Word Weaver - The ML Reading Revamp

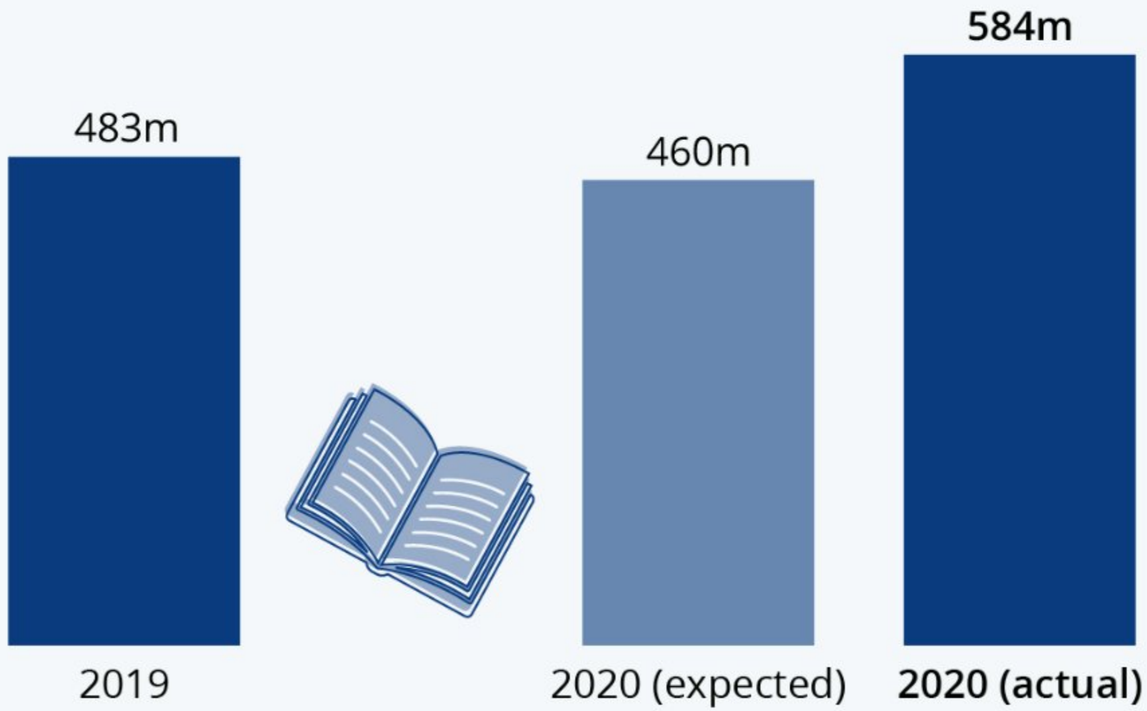
Enhancing Reading Proficiency through Machine Learning

Anirudh, Ronak & Vijeta

Problem Statement

Pandemic Causes Stark Rise in Child Reading Difficulties

Number of children in the world below the minimum proficiency level in reading



Source: UNESCO

statista

The COVID-19 pandemic dramatically increased global reading challenges, with the number of children not meeting minimum proficiency soaring from 460 million to 584 million, erasing two decades of educational progress.

UNICEF. (2021, March 26). Over 100 million children will fall below the minimum proficiency level in reading due to the impact of COVID-19 school closures. Culture and Education.

But this is not where it ends...

The World Bank report shows that approximately 70% of 10-year-olds worldwide now find themselves in what is termed "learning poverty," unable to read and understand a simple text.

Globally, only a third of 10-year-olds are estimated to be able to read and understand a simple written story.

Our project addresses these challenges head-on by introducing a simple tool designed to quickly assess and improve reading skills. This tool aims to help users from all walks of life communicate more effectively and confidently, transforming the way we approach learning and literacy enhancement across diverse communities.

World Bank. (2022, June 23). 70% of 10-Year-Olds now in learning poverty, unable to read and understand a simple text. Who We Are News.

Our Solution

It goes beyond simply checking accuracy but delving deeper into four key aspects that make a confident reader:

- **Expression and Volume:** Can you capture the emotions of the text with your voice? Our tool analyzes how you use volume and expression to bring the story to life.

- **Phrasing and Intonation:** Do you naturally pause and emphasize to guide the listener? Our tool listens for phrasing and intonation that keeps your audience engaged

- **Smoothness:** Can you read fluently without stumbling? Our tool helps identify areas where you might want to practice smoothing out your delivery.

- **Pace:** Are you reading at a comfortable speed that allows your audience to understand? Our tool analyzes your pace to ensure it's clear and engaging.

Potential Application and Impact

Customized Learning Experiences: Adapts learning content and pace to individual student levels, enhancing engagement and effectiveness.

Early Intervention and Support: Identifies reading difficulties early to enable timely support and prevent long-term educational setbacks.

Enhanced Teaching Strategies: Allows educators to refine instruction based on detailed insights, improving efficacy and outcomes.

Scalable Educational Impact: Can be expanded to various regions, significantly boosting literacy rates and educational quality, especially in under-resourced areas.

Literature Review

1. SPEAKER FLUENCY LEVEL CLASSIFICATION USING MACHINE LEARNING TECHNIQUES" BY ALAN PRECIADO-GRIJALVA AND RAMON F. BRENA

2. AUTOMATIC ASSESSMENT OF CHILDREN'S ORAL READING USING SPEECH RECOGNITION AND PROSODY MODELING

SPEAKER FLUENCY LEVEL CLASSIFICATION USING MACHINE LEARNING TECHNIQUES"
ALAN PRECIADO-GRIJALVA AND RAMON F. BRENA

Aim of Study: Automate fluency level classification in non-native English speakers using machine learning.

Dataset Used: Created the Avalinguo Audio Set, containing labeled English audio clips segmented into 5-second intervals.

Model Used and Features Extracted: Utilized five machine learning models including MLP, SVM, RF, CNN, and RNN. Key features extracted were MFCCs, zero-crossing rate, root mean square energy, and spectral flux.

Key Findings: The SVM model reached the highest accuracy of 94.39%, with enhanced performance when combining 20 MFCCs with other selected features.

Preciado-Grijalva, A., & Brena, R. F. (2018, August 31). Speaker fluency level classification using machine learning techniques.

AUTOMATIC ASSESSMENT OF CHILDREN'S ORAL READING USING SPEECH RECOGNITION AND PROSODY MODELING

- Aim of Study: Develop an automated system to provide feedback on children's oral reading using speech recognition and prosody modeling.

- Dataset Used: Utilized 200 recordings of English oral readings by children aged 10-14 from an urban school, annotated for prosodic and lexical features.

- Model Used and Features Extracted: Employed advanced speech recognition with features including pitch, intensity, spectral tilt, and duration measures for prosodic analysis.

- Key Findings: The system's assessments correlated well with human judges, highlighting the importance of combining lexical and prosodic analysis for comprehensive reading fluency evaluation.

Sabu, K., & Rao, P. (2018). Automatic assessment of children's oral reading using speech recognition and prosody modeling. CSI Publications, 201.

Gaps Identified

Influence of Native Language

Limited Data Sets

Speech Variability in Children

Comprehensive Feature Extraction

Real-World Application Challenges

Our Dataset

Our project leverages the LibriSpeech ASR corpus, which is an extensive dataset of about 1000 hours of 16kHz sampled English speech from public domain audiobooks via the LibriVox project. This dataset is ideal for developing automatic speech recognition (ASR) systems, featuring a diverse range of English accents and dialects.

Development Sets:
dev-clean and dev-other for tuning ASR models.

Test Sets:
test-clean and test-other for evaluating ASR performance.

Training Sets: train-clean-100, train-clean-360, and train-other-500 covering a total of 960 hours of speech for comprehensive ASR model training.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). LibriSpeech: An ASR corpus based on public domain audio books. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Retrieved from http://www.openslr.org/12/

Why are we using this Dataset?

- **Extensive Collection:** One of the largest available datasets with 1000 hours of English speech, offering a broad diversity of accents and styles.
- **High Quality:** Carefully segmented and aligned with text for accurate model training and evaluation.
- **Open Access:** Freely available under CC BY 4.0 license, facilitating widespread use and innovation in speech and language research.
- **Benchmarking Standard:** Recognized as a benchmark dataset, allowing for meaningful comparisons with state-of-the-art ASR models.
- **Comprehensive Resources:** Includes not only audio data but also transcripts providing a full suite of tools for our model development.

How was the Libri Speech data collected?

The data is derived from read audiobooks from the LibriVox project, and has been carefully segmented and aligned.

Were there and ethical concerns?

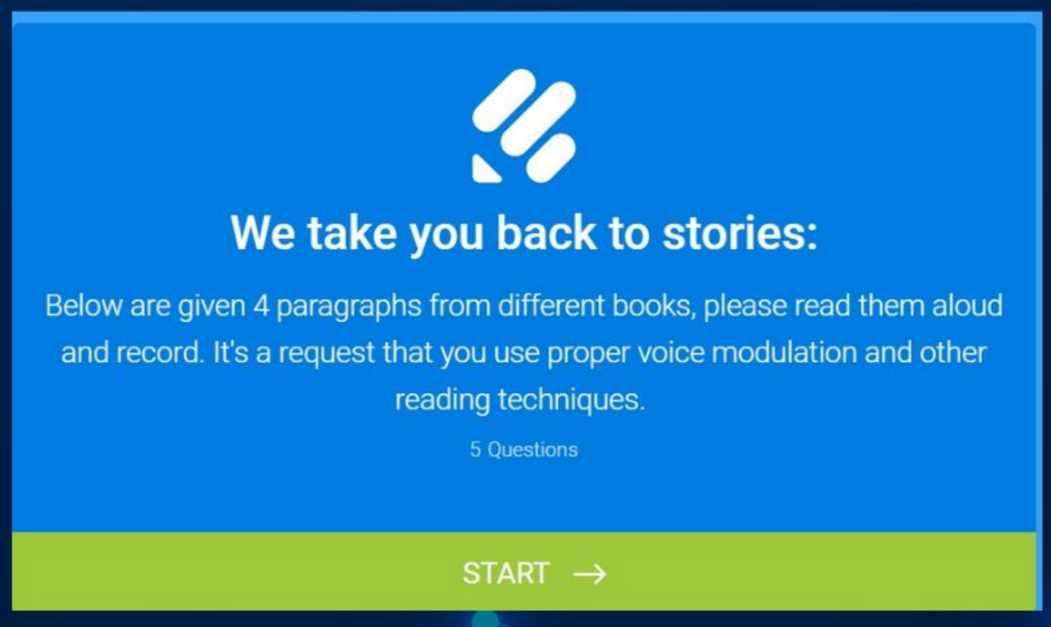
LibriVox volunteers narrate, proof listen, and upload chapters of books and other textual works in the public domain. These projects are then made available on the Internet for everyone to enjoy, for free.

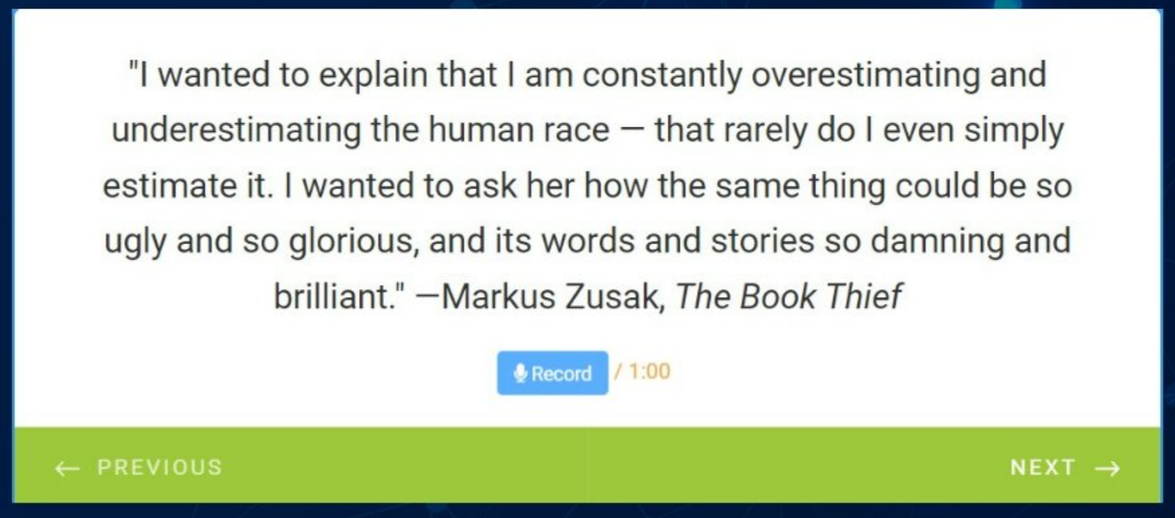
Challenges with the Dataset

1. *Only contains audio files and transcrips*
2. *No labels or evaluation metrics available*
3. *No features were available at all*

Data collected on our own

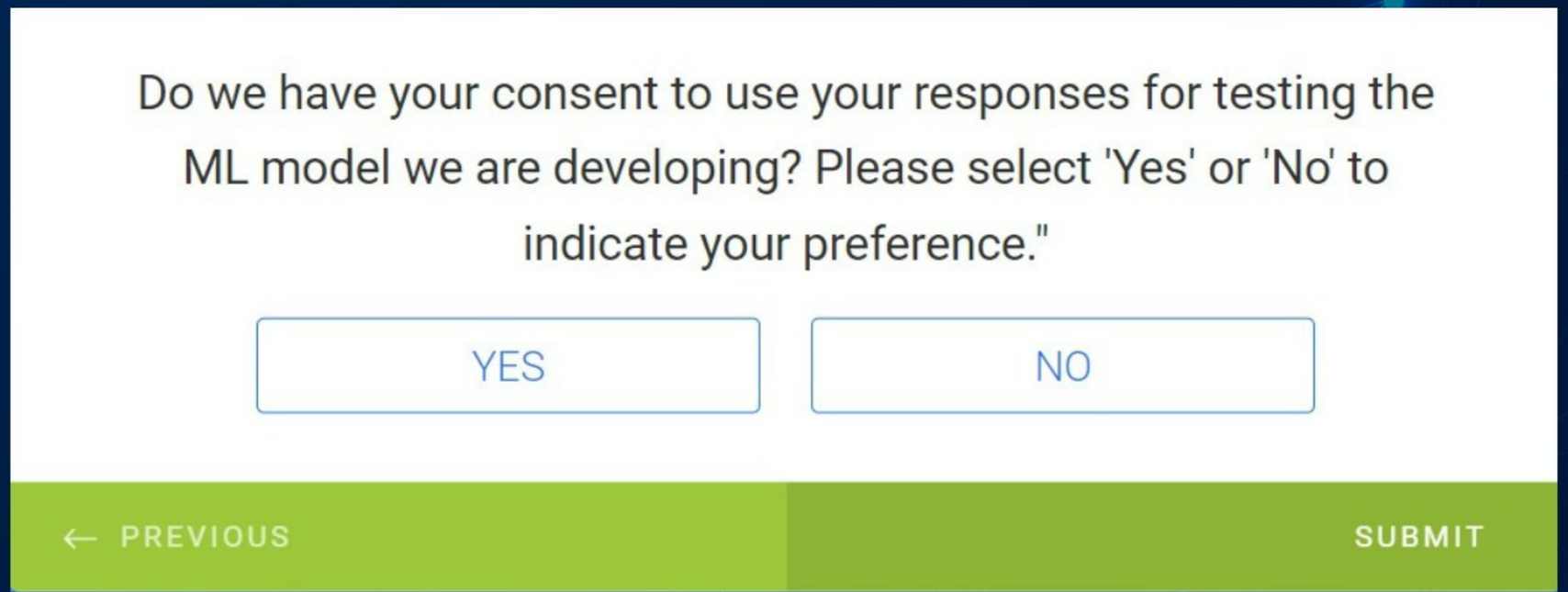
Data was collected using forms wherein the user was asked to record 4 audio samples.





Ehthical considerations for the same

When collecting audio at the university, we ensured to get explicit consent from participants via a form, which prioritized their anonymity and privacy without collecting any personal information.



Data Pre-processing

1. Labelling the data
2. Scale used for data labelling
3. Feature extraction

HOW DID WE LABEL THE DATA?

We had to do manual interventions to label our dataset by rolling out forms.
The target variables we provided to label our data was as follows:

- Expression and Volume
- Phrasing
- Smoothness
- Pace

Scale used to label the data:

Rating	Expression and Volume	Phrasing and Intonation	Smoothness	Pace
Circle one →	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4
1	Reads words as if simply to get them out. Little sense of trying to make text sound like natural language. Tends to read in a quiet voice.	Reads in monotone with little sense of phrase boundaries; frequently reads word-by-word.	Makes frequent extended pauses, hesitations, false starts, sound-outs, repetitions, and/or multiple attempts.	Reads slowly and laboriously.
2	Begins to use voice to make text sound like natural language in some areas but not in others. Focus remains largely on pronouncing words. Still reads in a quiet voice.	Frequently reads in two-and-three word phrases, giving the impression of choppy reading: improper stress and intonation fail to mark the ends of sentences and clauses.	Experiences several "rough spots" in text where extended pauses of hesitation are more frequent and disruptive.	Reads moderately slowly or too quickly.
3	Makes text sound like natural language throughout the better part of the passage. Occasionally slips into expressionless reading. Voice volume is generally appropriate throughout the test.	Reads with a mixture of run-ons, mid-sentence pauses for breath, and some choppiness; reasonable stress and intonation.	Occasionally breaks smooth rhythm because of difficulties with specific words and/or structures.	Reads with an uneven mixture of fast and slow pace.

Morrison, T. G., & Wilcox, B. (2020). Assessing Expressive Oral Reading Fluency. Education Sciences, 10(3), 59. https://doi.org/10.3390/educsci10030059

Our Training Dataset

Our labeled training dataset contains

- 984 audio samples
- The mean length of audio is 8 seconds
- Includes 30 unique readers who have read over 100 chapters. 22 - Female, 8-Male
- It’s a total of 3.5 hours of English Read Speech.

Our Testing Dataset

Our labeled testing dataset contains

- 200 audio samples
- Includes 50 unique readers

Feature Extraction and Preprocessing

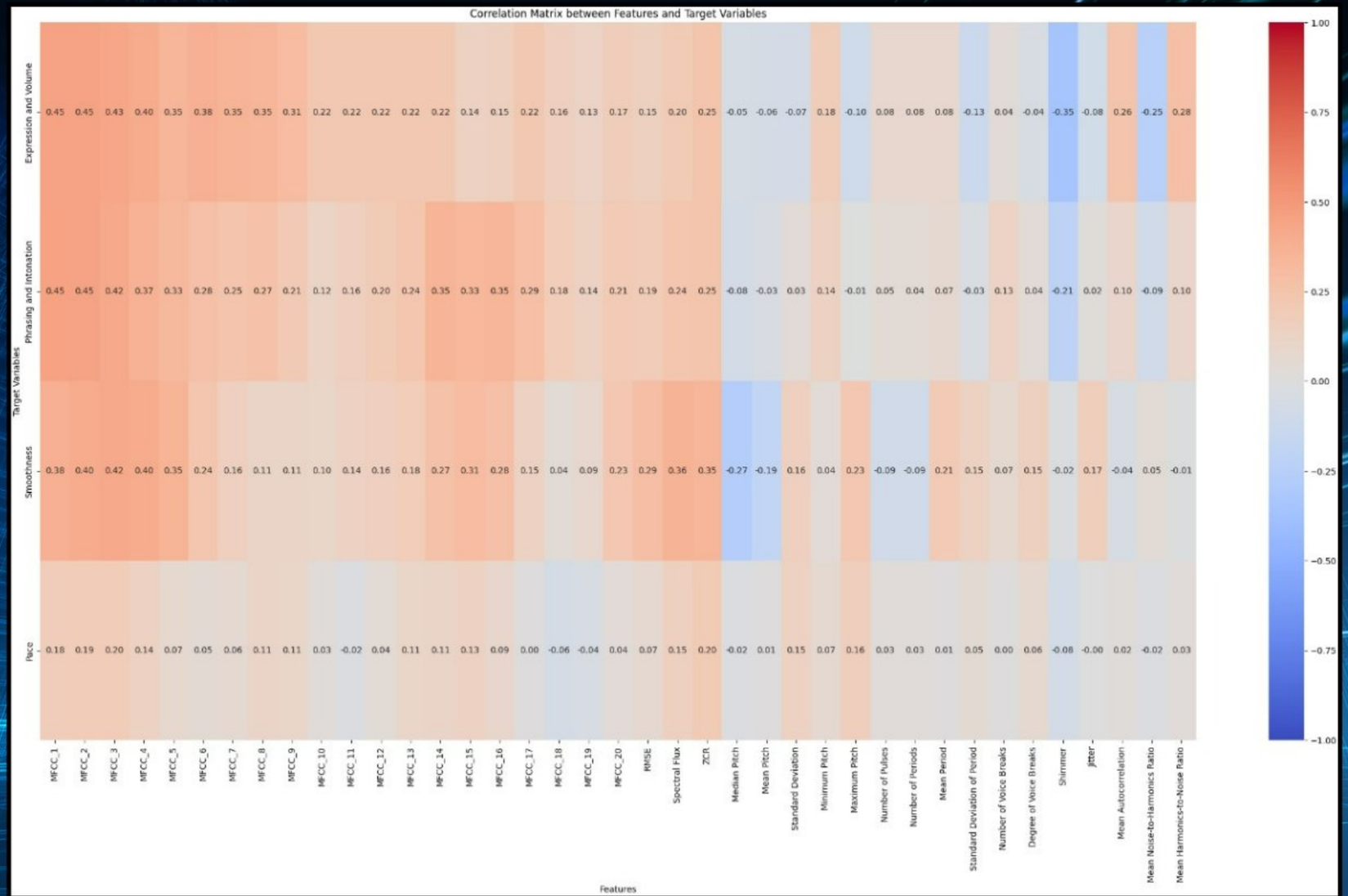
Prosodic Features Extracted

```
-- Voice report for  --
Date: Sat Mar 16 01:24:57 2024

WARNING: several of the following measurements will be incorrect,
because they are based on a sound from which higher frequencies have been filtered out.
For more correctness, go to "Pitch settings" and choose
the raw cross-correlation analysis method to optimize for voice research.

Time range of SELECTION
   From 3.651666 to 7.725000 seconds (duration: 4.073334 seconds)
Pitch:
   Median pitch: 177.826 Hz
   Mean pitch: 188.083 Hz
   Standard deviation: 37.998 Hz
   Minimum pitch: 75.729 Hz
   Maximum pitch: 281.485 Hz
Pulses:
   Number of pulses: 326
   Number of periods: 312
   Mean period: 5.374570E-3 seconds
   Standard deviation of period: 1.260542E-3 seconds
Voicing:
   Fraction of locally unvoiced frames: 53.309%   (145 / 272)
   Number of voice breaks: 8
   Degree of voice breaks: 35.931%   (1.463579 seconds / 4.073334 seconds)
Jitter:
   Jitter (local): 2.297%
   Jitter (local, absolute): 123.470E-6 seconds
   Jitter (rap): 0.927%
   Jitter (ppq5): 1.062%
   Jitter (ddp): 2.781%
Shimmer:
   Shimmer (local): 11.282%
   Shimmer (local, dB): 1.005 dB
   Shimmer (apq3): 3.839%
   Shimmer (apq5): 6.734%
   Shimmer (apq11): 13.511%
   Shimmer (dda): 11.518%
Harmonicity of the voiced parts only:
   Mean autocorrelation: 0.915432
   Mean noise-to-harmonics ratio: 0.105899
   Mean harmonics-to-noise ratio: 13.254 dB
```

The features we extracted were as follows:

- Median Pitch
- Mean Pitch
- Standard deviation
- Minimum pitch
- Maximum pitch
- Number of pulses
- Number of periods
- Mean period
- Standard deviation of period
- Number of voice breaks
- Degree of voice breaks
- Shimmer
- Jitter
- Mean autocorrelation
- Mean noise-to-harmonics ratio
- Mean harmonics-to-noise ratio

Significant Contribution only by MFCCs



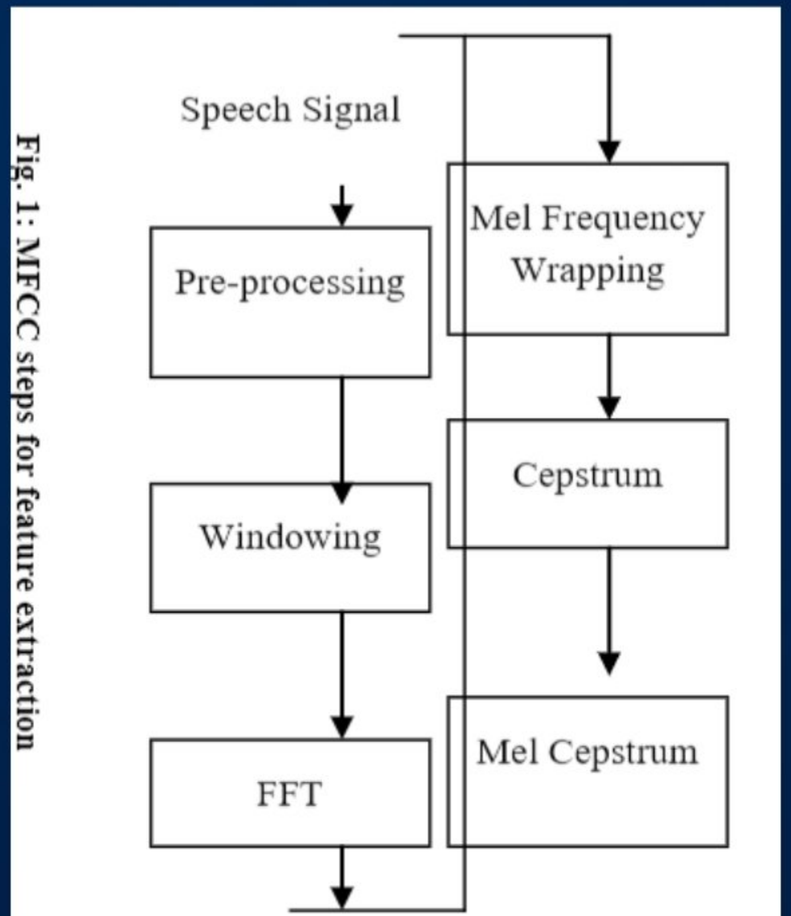


Fig. 1: MFCC steps for feature extraction

$$M = \frac{1000}{\log 2} \log\left(1+\frac{f}{1000}\right) \quad \ldots\ldots (1)$$

Where f is the frequency.

MFCC have the following steps when feature extracted from a speech signal-

MFCCs(Mel Frequency Cepstral Coefficients

MFCCs (Mel Frequency Cepstral Coefficients): A representation of the short-term power spectrum of sound, used in speech recognition and music analysis.

Pitch

Pitch in audio feature extraction refers to the perceived frequency of a sound, indicating how high or low the tone sounds.

Temporal Features Into Consideration

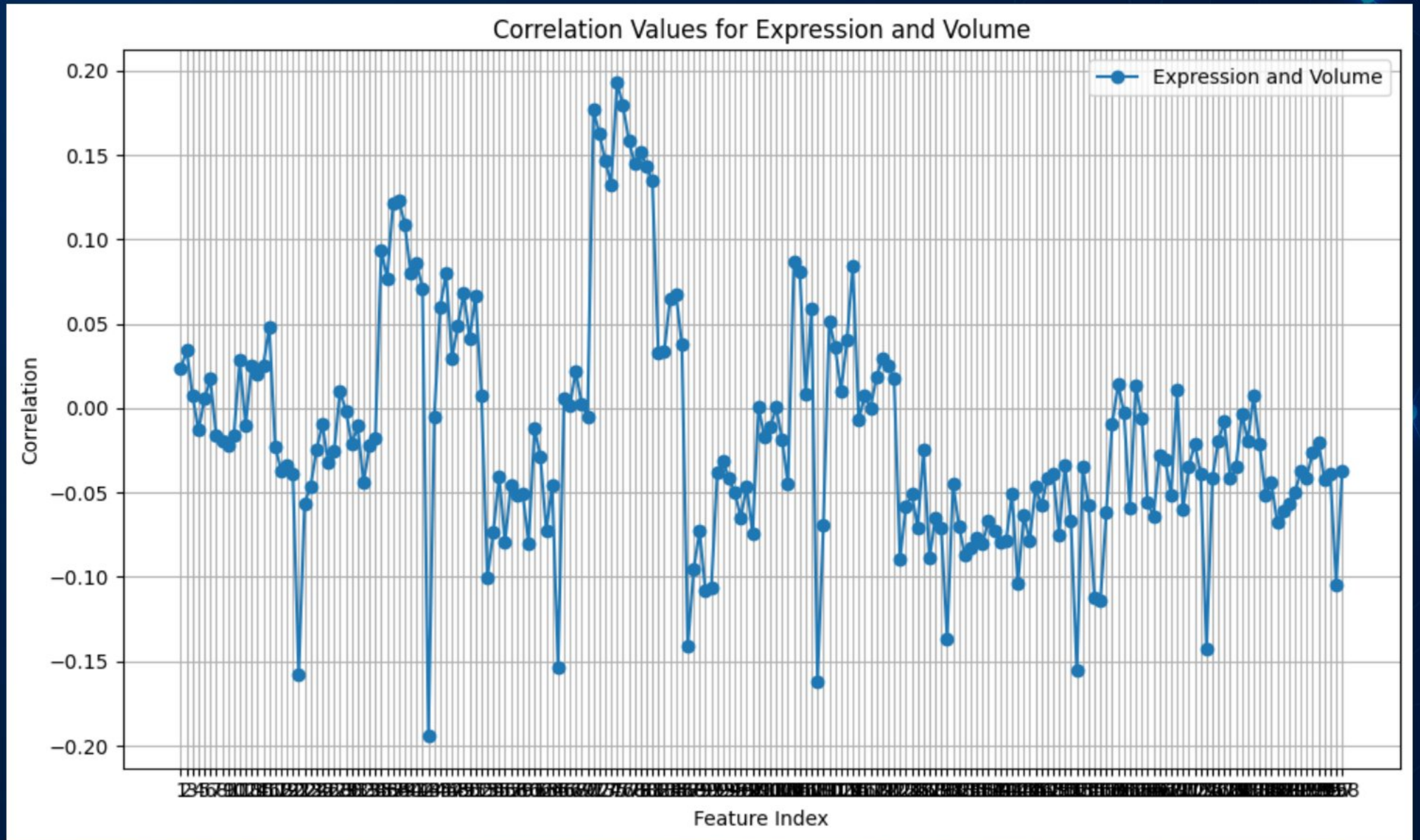
For each audio of 10 seconds , 20 MFCCs were extracted per second along with the mean pitch.

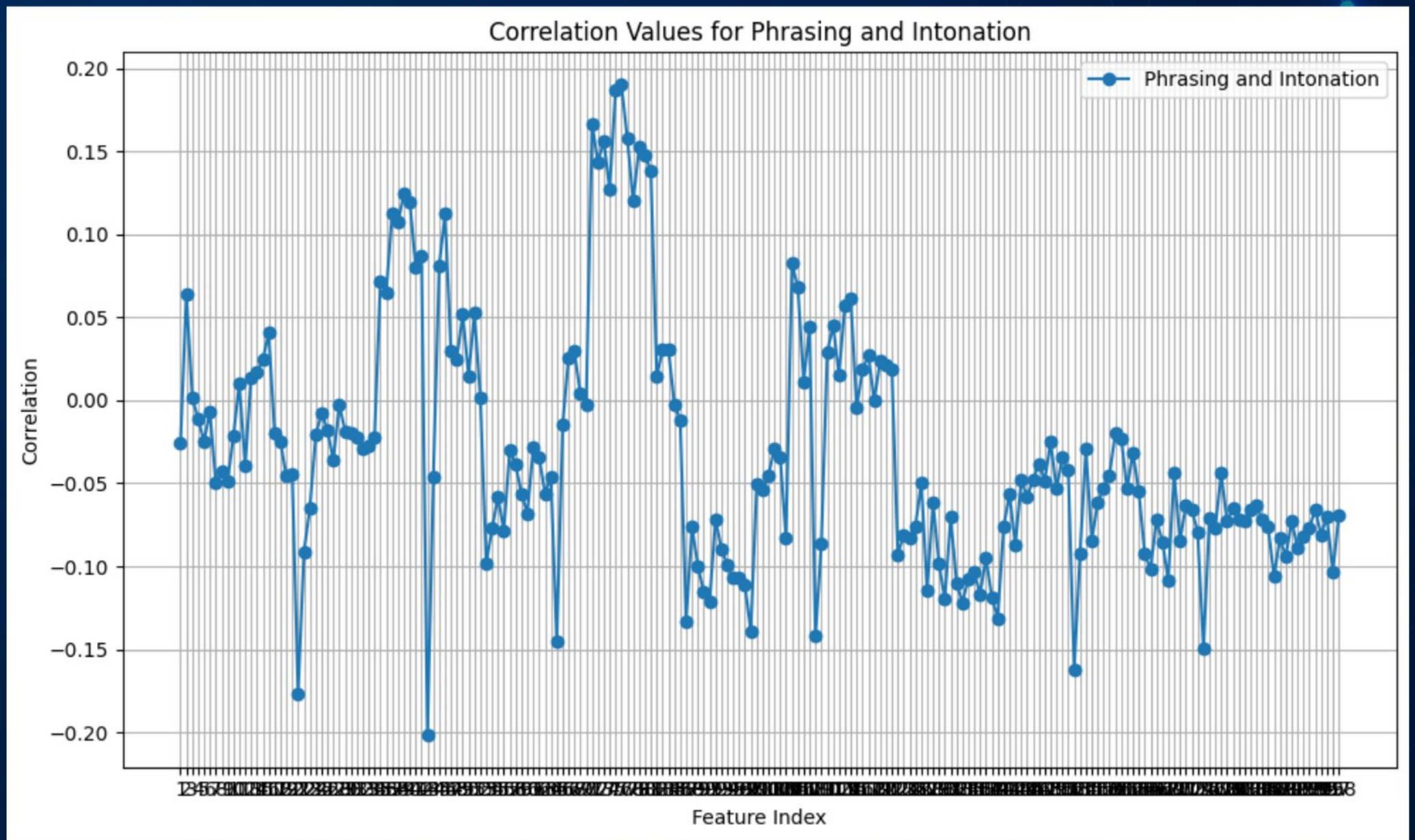
File Name	Expressior	Phrasing a	Smoothne	Pace	MFCC_1_1	MFCC_2_1	MFCC_3_1	MFCC_4_1	MFCC_5_1	MFCC_6_1	MFCC_7_1	MFCC_8_1	MFCC_9_1	MFCC_10_	MFCC_11_	MFCC_12
121-12172	2	2	2	1	-443.3	-407.148	-583.486	-208.521	-185.163	-395.759	47.94412	42.54451	24.36278	42.68267	73.20123	57.88136
121-12172	2	2	2	1	-287.658	-377.901	-507.379	-283.391	-881.832	89.29367	36.41231	25.42242	73.02939	50.86648	-32.4849	-6.29082
121-12172	2	2	2	2	-408.595	-382.095	-482.099	-232.658	-236.44	-218.619	-428.826	3.880723	44.67683	25.25409	87.97463	65.0114
121-12172	2	2	2	2	-453.757	-611.328	-469.072	-1131.37	11.52378	14.59939	17.96	0	-19.4225	-11.7845	-10.686	0
121-12172	3	3	3	3	-409.369	-588.293	-190.557	-357.119	46.79567	20.71331	69.68617	43.88192	-34.0116	5.850969	-38.1414	-10.4815
121-12172	2	2	2	2	-483.208	-446.507	-404.054	-239.066	-251.887	-232.621	-654.6	0.797091	-10.7236	32.76362	62.31522	49.90847
121-12172	2	2	2	2	-424.806	-684.411	-375.071	-236.489	-498.102	45.82055	6.075637	14.74754	63.71544	58.54764	-24.7852	5.680328
121-12172	3	3	3	3	-495.8	-502.961	-314.106	-199.3	-243.461	-305.897	-416.229	-1068.79	11.86448	57.05949	20.0903	77.12714
121-12172	3	3	3	3	-457.87	-497.947	-448.785	-218.719	-395.263	-273.95	-246.939	-289.318	-262.8	-591.122	-2.49689	49.357
121-12172	1	1	1	1	-382.515	-305.16	-430.965	-342.058	-849.993	28.62823	82.70094	77.58362	40.37636	44.83446	-0.38093	-27.5839
1688-1422	3	3	3	3	-332.11	-285.272	-388.907	-271.061	-260.475	-414.59	-193.832	-226.975	-357.263	-250.072	-309.574	-265.29
1688-1422	3	3	3	3	-291.251	-255.882	-483.979	83.6564	53.80949	60.41158	8.438785	61.97766	35.01836	54.57584	30.6779	55.04081
1688-1422	2	2	2	2	-327.783	-207.582	-324.799	-221.777	-399.554	-438.533	70.04159	108.8977	44.54787	90.312	36.20367	98.52591
1688-1422	2	2	2	2	-310.634	-283.665	-312.768	-281.324	-547.55	53.3235	97.25668	94.46826	122.9388	98.78751	24.2961	16.58396
1688-1422	2	2	2	1	-350.188	-250.669	-278.206	-299.025	-469.358	61.4999	76.06878	97.56253	88.44839	125.602	-1.82227	42.26729
1688-1422	1	1	1	1	-337.917	-279.794	-288.265	-278.908	-265.168	-313.571	-247.233	-387.662	-469.081	99.26572	98.86253	126.6008
1688-1422	1	2	1	1	-404.682	-249.804	-288.631	-236.247	-340.564	-275.772	-374.271	-570.802	81.66349	62.40535	75.56516	92.0382

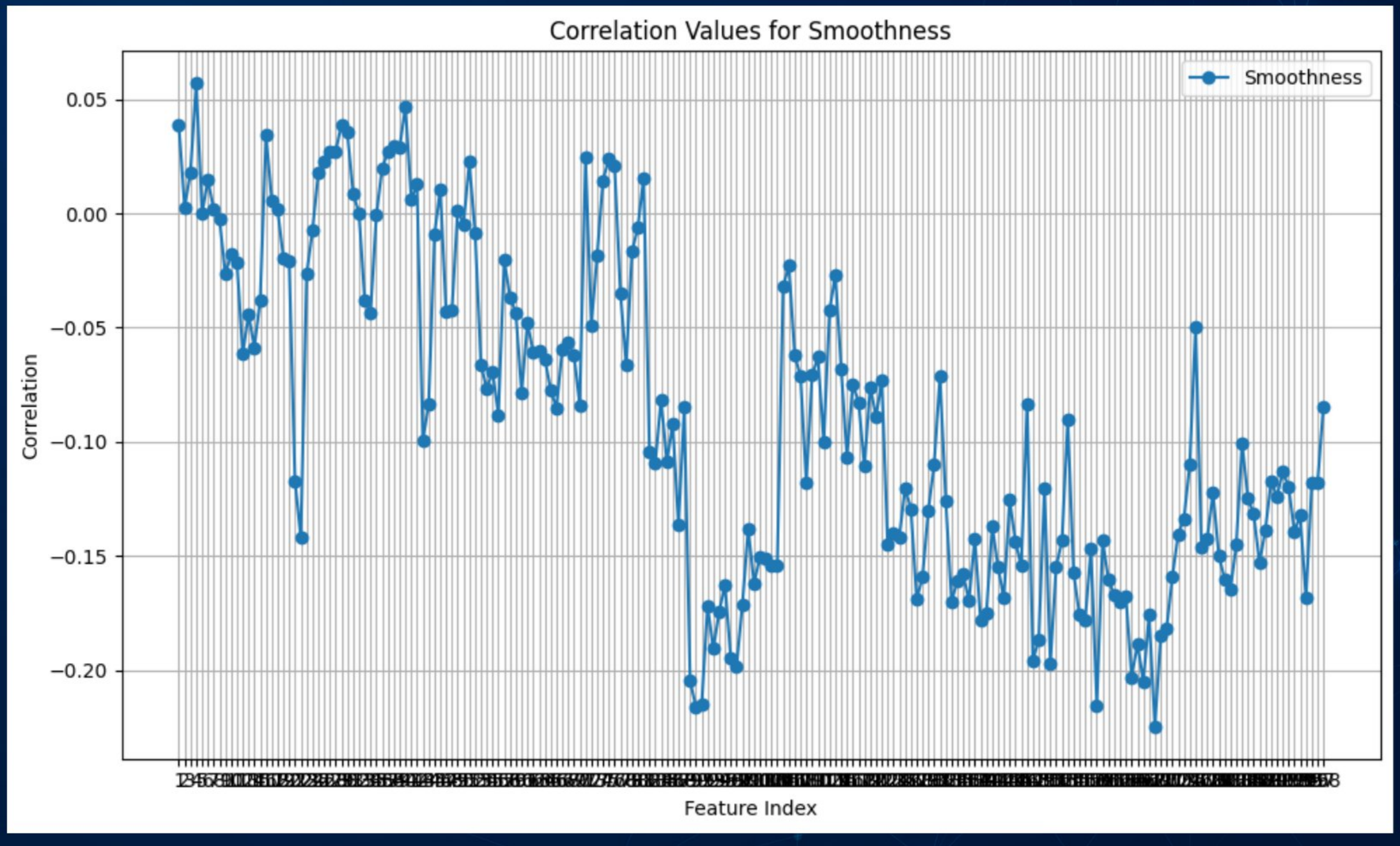
Handling Null Values for the features

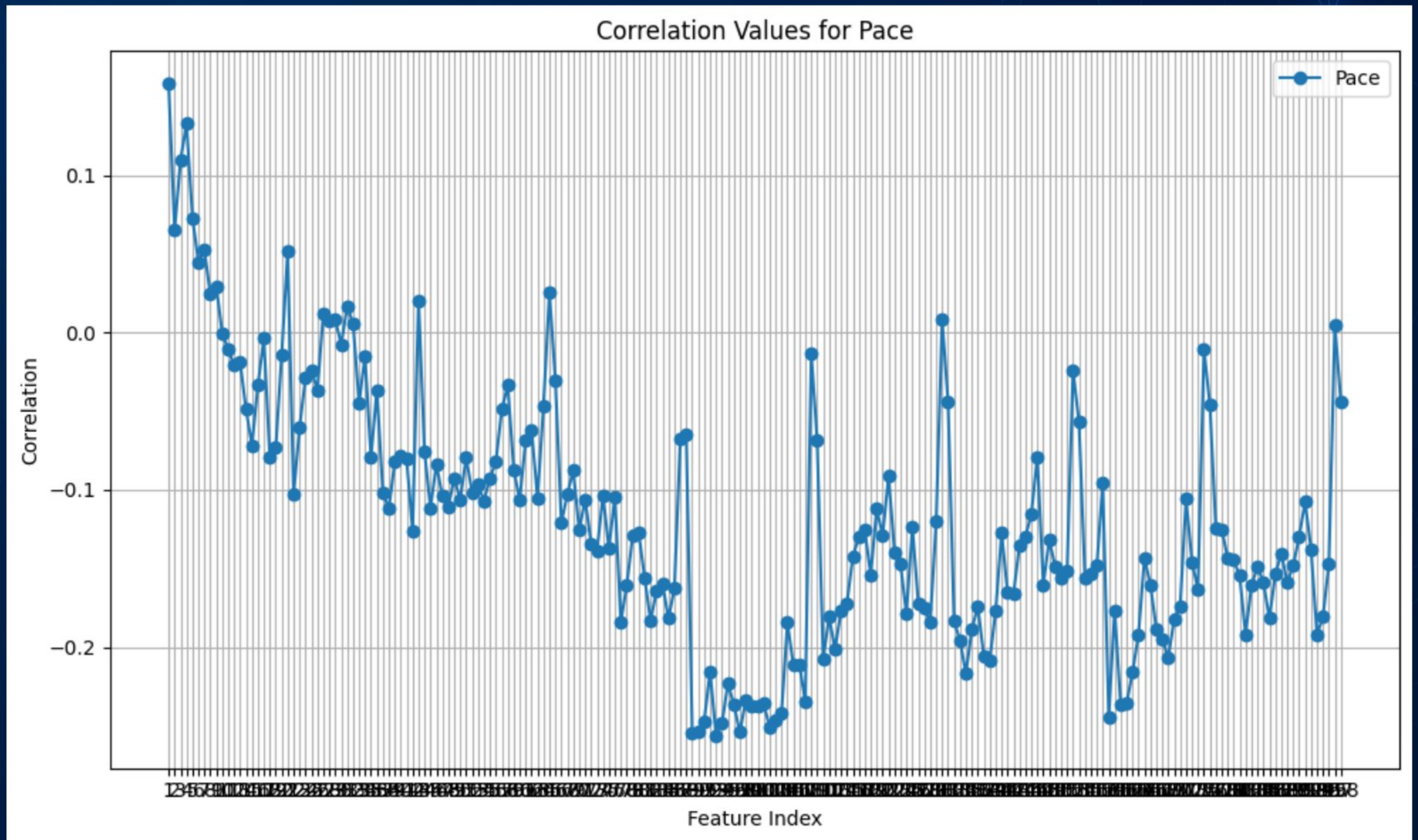
- Imputation Methods: SVM Imputer was used.
- Forward Fill Method

Correlation Graphs


Correlation Values for Expression and Volume
Expression and Volume
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
Correlation
Feature Index


Correlation Values for Phrasing and Intonation
Phrasing and Intonation
Correlation
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
Feature Index

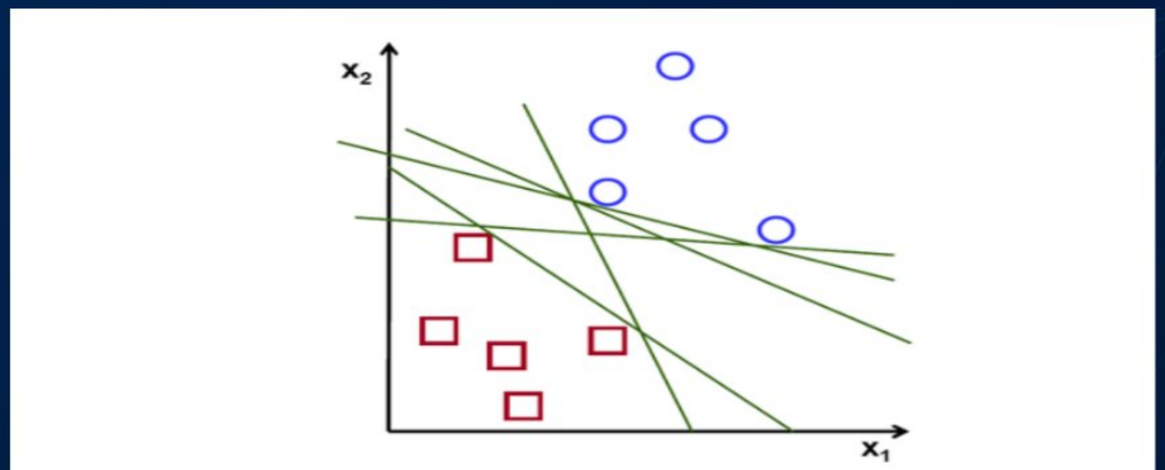

Correlation Values for Smoothness
Smoothness
0.05
0.00
−0.05
−0.10
−0.15
−0.20
Correlation
Feature Index


Correlation Values for Pace
Pace
0.1
0.0
−0.1
−0.2
Correlation
Feature Index

ML Methodologies Used: SVM

What is SVM?

Support Vector Machine (SVM) is a supervised machine learning model used for classification and regression tasks. It identifies the optimal hyperplane that separates different class labels with the largest possible margin, making it effective for complex decision boundaries in high-dimensional spaces.



SVM's goal is to find the "optimal" decision boundary line that can create best separation between the datapoints from two classes. This decision boundary is called a hyperplane.

Content Credit: MLPR Course

ML Methodologies Used: SVM

Why SVM?

High-Dimensional Efficacy: SVM excels in handling large feature vectors, typical in audio processing with features like MFCCs, pitch, and energy.

Margin Maximization: SVM minimizes overfitting through its ability to maximize the margin between differing class data points, crucial for subtle distinctions like fluency levels.

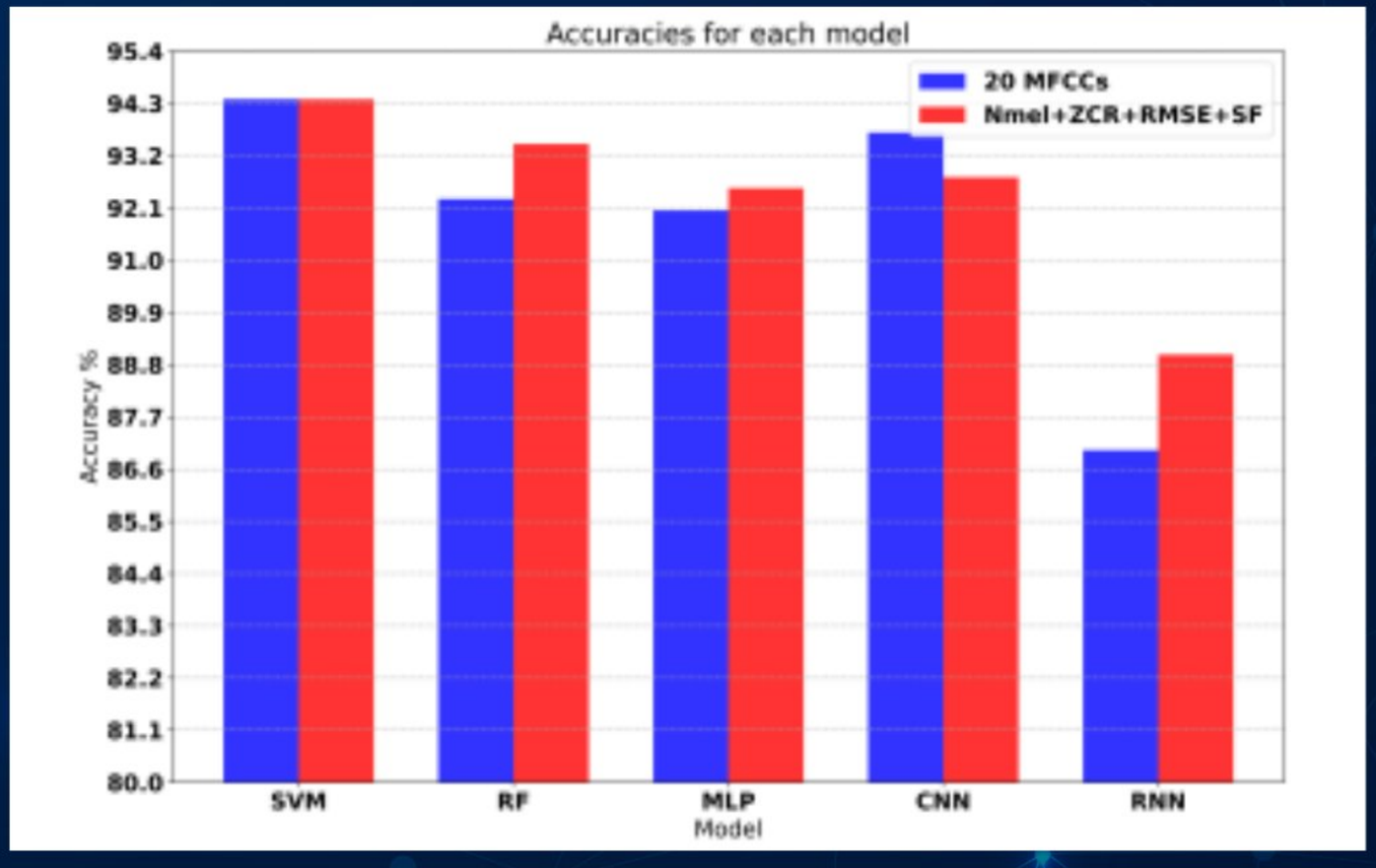
Memory Efficiency: By using only support vectors from the training data for its decisions, SVM reduces memory use, ideal for large datasets.

Overfitting Resistance: SVM's regularization helps prevent overfitting, ensuring the model performs well on new, unseen data.

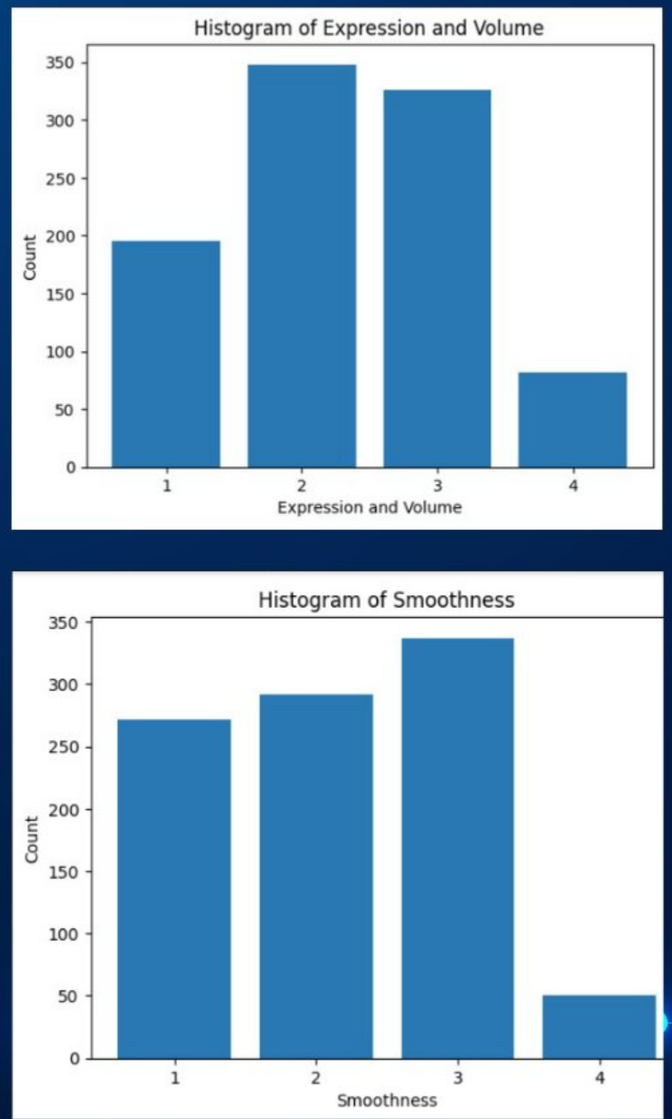
Also, amongst all the papers that we reviewed SVM was used in all of them and managed to get highest accuracy .

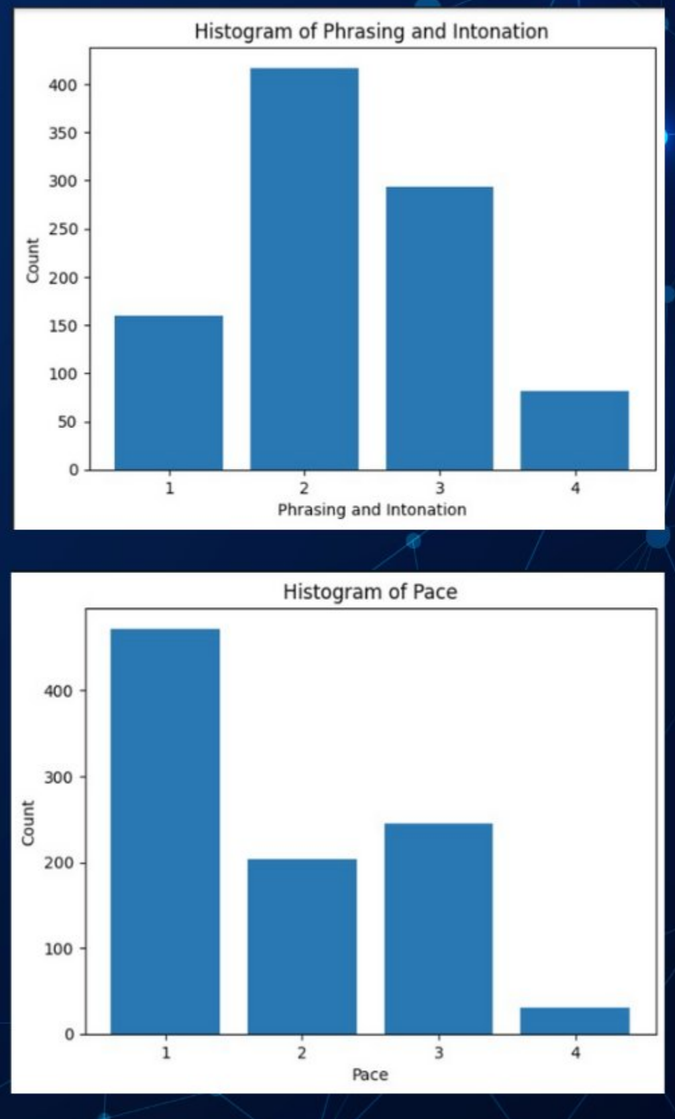
Features	SVM	RF	MLP	CNN	RNN
N_{mel}+ZCR+RMSE+SF	94.39%	93.45%	92.52%	92.75%	89.01%

Table 3: Accuracy performance of the classification models 20 MFCCs + extra features.



Data Distribution





Performance Metrics

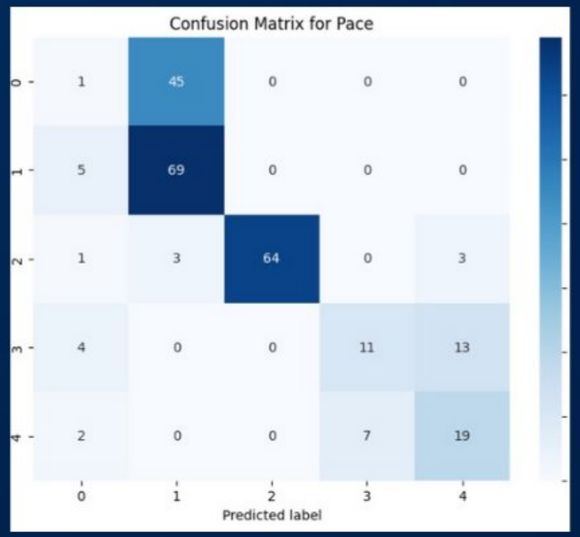
Accuracies

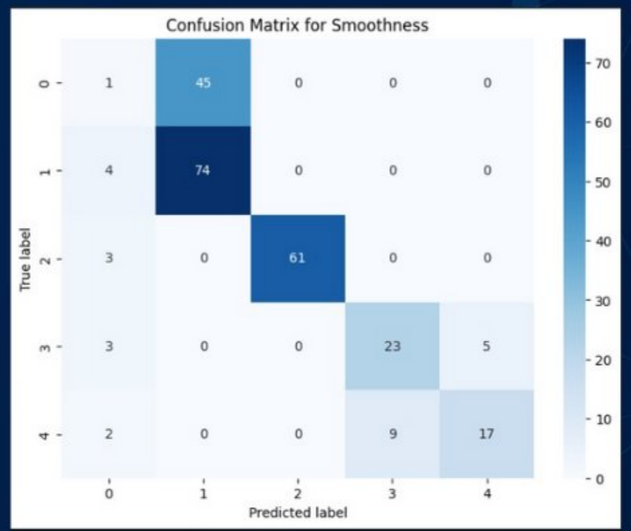
```
Accuracy for Expression and Volume: 0.7611336032388664
Accuracy for Phrasing and Intonation: 0.7651821862348178
Accuracy for Smoothness: 0.7125506072874493
Accuracy for Pace: 0.6639676113360324
```

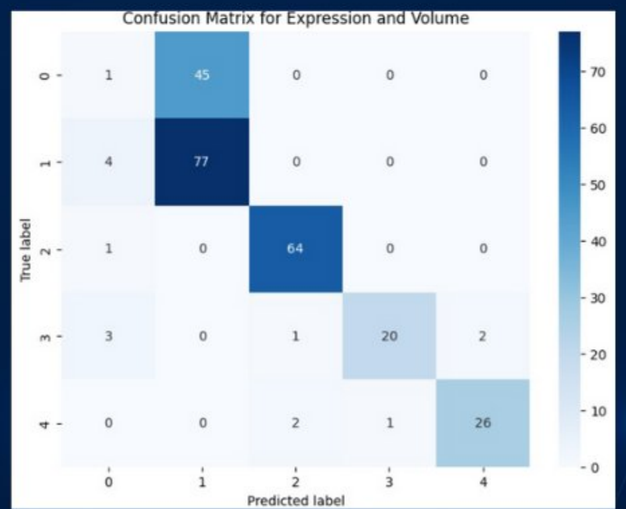
F1 - Scores

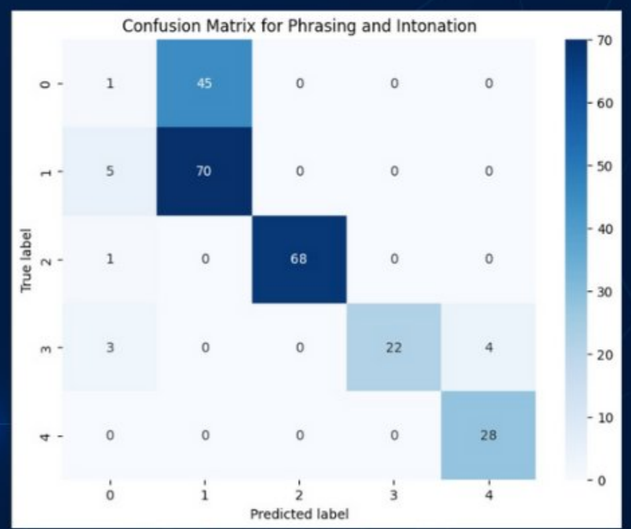
```
F1 Score for Expression and Volume: 0.7056051654514797
F1 Score for Phrasing and Intonation: 0.7122671104553999
F1 Score for Smoothness: 0.6342652141552942
F1 Score for Pace: 0.55719900299559
```

Confusion Matrices









Challenges

- **Data Acquisition:** Limited availability of labeled audio data posed a hurdle in model training.

- **Labeling Process:** Determining appropriate labeling scales and finding experts for accurate labeling were logistical challenges.

- **Feature Selection:** Identifying relevant audio features from a vast pool required extensive exploration.

- **Lack of Research:** Sparse existing literature in reading proficiency analysis using audio data necessitated pioneering efforts.

Solution deployement at Plaksha

1. **Academic Support Services:** Utilize the model to identify students who may require additional support in their academic journey, especially in reading-intensive subjects. This can help tailor tutoring programs and intervention strategies to improve student outcomes.

2. **Language Learning Programs:** Integrate the model into language courses to assess and enhance non-native speakers' proficiency. This could be particularly useful in courses teaching English as a second language, helping to customize learning plans based on individual fluency levels.

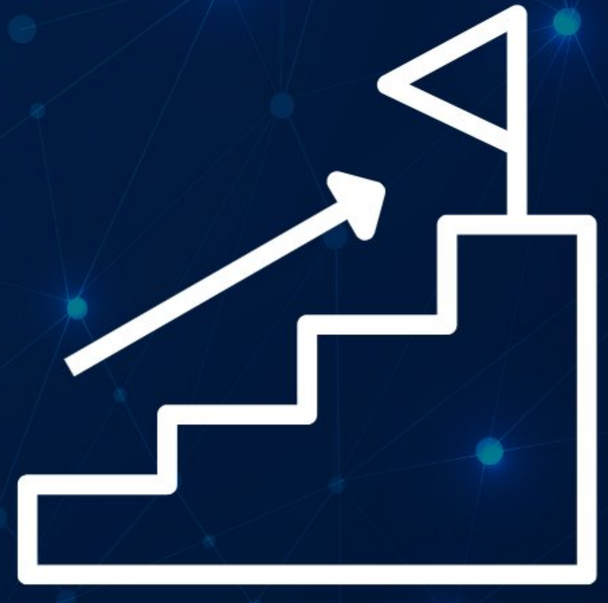
3. **Research Initiatives:** Researchers in linguistics, psychology, and education could use the model to study language acquisition, reading comprehension, and the effects of various teaching methods on reading skills. This could lead to more effective teaching strategies and educational tools.

4. **Admissions Processing:** The model could assist in the admissions process by analyzing applicants' written statements or conducting automated interviews where speech and fluency are evaluated, helping to ensure candidates meet the language proficiency standards of the university.

5. **Career Services:** Use the model within CPC to help students practice and improve their interviewing skills, particularly focusing on aspects like clarity of expression and fluency, which are crucial for job interviews.

What may some challenges be for the deployed solution when it will scale up?

- Training & Customization:
  - Train with diverse samples.
  - Adapt to linguistic nuances.
- Privacy & Consent Framework:
  - Ensure compliance with regulations.
  - Obtain explicit user consent.
- Data Expansion:
  - Continuously incorporate new data.
  - Maintain relevance and effectiveness.
- Model Evolution:
  - Periodic evaluation and adaptation.
  - Consider advanced models like LSTM.



Roadmap For Ahead

Moving forward, we plan to enhance our model's capabilities by training it on Long Short-Term Memory (LSTM) networks. Given the sequential nature of our data—particularly in the context of speech and reading patterns—LSTM models are well-suited for this task.

Their ability to remember and utilize past information makes them ideal for analyzing time-series data, where context and the order of events significantly impact the outcome.

By leveraging LSTMs, we aim to capture these dynamics more effectively, potentially leading to improvements in both accuracy and the applicability of our assessments in real-world settings.

Refernces

- http://roadtocomprehension.com/pdvideo/pdfs/D1_06.pdf
- https://eurasip.org/Proceedings/Eusipco/Eusipco2023/pdfs/0000231.pdf
- https://www.researchgate.net/publication/327392661_Speaker_Fluency_Level_Classification_ Using_Machine_Learning_Techniques
- https://www.researchgate.net/publication/258652444_MFCC_and_Prosodic_Feature_Extractio n_Techniques_A_Comparative_Study
- https://news.un.org/en/story/2021/03/1088392
- https://www.worldbank.org/en/news/press-release/2022/06/23/70-of-10-year-olds-now-in-learning-poverty-unable-to-read-and-understand-a-simple-text
- Morrison, T. G., & Wilcox, B. (2020). Assessing Expressive Oral Reading Fluency. Education Sciences, 10(3), 59. https://doi.org/10.3390/educsci10030059