Project Paper: Vocal Fluency Evaluation Using ML

By: Anant Ghuman, Arkadeep Dutta

Abstract

This project aims to develop a machine-learning algorithm for evaluating the vocal fluency of English learners. The system listens to the learner's speech, processes the audio data, and provides a fluency rating on a scale of 0-2, with 0 representing a novice speaker and 2 indicating an expert speaker. The primary objective is to assist English learners in improving their pronunciation skills by offering objective feedback. The algorithm combines Support Vector Machine (SVM) and Long Short-Term Memory (LSTM) models to achieve its classification task. The project demonstrates promising results, achieving an accuracy of approximately 83.5% on the test set. However, further improvements are possible with a larger and more diverse dataset.

Introduction

Fluency in a language is a crucial aspect of effective communication. For English learners, achieving vocal fluency can be challenging, and accurate evaluation of their progress is essential for improvement. Traditionally, human evaluators have manually judged vocal fluency, leading to subjectivity and inconsistency. This project seeks to address this limitation by developing a machine-learning algorithm capable of objectively assessing the fluency of English learners.

The system uses a dataset comprising audio recordings of non-native English speakers, categorized into three fluency levels: Novice, Intermediate, and Expert. Native English speaker audio recordings serve as reference data to evaluate the learners' pronunciation. By employing advanced machine learning techniques, the algorithm analyzes various features of the learners' speech, including phonetic accuracy, stress patterns, intonation, and overall fluency.

Related Works

This project builds upon the principles learned in class about machine learning and natural language processing. In particular, classification algorithms (such as SVM), and deep learning models (such as LSTM) have been utilized to achieve accurate vocal fluency assessment, and the LSTM and SVM classes created in the class were initially used for the sake of the project. Additionally, lessons in how to calculate various metrics such as precision, recall, etc were used in the final presentation and were also used to analyze the model during training.

Previous research in automatic speech recognition (ASR) and speaker identification has provided valuable insights into audio analysis and signal processing techniques. While not directly applicable to vocal fluency evaluation, these works have paved the way for the development of similar machine-learning systems that deal with audio data. This project draws inspiration from these related works to create a robust and objective vocal fluency evaluation system.

Methodology

Our approach to developing the vocal fluency evaluation model involved several key steps. First, we took audio input and preprocessed it by splitting the audio into five-second chunks and removing background noise. Next, we extracted various features such as MFCCs, spectral contrast, ZCR, RMSE, spectral flatness, and spectrogram frames to aid in classification.

The data was then divided into training and testing datasets. We optimized the training dataset to be compatible with model architectures like Convolutional Neural Networks (CNN) and evaluated different models, ultimately selecting SVM as it showed the highest accuracy. Additionally, we chose to incorporate LSTM models to improve the identification of certain types of audio files.

To assess the models' performance, we used the testing dataset and obtained predictions from SVM and reshaped features for LSTM. We averaged the SVM outputs with LSTM predictions for a more comprehensive evaluation. The averaged outputs were used to classify vocal fluency into low, medium, or high levels, and this process was repeated on every 5-second chunk until the inputted recording ended. The classifications were then all averaged into one big classification, providing learners with valuable feedback for improvement.

Results

In conclusion, our project successfully developed a machine learning algorithm to assess the vocal fluency of English learners quickly and efficiently. By combining Support Vector Machine (SVM) and Long Short-Term Memory (LSTM) models, we achieved the best results, handling both temporal data and spectrogram frame analysis effectively. However, there are important points to consider and areas for improvement in future work.

One limitation of our current model is its bias towards certain accents due to the limited diversity in our training dataset. To fix this, it's crucial to expand the dataset with audio samples from speakers with different accents, so the model can better understand various pronunciation styles.

Our model's performance is weaker when evaluating speakers with low to medium fluency levels. To enhance its accuracy in distinguishing between these levels, we can add more examples of learners with such fluency levels to our dataset. This way, the model can better learn the subtle differences between these classes.

Expanding the classification into more specific fluency levels, like Beginner, Advanced Beginner, and Advanced, would improve the precision of our system and cater to learners at different language proficiency stages.

Although our model achieved reasonable results, we can further improve accuracy by fine-tuning hyperparameters and optimizing feature extraction. Additionally, we should consider adding prosody and rhythm features for a more thorough evaluation of vocal fluency.

For future work, gathering more labeled data from a diverse set of speakers, introducing additional fluency classes, and expanding our feature set will enrich the model's understanding of vocal fluency. Extending our system to evaluate fluency in other languages would also be valuable in supporting language learners worldwide. Giving learners feedback on specific aspects of their vocal fluency, such as stress patterns or pronunciation of certain sounds, can offer targeted guidance for improvement. Implementing a chorusing evaluation function can assess learners' ability to speak in unison with a native speaker, encouraging better pronunciation and accent alignment.

Conclusion

The project successfully implements a machine learning algorithm for evaluating the vocal fluency of English learners. By providing objective and accurate fluency ratings, the system offers learners valuable feedback to improve their pronunciation skills effectively. The combination of SVM and LSTM models demonstrates a robust approach to handling audio data and speech temporal patterns.

However, additional data is necessary to further enhance the accuracy and generalizability of the algorithm. A more extensive and diverse dataset of non-native English speakers would enable the model to better distinguish between different fluency levels. Moreover, the inclusion of features like prosody, pitch, and rhythm could offer further insights into speech evaluation.