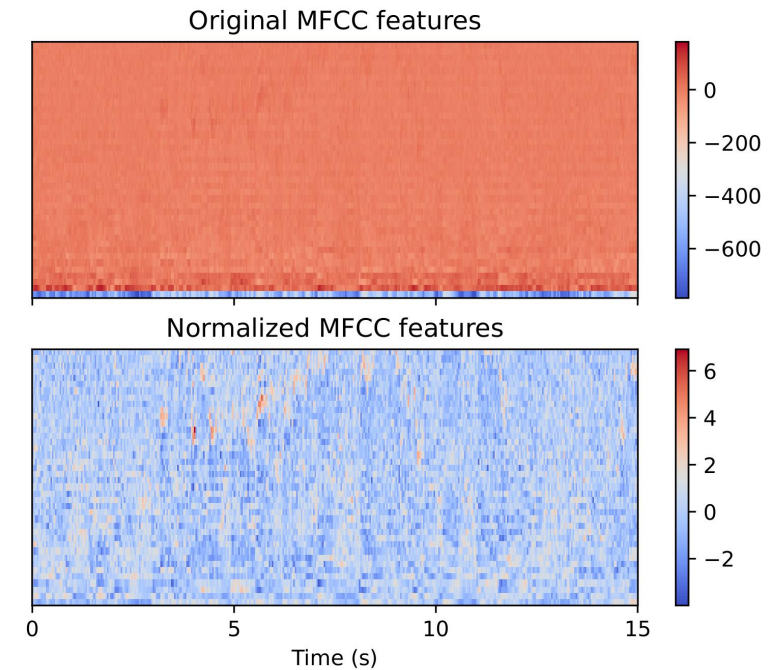
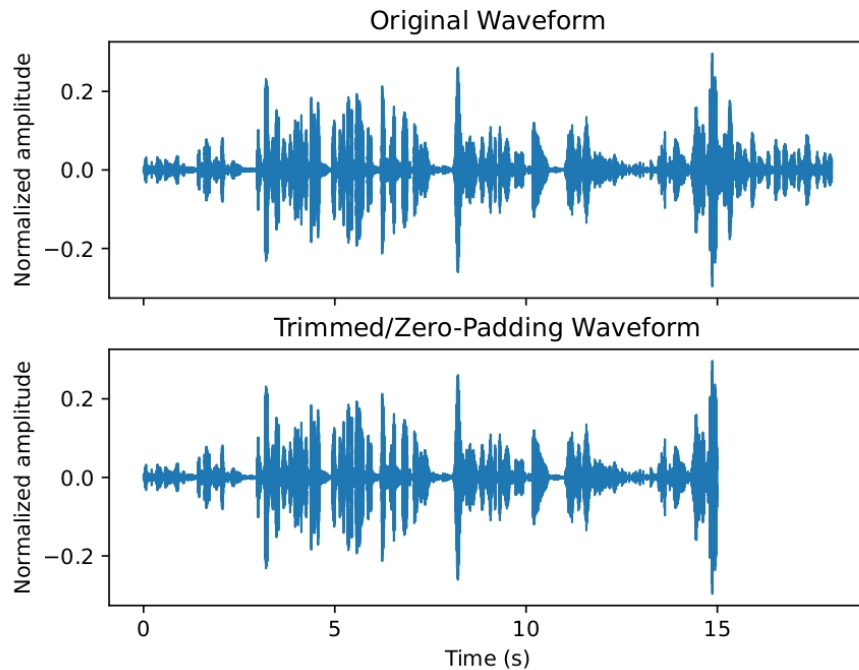
Dataset

- UT-Podcast [\[5\]](#)
- Accents:
 - Australian English (AU)
 - American English (US)
 - United Kingdom English (UK)
- Sampling Rate: 8000 Hz

<div>Dataset \ Class</div>	AU <div>🔊</div>	UK <div>🔊</div>	US <div>🔊</div>
Training	449	246	406
Testing	332	89	240

Methodology – Data Pre-processing

1. Fix the length of all audios to 15 seconds [\[6\]](#)
2. Extract Mel-frequency cepstral coefficients (MFCCs) from each audio [\[7\]](#). Window length = 28ms, stride = 10ms, # MFCCs for each window = 40.
3. Perform the Z-score normalization of MFCCs for each audio [\[8\]](#)



Methodology - MLP

Input layer

- # MFCCs for each window * # windows = $40 * 1501 = 60,040$ neurons

Two hidden layers

- 32 neurons with a ReLU activation function, and a dropout probability 0.2
- 16 neurons with a ReLU activation function, and a dropout probability 0.2

Output layer

- A linear layer with 3 neurons

Methodology – Transformer Encoder [\[1\]](#)

- Input data shape:
 - [# MFCCs for each window, # windows] = [40, 1501]
- Apply positional encoding to the input data [\[9\]](#)
- 2 Transformer Encoder layers, each includes:
 - A Multi-head Attention layer with # heads = 2
 - Dropout probability: 0.2
 - Activation function: relu
 - A feedforward layer dimension = 200
- Output layer:
 - A linear layer with 3 neurons

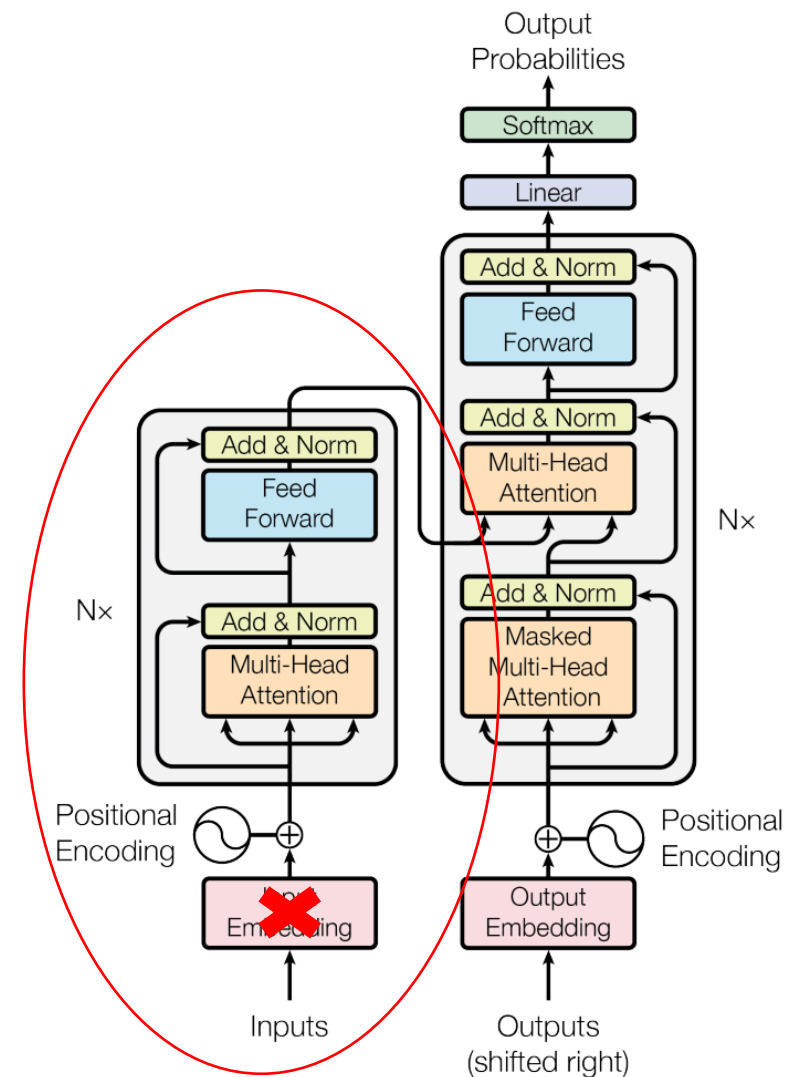


Figure 1: The Transformer - model architecture.

Experimentation and Results

Model Training Parameters and Values

Train/Valid Split	90/10
Batch Size	1
Learning Rate	1e-3
Epoch	100
Patience	20
Loss	Cross-Entropy Loss [10]
Optimizer	Adam [11]

Experimentation and Results

Model Performances on the Testing Dataset

Models	Labels	Precision	Recall	F1 score	Accuracy	Macro Average		
						Precision	Recall	F1 score
MLP	AU	0.56	0.45	0.50	0.42	0.40	0.41	0.39
	UK	0.18	0.40	0.25				
	US	0.46	0.38	0.41				
Transformer Encoder	AU	0.78	0.80	0.79	0.75	0.68	0.68	0.68
	UK	0.44	0.44	0.44				
	US	0.82	0.79	0.80				

Activity Table

Activity	Why	Time Used	Deliverable
Read articles and search materials	Gather knowledge about accent classification and relevant libraries	6h	
Explore dataset	For training and testing deep learning models	3h	The UT-Podcast corpus
Call the Google Cloud Speech-to-Text API and Calculate the WER	Present the accent information is important for the Speech Recognition module	7h	WER of ground truth text and API transcribed text
Data preprocessing	Required to make data into the proper format to feed into deep learning models	4h	Normalized MFCCs of the audio data
A MLP architecture	To have a baseline model to compare with	3h	A MLP architecture
A Transformer Encoder architecture	To have a refined model to compare with	5h	A Transformer Encoder architecture
Evaluate the baseline and refined models	To evaluate and analyze the performances of the two models	2h	Evaluation metrics and results

Challenges and Takeaways

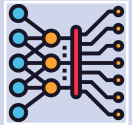
Challenges:

1. Text pre-processing before computing WER
2. Find a high-quality dataset
 - Quantity limitation: dataset choices, dataset sizes
 - Non-uniform sampling rate
 - Need to pay
3. Achieve satisfying results

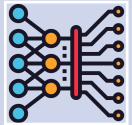
Takeaways:

1. Get hands-on experience in computing WER and discover previously unthought-of problems.
2. Gain practical experience in pre-processing audio data and playing with the Transformer Encoder model.

Conclusions



The project classified English accents using two deep learning models: MLP and Transformer Encoder.



The two models were tested on the UT-Podcast corpus, and the Transformer Encoder achieved a 74% improvement on the macro F1 score.



For coding details, please refer to the following GitHub link:

https://github.com/Jiacheng98/CSI5180_Accented_English_Speech_Classification

References

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. [1]
- *Speech-to-Text: Automatic speech recognition* | Google cloud. (n.d.). Google Cloud. <https://cloud.google.com/speech-to-text> [2]
- The Speech Accent Archive. http://accent.gmu.edu/browse_language.php?function=detail&speakerid=114 [3]
- Jiwer. (n.d.). PyPI. <https://pypi.org/project/jiwer/> [4]
- Hansen, J. H., & Liu, G. (2016). Unsupervised accent classification for deep data fusion of accent and language information. *Speech Communication*, 78, 19-33. [5]
- Librosa.util.fix_length — librosa 0.9.1 documentation. (n.d.). Librosa. https://librosa.org/doc/main/generated/librosa.util.fix_length.html [6]
- Librosa.feature.mfcc — librosa 0.9.1 documentation. (n.d.). Librosa. <https://librosa.org/doc/main/generated/librosa.feature.mfcc.html> [7]
- Freesound general-purpose audio tagging challenge | Kaggle. (n.d.). Kaggle: Your Machine Learning and Data Science Community. <https://www.kaggle.com/c/freesound-audio-tagging/discussion/54082> [8]
- Language modeling with nn.Transformer and TorchText — PyTorch tutorials 1.11.0+cu102 documentation. (n.d.). PyTorch. https://pytorch.org/tutorials/beginner/transformer_tutorial.html [9]
- CrossEntropyLoss — PyTorch 1.11.0 documentation. (n.d.). PyTorch. <https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html#torch.nn.> [10]
- Adam — PyTorch 1.11.0 documentation. (n.d.). PyTorch. <https://pytorch.org/docs/stable/generated/torch.optim.Adam.html> [11]