Speaker Fluency Level Classification Using Machine Learning Techniques

A. Preciado Grijalva[1] , R. F. Brena[1]

[1] Grupo de Investigación en Sistemas Inteligentes, Tecnológico de Monterrey, México

Introduction

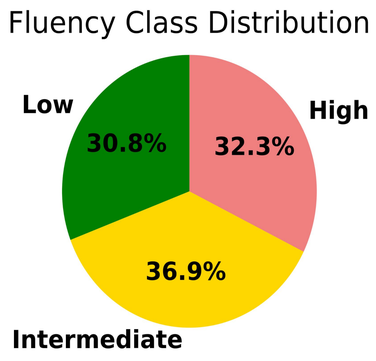
Level assessment for foreign language students is necessary for putting them in the right learning group, but it is also a very time-consuming task, so we propose to automate the evaluation of speaker fluency level by implementing machine learning techniques.

This work presents an audio processing system capable of classifying the level of fluency of non-native English speakers. Our approach to this problem is based on audio analysis, starting with audio feature extraction and afterwards training machine learning models to perform classification of audio segments provided labeled target classes.

Audio Dataset

The **Avalinguo Audio Dataset** is a labeled collection of 1424 audio recordings of non-native English speaking persons. This public dataset was constructed to evaluate the classification performance of the analyzed machine learning models. [2]

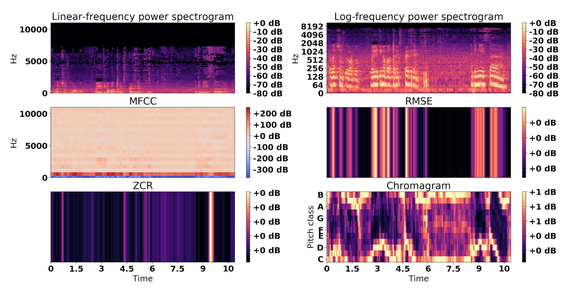
- Total 1424 audio files of non-native English speakers
- 5s non-overlapped segments
- Three classes: low, intermediate, high
- ~2 hours of recordings
- 450 audio clips per class (approx.)
- Sample rates 22050 Hz to 48000 Hz
- MP3 format

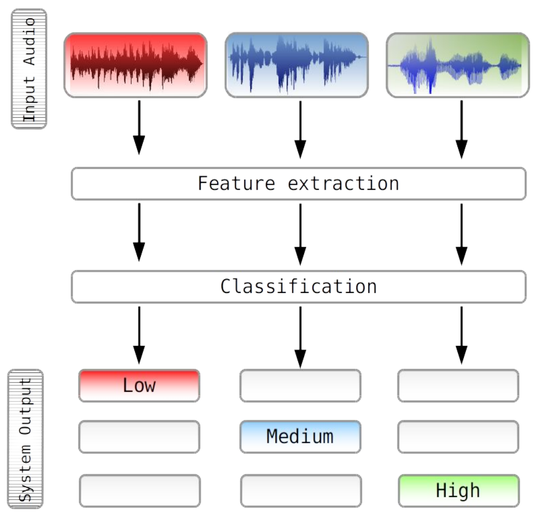


Experimental Framework

Our experimental procedure is the standard ML approach for analysis of sound events [3].

Feature Extraction





Audio Classification

We have compared five different classification frameworks to compare their performance as we vary the number of extracted features:

- Multilayer perceptron
- Convolutional neural network
- Recurrent neural network
- Support vector machine
- Random forest

Neural Network	Hidden layers	Neurons	Activation Function
MLP	2	512x512x3	relu, softmax
CNN	4	64x64x32x32x3	relu, softmax
RNN	2	256x32x3	relu, softmax

Experimental Results

As a first approach, we have trained and evaluated accuracy performance with increasing values of mel coefficients (MFCCs).

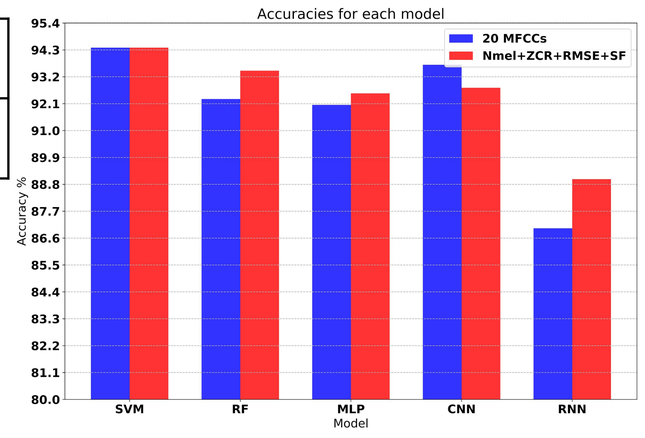
The graph contains the achieved accuracies as we set MFCCs = 5, 10, 12, 20. In each case, the accuracy improved considerably as the MFCCs increased.

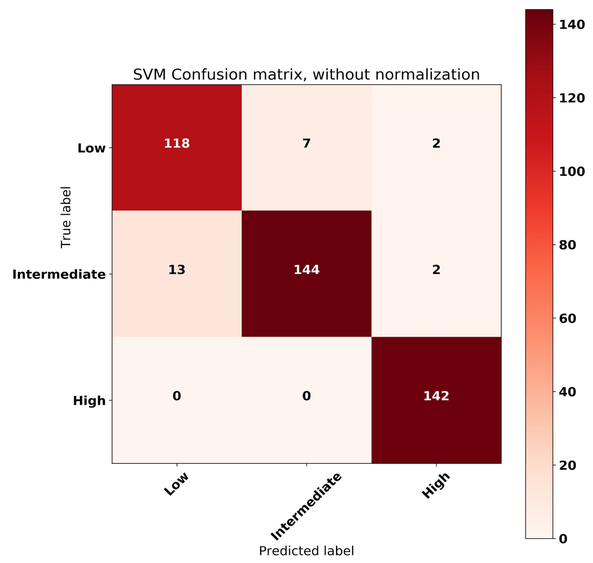
We obtained an accuracy as high as 94.39% with the SVM for MFCCs = 20.

Feature	SVM	RF	MLP	CNN	RNN
Nmel + ZCR + RMSE + SF	94.39%	93.45%	92.52%	92.75%	89.01%

Accuracy performance with 20 MFCCs + extra features. The table above shows the accuracy results for each of our models.

There is an increase in performance in most of our models. The bar plot presents the accuracy results before and after adding extra features.





In order to evaluate the quality of the output of the classifiers we have used a confusion matrix. Here we show a map corresponding to the SVM prediction results.

This model, trained to classify among our three fluency levels, has achieved a classification accuracy of 94.39%.

Conclusions

- We have developed an audio processing system capable of determining the level of fluency of non-native English speakers using the Avalinguo audio dataset.
- We have used five different ML models to classify audio segments into low, intermediate or high fluency levels. Each model was capable of classifying audio frames with an accuracy of more than 90% (except one classifier that reached 89%).
- This project will be integrated to the bigger Avalinguo system. Here, we will have to deal with the classification of live conversations, for example.

References

[1] Link to project decription: https://tec.mx/es/noticias/nacional/investigacion/realidad-virtual-para-popularizar-el-aprendizaje-de-idiomas
[2] Github dataset repository: https://github.com/agrija9/Avalinguo-Audio-Set
[3] Maxime Jumelle and Taqiyeddine Sakmeche. *Speaker Clustering With Neural Networks And Audio Processing*. arXiv: 1803.08276v1 (2018).