Speech Fluency Features for Robust Automatic Classification of L2 English Speech using a Small Dataset

Joaquín Jordán
Department of Computer Science
UTEC
joaquin.jordan@utec.edu.pe

Renato Cernades
Department of Computer Science
UTEC
renato.cernades@utec.edu.pe

José Chachi
Department of Computer Science
UTEC
jose.chachi@utec.edu.pe

Abstract

This work presents an automatic fluency level classifier of L2 English speech. Using models such as Support Vector Machines, and a Multilayer Perceptron, comparing features of three previous works [3], [20], [12] such as speech rate, filled-pauses, phonation ratio, effective speech rate, and the well know Mel Frequency Cepstral Coefficients (MFCC). Furthermore, we utilize the Avalinguo dataset, which [12] proposed. It comprises 1424 audio samples of individuals speaking, categorized into three fluency levels: Basic, Intermediate, and Advanced. Additionally, we introduce two experiments that demonstrate the robustness of some of these features and models. Our findings indicate that SVM with Yu & Van Heuven [20] features provides the highest accuracy and demonstrates the most robustness among all tested models and features.

1 Introduction

1.1 Motivation

The growing demand for learning English as a second language (L2) has led to an increasing interest in automatic spoken language assessment, whether for use in computer-assisted language learning (CALL) tools or for grading candidates for formal qualifications [6]. In particular, oral fluency is viewed as an important characteristic of second language speech, and is an important object of evaluation in testing second language skill [17]. According to the bibliography review shown in [15], there are multiple definitions of fluency. For instance, Richards points out that fluency is the use of naturally occurring language when a speaker engages and maintains meaningful communication. According to Fillmore, a fluent speaker knows what to say and how to say it without frequent pauses to think. On the other hand, Harmer mentions that fluency refers to focusing on the content of speech to communicate as effectively as possible. Finally, Baily defines fluency as using language quickly and confidently, with limited hesitations, unnatural pauses, etc. For this work, we will be focusing on the definition of fluency as proposed by Baily.

Final Paper for the ML3061-2023.2 - Machine Learning Class at UTEC. Lima, Peru.

1.2 Related work

There are multiple works performing English fluency classification in L2 speech. On the one hand, Kobayashi & Wilson [5] use deep RNNs with MFCC features for the task with a fairly small dataset containing 1139 2-second files for training and achieves an accuracy of 25.9% on their first experiment, which is slightly better than random chance (20%). Their results show that using MFCCs with a small dataset can lead to low accuracies. On the other hand, Preciado-Grijalva & Brena [12] compare different classification models such as SVMs, RF, MLP, CNNs, and RNNs. They use MFCC features and a dataset containing 1424 5-second files called the Avalinguo dataset, which will also be used in the present work. Their most accurate classifier achieves 94.39% accuracy. Finally, Deshmukh et al. [3] use an SVM with their proposed lexical and prosodical features. Their model is trained on a dataset with 112 1-minute audios and achieves an accuracy of 71.43% when classifying the speakers in 3 fluency levels.

Concerning audio classification tasks in general, Phan et al. [10] and Piczak [11] have successfully used MFCCs with deep CNNs and RNNs. However, their datasets are considerably bigger than the Avalinguo dataset. As such, in developments such as Valdiviezo Mora et al. [19] and Preciado-Grijalva & Brena [12] that use this dataset, which is small, the preferred choice is to use simpler Machine Learning approaches, such as SVMs or MLPs.

1.3 Dataset

For this work, we will use the Avalinguo data set proposed in [12]. It consists of 1424 five-second long audio fragments which are distributed in three categories: 438 of them are basic, 527 intermediate, and 459 advanced.

1.4 Our approach

What we find particularly intriguing is the high accuracy found in the work of Preciado-Grijalva & Brena [12] using MFCCs. We believe that this result might be affected by the particular patterns found within the Avalinguo dataset since machine learning methods are exceedingly adept at finding patterns (which in some cases are brittle and spurious) to boost performance on held-out data from the same dataset [14]. Our experimentation shows that models trained using MFCCs in a modified version of the Avalinguo dataset that exploit patterns present within it are prone to misclassify held-out data. Furthermore, we propose alternative features found in the literature that are more robust in this same scenario to build a classifier with better generalization.

2 Robust characterizations of fluency

Plenty of research has been done to quantitatively measure English speech fluency with the thought of future applications in automatic classification.

A first approach to quantitatively measure fluency was done in Towell et al. [18], where they use the model of Levelt [7] that provides the descriptive base for the sub-processes of language production. They implemented it with *temporal variables* as correlates of fluency which can measure the global level of fluency in French speech and the contribution of the sub-processes in the model.

Temporal variables	
1	Articulation rate
2	Speech rate
3	Mean length of fluent runs
4	Phonation/time ratio

Table 1: Features according to [18]

Yu & Van Heuven [20] builds upon the work of Towell et al. [18] to make an experiment whose main goal was to determine whether judged fluency in English can be predicted from computer-based measurements such as articulation rate. To do so, the work utilizes additional acoustic measures of fluency. These are separated into the following groups: speed fluency (1, 2), breakdown fluency (2, 4,

5, 6, 7, 8, 9, 11, and 12), repair fluency (10), or all three categories (3). Their study concludes that effective speech rate (3) appears to be the best predictor for judged fluency. From this point onward, we will be referring to these acoustic measures of fluency as just *features*, and we will consider all twelve of these features in this study.

Acoustic correlates of fluency	
1	Articulation rate
2	Speech rate
3	Effective speech rate
4	Number of silent pauses above 0.25 seconds in duration
5	Mean length of silent pauses longer than 0.25 seconds
6	Number of filled pauses (uh, er, mm, etc)
7	Mean length of all filled pauses
8	Number of pauses = (4)+(6)
9	Mean length of pauses
10	Number of other disfluencies
11	Mean length of fluent runs
12	Phonation/time ratio

Table 2: Features according to [20]

Likewise, Deshmukh et al. [3] evaluate 16 features separated into two groups: prosodic and lexical features. Prosodic features (see 3) are computed directly from the speech and are used to detect disfluencies such as filled pauses. Meanwhile, lexical features (see 4) are computed from the speech's text. An important thing to take into account is that lexical features are based on the transcription of the speech, particularly, features (l.d), (l.g) and (l.h) are very sensitive to the quality of the transcription. In other words, lexical feature extraction highly depends on the quality of ASR (Automatic Speech Recognition). In this work, we automated the extraction of two features using NLP techniques: filled pauses (CFP) as described in [13], and ClUni which can be interpreted as stuttering as described in [16]. However, automatically extracting ClTri requires reparandum detection as shown in [4], and was not implemented in this work.

Prosodic features		
p.a	Avg. number of filled pauses per sec.	AvgFP
p.b	Avg. duration of a filled-pause	DurFP
p.c	Avg. distance between filled-pauses	DistFP
p.d	Length of the longest filled pause	MaxFP
p.e	Fraction of silence	FracSIL
p.f	Avg. duration of contiguous silence	DurSIL
p.g	Avg. duration of contiguous speech	DurSP
p.h	Avg. distance between silences	DistSIL

Table 3: Prosodic features by Deshmukh et al. [3]

3 Experimentation

In this section, we evaluate the effectiveness of these features in fluency classification. The code for this section can be found in our GitHub repository A. The main goals are as follows:

1. **Feature Efficacy**: Evaluate the performance of the models using Yu & Van Heuven [20] and Deshmukh et al. [3] features in capturing relevant fluency-related information, surpassing the capabilities of conventional methods such as MFCC used in [12].
2. **Comparison with Previous Work**: Conduct a comparative analysis with existing fluency classification techniques and highlight the advancements achieved.

Lexical features		
l.a	Count of most frequent word	FreqW
l.b	Total words	TW
l.c	Total unique words	TUW
l.d	Count of filled-pauses	CFP
l.e	Count of dictionary words	Cwrd
l.f	Total repeated 'similar' trigrams	RepTri
l.g	No. of closely occurring unigrams	ClUni
l.h	No. of closely occurring similar trigrams	ClTri

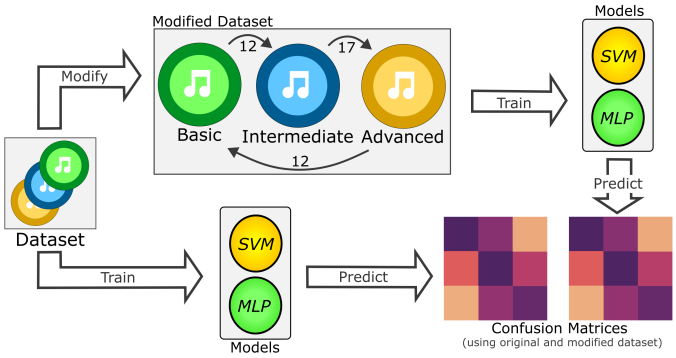
Table 4: Lexical features by Deshmukh et al. [3]

3.1 Feature Extraction

For MFCC, these features were extracted using the Librosa Python library [1]. Additionally, Yu & Van Heuven [20] features were extracted using a Python script that computed features such as the duration, number of words (transcribed using Faster-Whisper Python for audio transcription [2]), and duration of filled pauses (these were captured using a regex expression to identify filler sounds). In the same way, we also compute the Prosodic and Lexical features [3].

3.2 Impact on accuracy

In the first experiment, the process involved modifying the dataset by altering certain labels within the data. Specifically, we selected 12 segments of audio from individuals at the basic level and moved them to the intermediate level. Similarly, we transferred 17 audio segments from the intermediate level to the advanced level and 12 from the advanced level to the basic level. Besides, we extract the MFCC, Yu & Van Heuven [20], and lexical and prosodic features [3] from the dataset. Then, we train two models: Support Vector Machine (SVM) and the fully-connected neural network Multilayer Perceptron (MLP), using the modified and original dataset. Finally, we compute the confusion matrices for each model and dataset. In the following figure, the experiment's pipeline is illustrated.



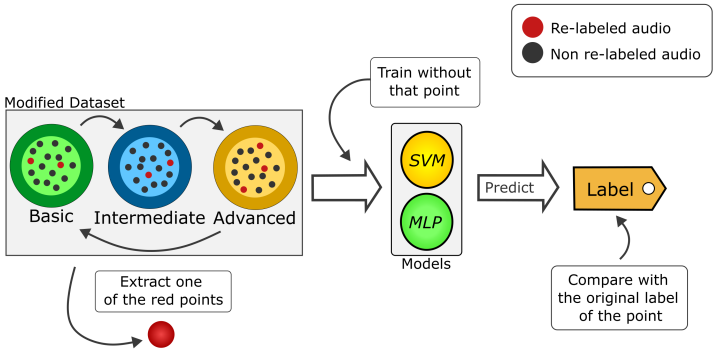
Experiment 1: The impact of modifying labels on accuracy

After implementing these changes, we anticipate a shift in the model's accuracy, even if only slightly. This expectation arises from the fact that we are intentionally introducing confusion by re-labeling the audios, such as those originally categorized as advanced now being labeled as basic.

3.3 Evaluation of robustness

In this second experiment, the process consists of extracting one re-labeled audio from the modified dataset. More specifically, we will be extracting the audio segments corresponding to *"Luis Suarez interview in English after being awarded November player of the month segment x - W"*, an advanced audio re-labeled as basic. Using the same models and features as in the first experiment, we attempt to predict the fluency level of the extracted audio and compare it with the original. This process will

be repeated for each re-labeled audio in the modified dataset. Once again, the experiment's pipeline is illustrated.



Experiment 2: Testing the robustness of the models and features

As a result, we expect to achieve the original fluency level of the audio. Although we are using a modified dataset, the model should be robust enough to make accurate predictions.

4 Results

The results of the first experiment (see 3 and 4) show that the impact on accuracy of modifying the dataset in both of the models is minimal. However, the SVM model (see 3) yields better classification results for practical applications since it rarely confuses the basic and advanced classes. Meanwhile, the MLP model (see 4) indicates the presence of noise in the labels because the loss (see 5) converges at a high value. Additionally, the accuracy of the models trained with our features is low even in the original dataset, while struggling to identify the intermediate category in particular. We believe that this is due to two factors: 1) the precision of the ASR system used in our feature extraction methods (faster-whisper), which we found can sometimes err upon manual inspection of the features, and 2) the label noise present in the dataset since we disagree with the judged fluency of some audios.

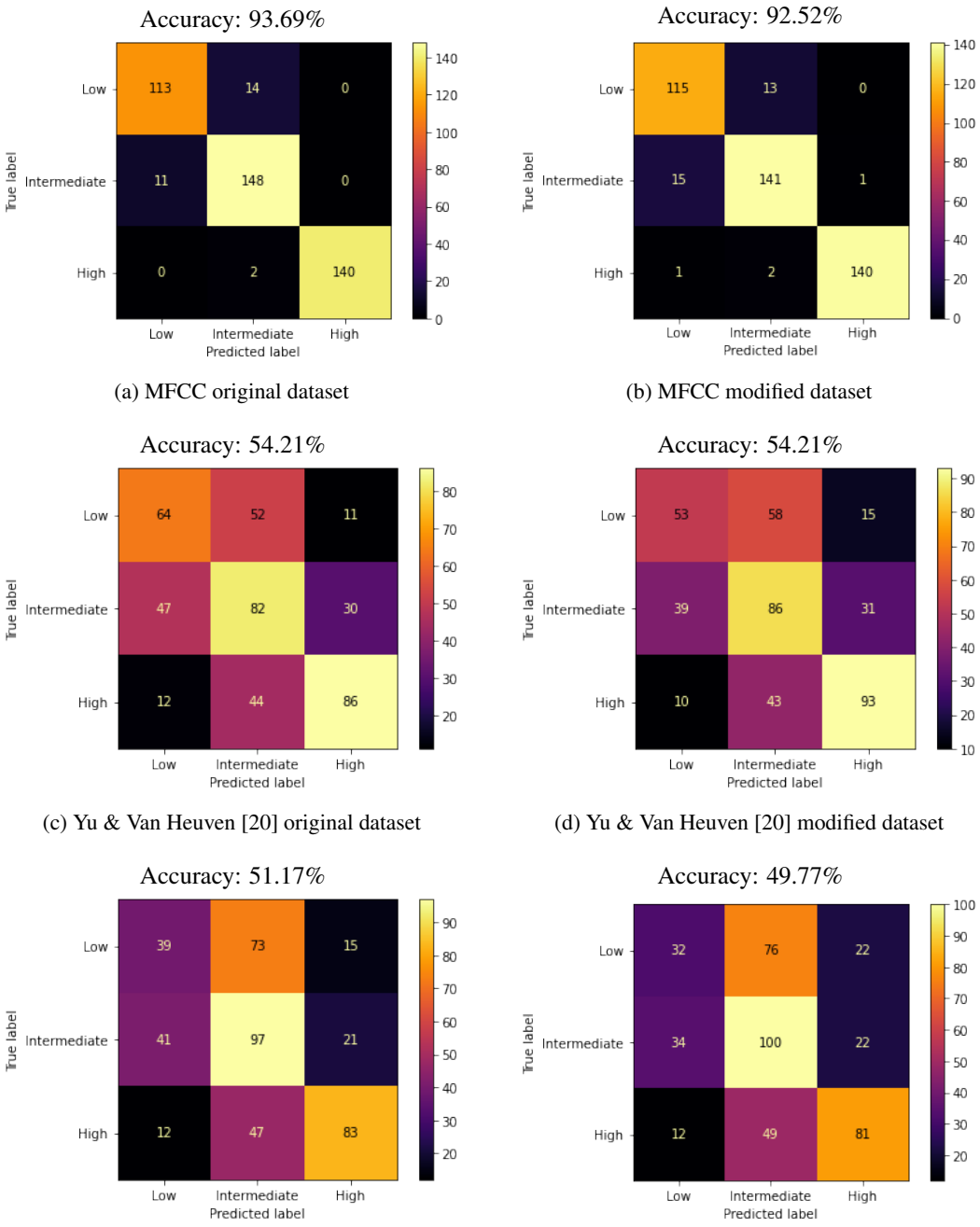
Level	MFCCs	Yu & Van Heuven [20]	Deshmukh et al. [3]
Basic	9	0	0
Intermediate	3	0	3
Advanced	0	12	9

Table 5: Held-out data point classification (12 advanced audios that were re-labeled as basic)

The results of the second experiment (see 5) show that the proposed features are more robust. This implies that the measures of fluency from Yu & Van Heuven [20] are the most adept for the task. Furthermore, these results indicate the presence of a particular pattern in the dataset that is being exploited by the SVM classifier trained with MFCC features [12] to achieve higher accuracies. Since the dataset consists of audio recordings cut into 5-second fragments, each class has multiple audios of the same speaker in the same settings (environmental noise, microphone quality, etc), and this combination of speaker and settings does not appear in other classes. As such, the resulting model predicts almost all of the 12 advanced audios that were re-labeled as basic (and held out one at a time) to be basic. Some of them were predicted as intermediate, and upon inspection of the dataset we found that in this class there are fragments coming from the same original recording as the re-labeled data points.

5 Conclusions and future work

We have successfully compared the performance of different speech fluency features proposed in the literature in developing a robust automatic fluency classifier given the constraint that our dataset is particularly small. Our experiments show that the classifiers trained with our features, which are



(a) MFCC original dataset

(b) MFCC modified dataset

(c) Yu & Van Heuven [20] original dataset

(d) Yu & Van Heuven [20] modified dataset

(e) Deshmukh et al. [3] original dataset

(f) Deshmukh et al. [3] modified dataset

Figure 3: Confusion matrices using different features with the SVM model trained in the original (left), and modified (right) datasets

tailored to measure a speaker's fluency, are indeed more robust. However, it was also shown that the classification accuracy is rather poor, given that our best (most robust and accurate) model was the SVM trained with the features proposed in Yu & Van Heuven [20], giving an accuracy of 54.21% on the original dataset. In order to improve the performance of our classifier without building a new dataset, a label noise correction technique could be used [9], since we believe there is an important amount of noise in the class labels, and the effect of noise on accuracy is widely acknowledged [21], [8].

References

[1] Librosa 0.10.1. URL `https://pypi.org/project/librosa/`.

[2] Faster-whisper. URL `https://github.com/SYSTRAN/faster-whisper`.

[3] Om D Deshmukh, Kundan Kandhway, Ashish Verma, and Kartik Audhkhasi. Automatic evaluation of spoken english fluency. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4829–4832. IEEE, 2009.

[4] Julian Hough, David Schlangen, et al. Recurrent neural networks for incremental disfluency detection. 2015.

[5] Aozora Kobayashi and Ian Wilson. Using deep learning to classify english native pronunciation level from acoustic information. In *SHS Web of Conferences*, volume 77, pp. 02004. EDP Sciences, 2020.

[6] Konstantinos Kyriakopoulos. *Deep learning for automatic assessment and feedback of spoken english*. PhD thesis, University of Cambridge, 2022.

[7] Willem JM Levelt. *Speaking: From intention to articulation*. MIT press, 1993.

[8] David F Nettleton, Albert Orriols-Puig, and Albert Fornells. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial intelligence review*, 33:275–306, 2010.

[9] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1944–1952, 2017.

[10] Huy Phan, Philipp Koch, Fabrice Katzberg, Marco Maass, Radoslaw Mazur, and Alfred Mertins. Audio scene classification with deep recurrent neural networks. *arXiv preprint arXiv:1703.04770*, 2017.

[11] Karol J Piczak. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*, pp. 1–6. IEEE, 2015.

[12] Alan Preciado-Grijalva and Ramon F Brena. Speaker fluency level classification using machine learning techniques. *arXiv preprint arXiv:1808.10556*, 2018.

[13] Anna Pylypiuk. Filled pauses in learner and native english. 2019.

[14] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023.

[15] Gholamhossein Shahini and Fatemeh Shahamirian. Improving english speaking fluency: The role of six factors. *Advances in Language and Literary Studies*, 8(6), 2017.

[16] Shakeel A Sheikh, Md Sahidullah, Fabrice Hirsch, and Slim Ouni. Machine learning for stuttering identification: Review, challenges and future directions. *Neurocomputing*, 2022.

[17] Helmer Strik and Catia Cucchiarini. Automatic assessment of second language learners' fluency. 1999.

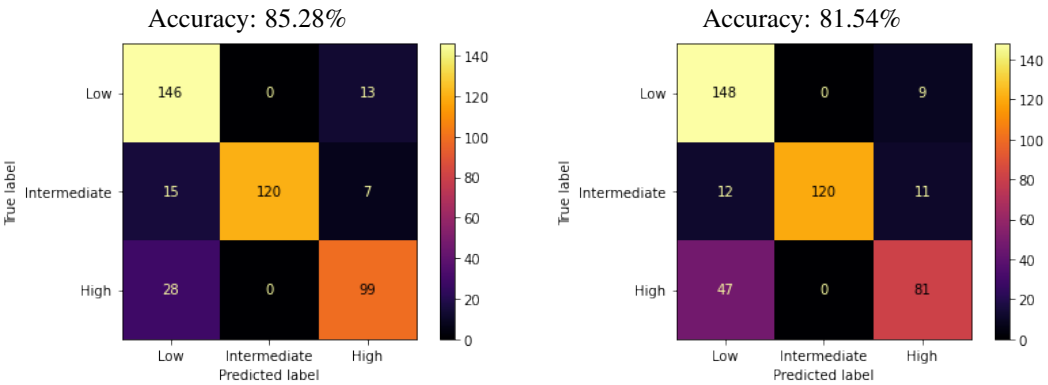
[18] Richard Towell, Roger Hawkins, and Nives Bazergui. The development of fluency in advanced learners of french. *Applied linguistics*, 17(1), 1996.

[19] José Aristh Valdiviezo Mora et al. Identification of pronunciation errors in l2 english speech by spanish speaking natives for s-impure sounds. 2019.

[20] Wenting Yu and Vincent J Van Heuven. Predicting judged fluency of consecutive interpreting from acoustic measures: Potential for automatic assessment and pedagogic implications. *Interpreting*, 19(1):47–68, 2017.

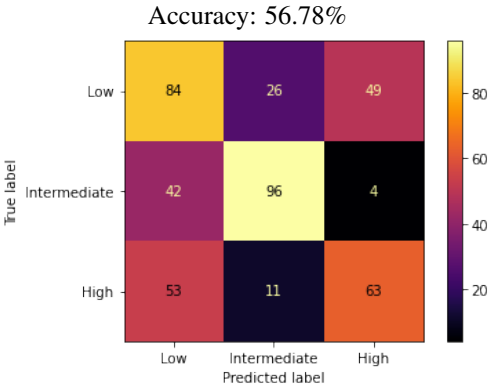
[21] Xingquan Zhu and Xindong Wu. Class noise vs. attribute noise: A quantitative study. *Artificial intelligence review*, 22:177–210, 2004.

A Appendix

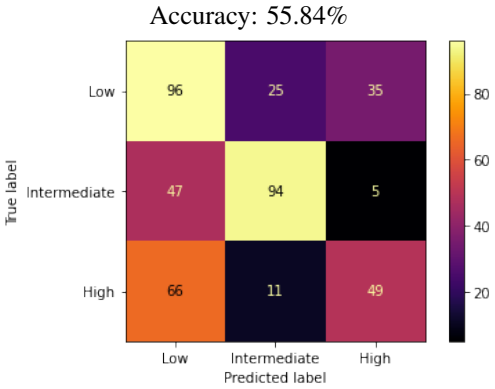


(a) MFCC original dataset

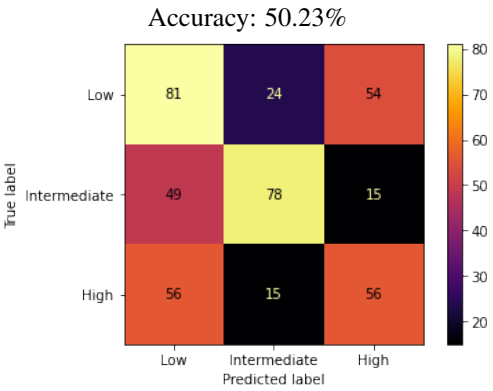
(b) MFCC modified dataset



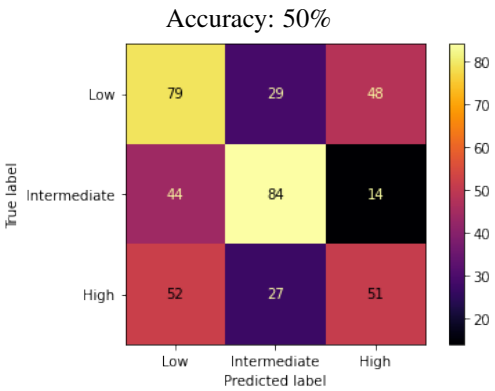
(c) Yu & Van Heuven [20] original dataset



(d) Yu & Van Heuven [20] modified dataset



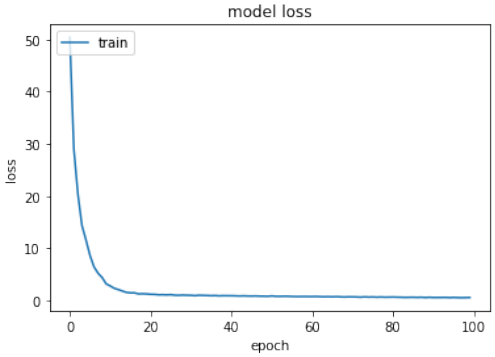
(e) Deshmukh et al. [3] original dataset



(f) Deshmukh et al. [3] modified dataset

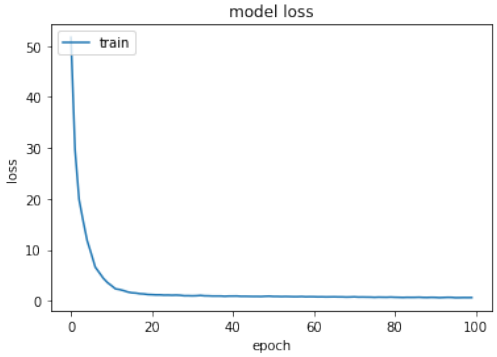
Figure 4: Confusion matrices using different features with the MLP model trained in the original (left), and modified (right) datasets

Final loss: 0.6021



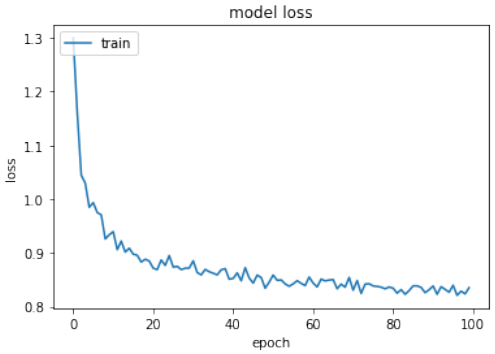
(a) MFCC original dataset

Final loss: 0.6642



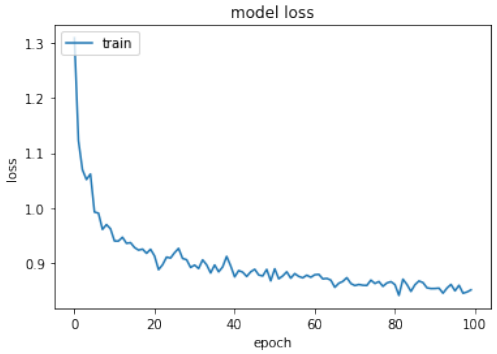
(b) MFCC modified dataset

Final loss: 0.8353



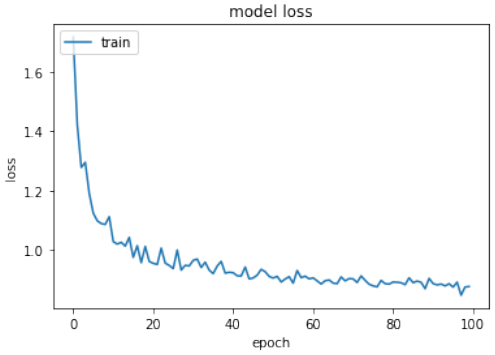
(c) Yu & Van Heuven [20] original dataset

Final loss: 0.8518



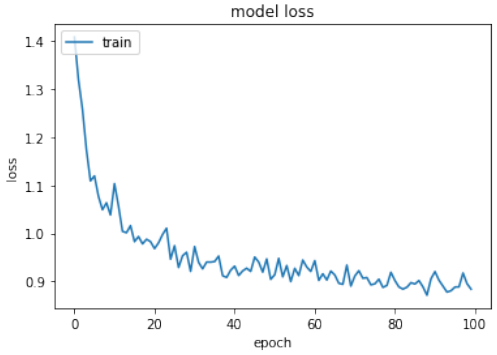
(d) Yu & Van Heuven [20] modified dataset

Final loss: 0.8769



(e) Deshmukh et al. [3] original dataset

Final loss: 0.8843



(f) Deshmukh et al. [3] modified dataset

Figure 5: Training loss using different features with the MLP model trained in the original (left), and modified (right) datasets

GitHub repository of this work: https://github.com/jjordanoc/robust-english-speech-fluency-classification