Vocal Fluency Classification using SVM and LSTM

Arkadeep Dutta
Anant Ghuman

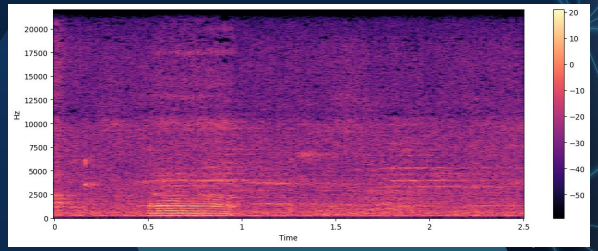
Introduction

- Vocal Fluency Analysis: How fluent is the speaker?
- Assess learners' pronunciation and fluency and return their level of English expertise
- Enhances pronunciation skills, automates language assessment and evaluation
- Current implementations use solely SVM or solely CNN
- Existing research and attempts are:
  - Biased
  - Inconsistent
  - Expensive

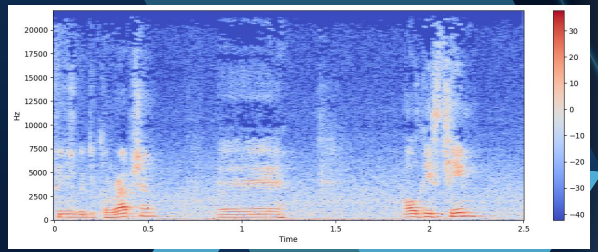
Data

- 5-second audio clips of non-native and native English speakers
- Three requirements:
  - Little background noise
  - Random conversation topics
  - Unscripted conversations
- Data classified into Low, Med, High Fluency folders
- Extracted features: MFCCs, spectral contrast, ZCR, RMSE, spectral flatness, and spectrogram frames

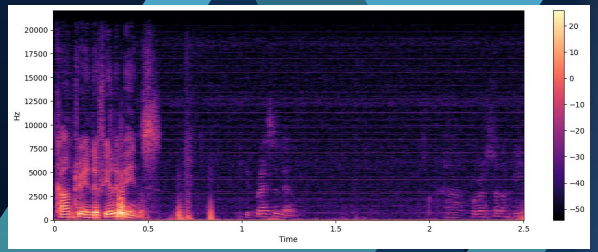
Low Fluency

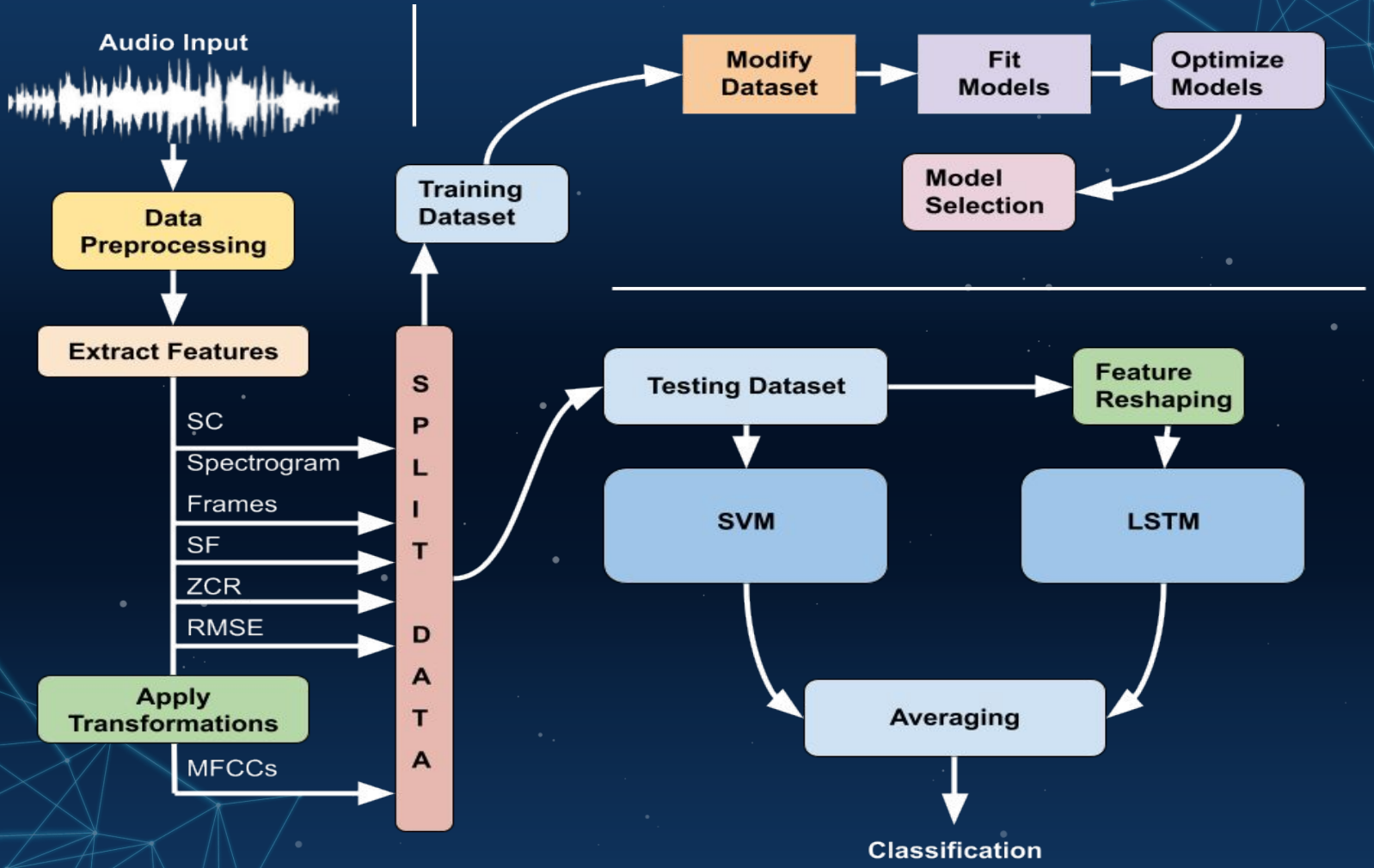


Med Fluency



High Fluency




Audio Input
Data Preprocessing
Extract Features
SC
Spectrogram
Frames
SF
ZCR
RMSE
Apply Transformations
MFCCs
SPLIT DATA
Training Dataset
Modify Dataset
Fit Models
Optimize Models
Model Selection
Testing Dataset
Feature Reshaping
SVM
LSTM
Averaging
Classification

Results

- Accuracy: 83.51%
- Precision: [0.75, 0.80, 0.97]
  - Macro Average: 0.839
  - Weighted Average: 0.837
- Sensitivity/Recall: [0.77, 0.81, 0.93]
  - Macro Average: 0.836
  - Weighted Average: 0.836
- F-Score: [0.76, 0.80, 0.95]
  - Macro Average: 0.837
  - Weighted Average: 0.836
- Specificity: [0.77, 0.81 0.93]
  - Macro Average: 0.837
  - Weighted Average: 0.836

Results (Cont.)

Clip:	Clip 1	Clip 2	Clip 3
SVM	[1,0,0]	[0,1,0]	[0,0,1]
LSTM	[0.65, 0.22, 0.13]	[0.24, 0.43, 0.33]	[0.16, 0.21, 0.62]
Output	Low	Med	High
Confidence	75.5%	60.1%	73.4%

Conclusion

- Managed to create somewhat accurate model in a quick and efficient manner
- SVM and LSTM combination most efficient for both temporal data and spectrogram frame analysis
- Biased towards certain accents
- Works worse for low/medium frequency users
- Only 3 classes
- Could be more accurate

Future Steps

- More labelled data necessary
- More classes in between low/med, med/high
- More features to analyze
- Expand to more languages (Spanish, French, etc)
- Provide feedback on specific aspects of vocal fluency
- Implement a chorusing evaluation function
- Turn into an app