Lexical Stress Detection in Spoken English

Aditi Singhal, Madhavi Srinivasan, Vivek Kumar

Introduction

- In spoken English, *Stress is essential to being understood*. Misplacing the syllabic stress can alter a word's part of speech or even alter the meaning of the word
- We have designed a deep learning model that will take spoken sentences as input, slice them at phoneme-level, and classify each vowel phoneme into a binary-class output - primary stress or no stress
- Finally we have processed the output to report errors in the following cases:
  i) The user interchanges stress between primary and unstressed vowel phonemes
  ii) The user doesn't stress any vowel phoneme

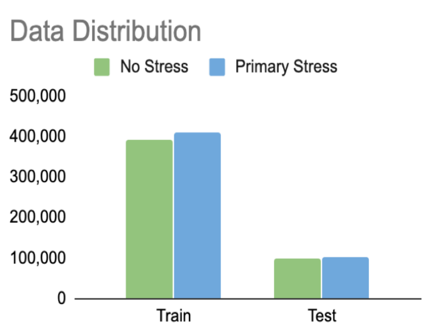
Data Summary

LibriSpeech [2]

- Contains 460 hours of speech of ~1100 native English speakers

OSCAAR [3]

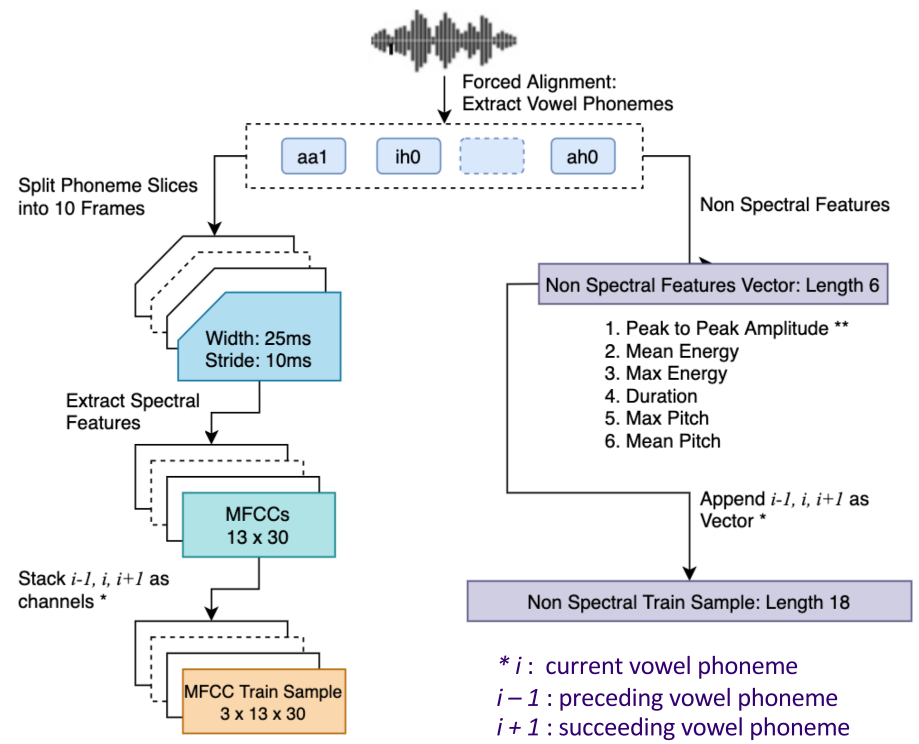
- 4907 audio files, Speaker variation - 10 females, 2 males



Experiments

- Initially we started the modelling process with 3 classes – Primary, Secondary and No-Stress. Since very few words have secondary stressed vowel, we had a class imbalance problem. Additionally secondary stress is very similar to primary stress, hence, we decided to model this as a binary classification problem with classes primary stress and no-stress
- During initial iterations, we observed most of the misclassified phonemes belonged to high-frequency common words like *the, at, to*. To address this problem, an additional preprocessing step was added to remove 80 stop words. This increased the accuracy from 89% to 96%.

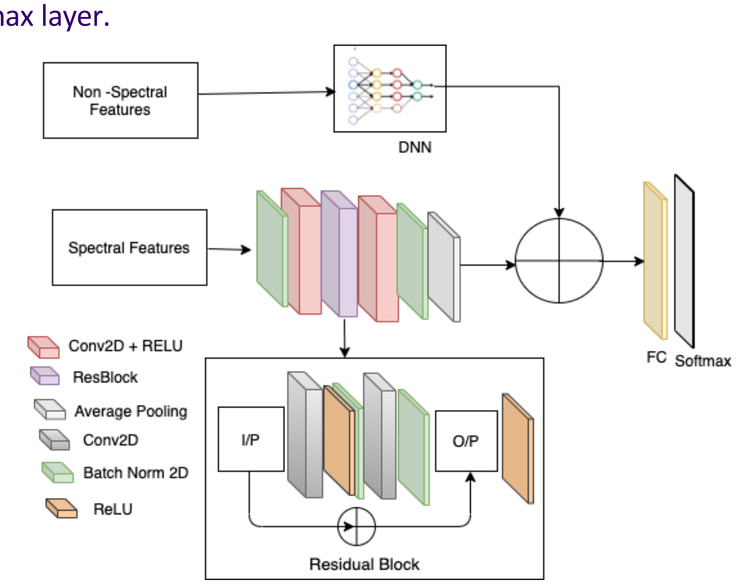
Data Preprocessing



*i : current vowel phoneme
$i-1$: preceding vowel phoneme
$i+1$: succeeding vowel phoneme
MFCC - Mel Frequency Cepstral Coefficients

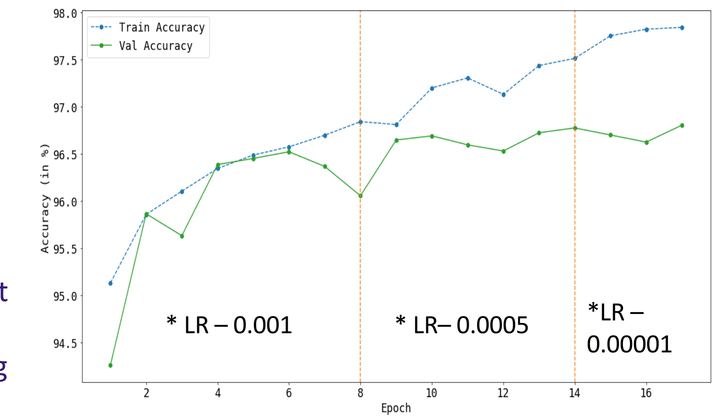
Model Architecture

The model is inspired from [1]. Spectral features are passed through a Residual Convolutional Neural Network and Non-Spectral features are fed into a Deep Neural Network. Their output is concatenated before the softmax layer.



Results

- The baseline accuracy without any training was 49.78% on validation set
- The architecture achieves a maximum accuracy of 96.80% on the validation set and 97.84% on train set after 17 epochs with varying learning rates



* LR– Learning rate

Future Work

- Use Data sources with more speaker variation
- Model the problem using sequence to sequence models to capture stress variation across words in continuous speech

References

[1] Shahin, Mostafa Ali, Julien Epps, and Beena Ahmed. "Automatic Classification of Lexical Stress in English and Arabic Languages Using Deep Learning." INTERSPEECH. 2016.
[2] Librispeech: An ASR corpus based on public domain audio books, retrieved from http://www.openslr.org/12/
[3] https://oscaar3.ling.northwestern.edu/ALLSSTARcentral/#!/recordings