Oral Fluency Classification for Speech Assessment

Ashish Panda
TCS Research – Mumbai
ashish.panda@tcs.com

Rajul Acharya
TCS Research – Mumbai
rajul.acharaya@tcs.com

Sunil Kumar Kopparapu
TCS Research – Mumbai
sunilkumar.kopparapu@tcs.com

***Abstract*—Automatic speech quality assessment finds importance in evaluating the quality of spoken speech, especially by L2 speakers. Goodness of pronunciation, stressing on the right syllable in a multi-syllable word, and oral fluency are a few main components which are assessed for a speaker. While gauging pronunciation and identifying the syllable stress is relatively standard, oral fluency assessment has large variation in the rubrics used in addition to qualitative dimension to measure the quality of fluency. In this paper, we explore, using Statistical Machine Leaning (SML) and Deep Learning (DL) models, to classify oral fluency using two publicly available datasets, namely, Avalinguo Audio Dataset (AAD) and SpeechOcean762 (SO762). We introduce pre-trained DeepSpeech model embeddings in conjunction with known speech features like Mel- Frequency Cepstral Features (MFCC) and Fluency Features (FF) to correctly predict the fluency class. The best classification accuracy obtained for AAD was 95.04%, while the same for SO762 was 77.12%.**

***Index Terms*—oral fluency, goodness of pronunciation, machine learning, deep learning, speech features, DeepSpeech**

I. Introduction

The rapid increase of demand for second language (L2) learners to study foreign languages leads to the imminent need for an automatic L2 speech proficiency assessment. Good *oral fluency* in the absence of *speaking disturbances* not only enhances L2 language proficiency but also makes it sound more natural and native-like to listeners. As a consequence, L2 speech fluency assessment has been mostly embodied in designing efficient Computer-Assisted Pronunciation Tools (CAPT) for L2 learners. Though oral fluency suggests attaining mastery over a language, there exists a debate in the definition and measurement of fluency [1] . Therefore, developing an automatic and reliable measure of oral fluency is essential in language pedagogy, various fields of applied linguistics, and language assessment.

The non-standard interpretation of oral fluency in different disciplines render oral fluency assessment as an extremely challenging task. Some technical challenges associated with the oral fluency assessment are: (a) Oral fluency is often conflated with the notion of language proficiency and therefore assessing fluency in midst of other pronunciation problems makes oral fluency assessment a challenging task, (b) Fluency assessment is carried out by human experts on a given perceptual scale. However, there is no consensus standard scale for perceptual evaluation of oral fluency. And in the absence of any standard training fluency assessment process the perceptual assessment by human experts might be strongly biased [1].

Fast and automatic assessment tool of fluency by means of computerized programs has attracted a lot of attention. This kind of assessment has been part of the most CAPT and Computer-Assisted Language Learning (CALL) systems for L2 learners. However, such systems which are actually available to L2 learners and teachers are still scarce. Furthermore, the prerequisite for developing CAPT and CALL systems are designing cost-effective and efficient algorithms that can accurately and consistently predict the fluency score. Due to the lack of availability of specific standard fluency assessment tools, there is a tendency of relying on automatic speech recognition (ASR) systems, either trained on native (L1) or non-native (L2) or both speaker data which renders it unusable [2], [3]. Moreover, most of the research studies focused on L2 English as the non-native language. There exist very few works on fluency assessment that consider non-English as the L2 language [4]–[7]. Another challenge, especially in the era of deep learning, is the lack of annotated training data. Many previous works often rely on their own manually recorded, and labelled (by human experts of fluency evaluation) datasets, which are often quite small and also not publicly available [8] [9].

In this paper, we explore and experiment with different speech features and different classifiers to identify the oral fluency class for speech assessment. We specifically concentrate on two publicly available datasets, namely Avalinguo Audio Dataset (AAD) [10] and SpeechOcean762 (SO762) [11]. The main contribution of this paper is that for oral fluency classification task:

- We implement and compare the performance of various Statistical Machine Learning (SML) and Deep Learning (DL) algorithms.
- We motivate the use of DeepSpeech embeddings as additional features and demonstrate their performance.
- We report fluency classification results on SO762 database. To the best of our knowledge, no prior work has reported fluency classification results on this database.

The rest of the paper is organized as follows. Section II briefly describes the various statistical machine learning and deep learning algorithms that we have used in this work. Section II-B describes the various features that have been used in this paper for oral fluency classification task. Section III presents the experimental set-ups and results, while Section IV concludes this paper.

II. Proposed Approach

Figure 1 illustrates our proposed method for fluency classification.

A. Classifiers

We have experimented with and compared the performance of various classifier algorithms such as: Gaussian Mixture Model (GMM) [12], Support Vector Machine (SVM) [13], Random Forest (RF) [14] and 1D CNN. Table I describes the specifics of the SVM, RF, and GMM models used. Table II describes the model architecture of 1D CNN. The model consists of 8 blocks stacked together of which both Block A and Block B are further made out of layers as shown in the table. The last 2 layers are linear layers containing 32 neurons and 3 or 4 neurons depending on the number of classes in the working dataset.

TABLE I
Model Description of SVM, GMM, and RF model

Model	Parameters
SVM	Radial Basis Function (RBF) Kernel, Penalty Parameter (C) = 200, Gamma = 0.001
RF	Maximum Depth = 18, Random State = None
GMM	No. of Components = 16, Random State = None

TABLE II
Model Description of 1D CNN model

Unit	Output Channels
Block A	32
Block B	64
Block A	128
Block B	128
Block A	128
Block B	128
Block A	64
Block B	32
Linear A	32 (neurons)
Linear B	3 or 4 (neurons)

Block A
Conv1d (3 × 3) Relu BatchNorm

Block B
Conv1d (3 × 3) Relu BatchNorm Max Pool (2 × 2) Dropout (0.25)

B. Features for Oral Fluency

Mel Frequency Cepstral Coefficients (MFCCs) features are the most widely used speech representation in applications such as ASR systems, Speech Translation (ST), Speaker Recognition (SR), Speaker Diarization (SD), etc. The MFCC feature extraction process is explained in more details in [12].

Fluency Features (FF) comprise of signal-specific measurements, such as number of syllables, number of pauses, rate of speech, articulation rate, speaking duration, original duration, and other fundamental frequency ($f0$) measures such as mean, variance, median, minima, maxima, etc. of the $f0$ contour.

`DeepSpeech` Embedding Features (DSEF) have been used extensively for speech analysis. It has also been shown in [15] that number of $<$ `unk` $>$ outputs from `DeepSpeech` can be a measure of speech intelligibility and ASR accuracy is correlated with fluency as shown in [3]. Therefore, it is conceivable that embeddings from `DeepSpeech` (an end to end ASR) model will encode some fluency characteristics. We chose `DeepSpeech` model to extract embeddings, which are then used as features for oral fluency classification. Note that `DeepSpeech` [16] is an end-to-end speech recognition system comprising of a recurrent neural network (RNN) model. This model is trained on thousands of hours of training data, which is a combination of collected and synthesized data, that induces robustness to noisy environments and speaker variation [16]. The model consists of 5 hidden layers first 3 of which are non-recurrent, 4^{th} layer is a bidirectional recurrent layer, 5^{th} layer takes the outputs from both forward and backward layer of the bidirectional layer, and finally a 6^{th} *softmax* layer yielding character probabilities.

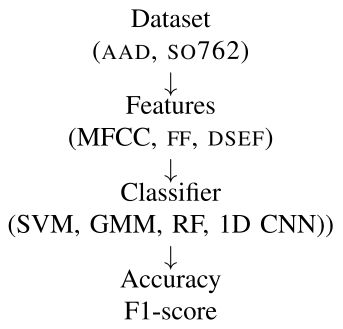

Fig. 1. Flow diagram showing experimental setup.

III. Experimental Results and Analysis

A. Datasets

The Avalinguo Audio Dataset (AAD) [10] consists of 1424 audio recordings of non-native English speakers labeled into three classes, namely, low, intermediate, high fluency labels. The audio is part of spontaneous (non-scripted) conversational speech with low or no background noise. Each audio recording comprises of a conversation around 10 minutes and this is cut into 5 sec equal-length segments resulting in 120 audio segments per 10 minute recording. In all, a total of 1420 non-overlapped audio segments ($\approx$ 2 hours) manually labelled into one of the three fluency classes: low, intermediate, high.

In AAD, the fluency levels are annotated by the human experts and the fluency classes are defined as low, intermediate, and high fluency. The low fluency class expects a person to talk about on familiar topics but the speech has unnatural pauses. The intermediate fluency class expects the person to be able to describe experiences and events and is also capable of giving reasons, opinions. However, there can still be some unnatural pauses. The high fluency class expects the person to (a) speak without unnatural pauses (no hesitation), and expects them to not pause long to find expressions.

SpeechOcean762 (SO762) is a new open-source read speech corpus designed for pronunciation assessment usage[1]. The database comprises of 5000 English utterances from 250 non-native (L2) Mandarin speakers, where half of the speakers are children. This corpus's text script is selected from daily life text, containing about $2,600$ common English words. The number of sentences read aloud by each speaker is 20, and the total duration of the audio is about 6 hours. The training and test set are predefined in the database, with 125 speakers for each. One of the key challenge associated with this database for speech fluency assessment is the class-imbalance problem. The details about the database can be found in [11].

Five human experts annotated the pronunciation of each of the utterances at sentence-level, word-level and phoneme-level in SO762. The sentence-level fluency annotations represent the oral fluency. The sentence-level fluency according to different score ranges are divided into four classes,

- 0-3: The speaker is not able to read the sentence as a whole or there is no voice
- 4-5: The speech is incoherent, with many pauses, repetition and stammering
- 6-7: Coherent speech in general, with a few pauses, repetition and stammering
- 8-10: Coherent speech, without noticeable pauses, repetition or stammering

B. Experimental Setup

The SVM [17], RF, GMM and 1D CNN models trained on AAD and SO762 try to learn the correct fluency class from one or a combination of the input feature vectors and performance metrics namely, accuracy and F1-score are computed as shown in Fig. 1.

The 22 MFCC features are extracted with a frame length of 25 ms and a window hop of 10 ms. These MFCC features along with their Δ and Δ^2 features are vertically stacked together resulting in 66-D MFCC features. Let $\{X_i = x_1, x_2, \cdots, x_t\}$ be a $(t \times 66)$ feature representation of a i^{th} speech sample. We take the mean of X_i across all the t frames to get a 66-D feature vector. This averaging of the MFCC features was done so that they can be stacked with the fluency features which are computed for the entire utterances rather than per speech frame.

Along with the 66-D MFCC features, we extract specific features, called the fluency features (FF) from the speech (see Section II-B). The 15-D FF are computed using the publicly available `myvoice` analysis toolkit[2]. While all the 15-D FF are used in case of SO762, for AAD we found that 4 of the 15 FF gave best classification accuracies.

Embeddings extracted from the `DeepSpeech` model, are also used as features for the classification algorithms. Here again, we take the mean of the extracted embeddings across all frames to get 2048-D feature vector. To choose amongst the best features suited for this problem, we selected the layer that showed the most distinction in the 2-D projected space, via t-SNE plot. Figure 2 shows t-SNE plots for different `DeepSpeech` layer outputs, for AAD. It can be observed that, comparatively more distinct class clusters are present in Fig. 2(c) that represents spatial dissimilarity amongst layer-3 embeddings of various classes. Hence, layer-3 DSEF are used in all our further experiments on AAD and SO762 dataset.

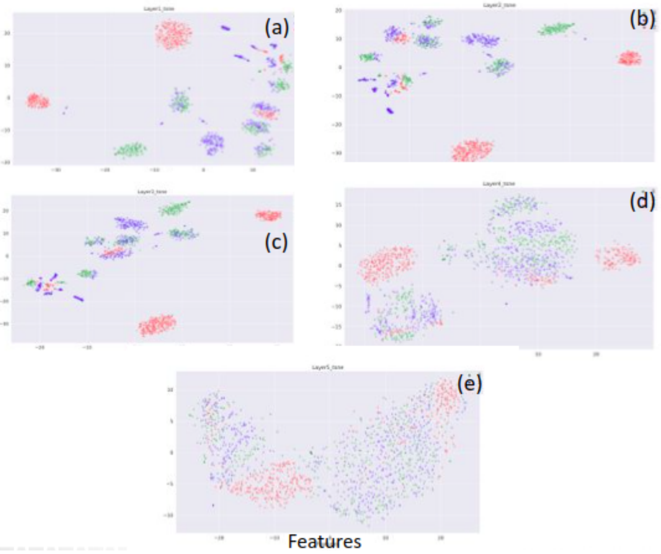

Fig. 2. t-SNE plots of DSEF for AAD, (a) layer-1, (b) layer-2, (c) layer-3, (d) layer-4, and (e) layer-5, are different layer outputs. Colours represent different classes, namely, low, intermediate and high speech samples.

We have employed mean and variance normalization for all features in our study, i.e., the features were normalized so that the mean for the entire set is a zero vector while the covariance matrix is an identity matrix for the entire set.

The 1D CNN model, as described in Section II, has been trained with cross-entropy loss function and a learning rate of $5e$-5. The model is optimised using Adam optimizer with β-1 and β-2 being 0.9 and 0.999, respectively. The model is trained for 200 epochs in the absence of DSEF. However, results as reported in Table III and Table IV that involve DSEF, are trained on one-tenth the number of epochs, i.e., only 20 epochs.

Finally, the system performance is evaluated using prediction accuracy and F1-scores. Accuracy represents the correctly classified data instances over total number of data instances whereas F1-score captures both precision and recall of a classification problem [18]. In case of class-wise balanced data as in case of AAD, both accuracy and *macro* F1-score are useful for system performance evaluation. However, in case of imbalanced classes as in case of SO762, accuracy and *weighted* F1-score is used. These metrics are evaluated using the sklearn python library [3]. The following sections discusses more in depth about the results and observations.

[1] https://www.openslr.org/101/

[2] https://github.com/Shahabks/my-voice-analysis

[3] https://scikit-learn.org/stable/

TABLE III
RESULTS ON AVALINGUO AUDIO DATASET (AAD).

Features			SVM		GMM		RF		1D CNN	
MFCC	FF (4-D)	DSEF	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
✓	×	×	81.07	81.15	89.02	88.90	87.85	87.94	91.12	91.06
✓	✓	×	81.07	81.13	84.81	84.73	89.95	89.88	89.49	89.45
×	×	✓	92.52	92.46	89.95	89.92	85.98	86.10	91.12	91.08
×	✓	✓	92.52	92.45	89.25	89.33	87.38	87.46	90.65	90.63
✓	✓	✓	**93.22**	**93.16**	**90.65**	**90.59**	**90.42**	**90.40**	**95.09**	**95.04**

TABLE IV
RESULTS ON SPEECHOCEAN762 DATASET (SO762).

Features			SVM		GMM		RF		1D CNN	
MFCC	FF (15-D)	DSEF	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
✓	×	×	70.16	67.80	67.96	67.38	69.84	62.69	63.08	64.09
✓	✓	×	75.16	**73.95**	**73.00**	**72.00**	**77.12**	**74.07**	**71.56**	**71.73**
×	×	✓	76.80	73.00	72.24	67.64	74.52	69.01	61.28	64.06
×	✓	✓	76.92	73.15	71.80	67.49	75.40	70.41	60.48	63.61
✓	✓	✓	**76.92**	73.06	71.96	67.60	75.72	70.89	61.36	64.33

C. *Experimental Results*

Table III presents the results from our experiments with AAD. It can be observed that with only 66-D MFCC features, 1D CNN provides the best performance with accuracy of 91.12% and F1-score of 91.06%. When FF are used along with MFCC features, however, RF and 1D CNN provide almost equal performance. For DSEF and DSEF with FF SVM provides the best performance. If all the features: MFCC, DeepSpeech embeddings and FF are stacked together, however, 1D CNN provides the best performance of 95.09% accuracy and 95.04% F1-score. For AAD, we have used 5-fold cross validation and the table shows mean of the 5-fold experiments. 5-fold cross-validation is a procedure where the entire data is split into 5 equal parts. Out of the 5 parts, the model is trained on 4 parts and tested on the remaining part. This process is repeated 5 times, till all the data points have been tested once [19], [20]. These experiments present results that follow a different trend than what has been reported in [21]. This shows that features selected have an impact on the performance of a particular classification method. In general, DSEF along with MFCC and FF provide the best performance for all classification algorithms. It should be noted that the performance reported here for 1D CNN is better than the best performance obtained in [21].

Table IV reports the results from the experiments with SO762 database. In case of AAD, 1D CNN provided the best results. In SO762, however, RF classifier provides the best accuracy and F1 score of 77.12% and 74.07% respectively. In case of this database, the DeepSpeech features do not provide improvement over MFCC and FF, in general. This is possibly because the speech type of SO762 mismatched with the training data of DeepSpeech model. For example, DeepSpeech model is not trained with children speech, while about half of the SO762 is child speech. It could also be interesting to experiment with embeddings from other layers of DeepSpeech model and other semi-supervised speech models such as Wav2Vec [22], HuBERT [23], etc. As some classes are under-represented in the training data of SO762, the performance, in general, is not as high as what we have seen in case of AAD. To the best of our knowledge, this is the first time fluency classification accuracy has been reported on this database.

IV. CONCLUSIONS

In this paper, we explore oral fluency classification on two openly available datasets. We have reported results using 4 types of classifiers and different combinations of 3 types of features. We have obtained the best reported results so far on AAD of above 95% accuracy, which we obtained using 1D CNN classifier and a combination of MFCC, FF and DSEF. We have also reported, oral fluency classification results on SO762 dataset for the first time. The DeepSpeech features represent one aspect of transfer learning technique that we have used in this study. In the future it will be interesting to study how we can use transfer learning more effectively for fluency classification task.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Monisankha Pal for his contribution in the initial phases of this project. A few baseline experiments were conducted by him.

REFERENCES

[1] M. van Os, N. H. de Jong, and H. R. Bosker, "Fluency in dialogue: Turn-taking behavior shapes perceived fluency in native and nonnative speech," *Language Learning*, vol. 70, 2020.

[2] L. Fontan, M. Le Coz, and C. Alazard, "Using the forward-backward divergence segmentation algorithm and a neural network to predict l2 speech fluency," in *Proc. 10th International Conference on Speech Prosody*, vol. 2020, 2020, pp. 925–929.

[3] J. Tao, K. Evanini, and X. Wang, "The influence of automatic speech recognition accuracy on the performance of an automated speech assessment system," in *IEEE Spoken Language Technology Workshop (SLT)*, 2014.

[4] S. Detey, L. Fontan, M. Le Coz, and S. Jmel, "Computer-assisted assessment of phonetic fluency in a second language: a longitudinal study of japanese learners of french," *Speech Communication*, vol. 125, pp. 69–79, 2020.

[5] V. De Fino, L. Fontan, J. Pinquier, I. Ferrané, and S. Detey, "Prediction of l2 speech proficiency based on multi-level linguistic features," in *23rd INTERSPEECH Conference: Human and Humanizing Speech Technology*, 2022.

[6] C. Cucchiarini, H. Strik, and L. Boves, “Quantitative assessment of second language learners’ fluency: an automatic approach,” in *5th International Conference on Spoken Language Processing: ICSLP 98*, 1998, pp. 2619 – 2622.

[7] L. Fontan, S. Kim, V. De Fino, and S. Detey, “Predicting speech fluency in children using automatic acoustic features,” in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 1085–1090.

[8] H. Deng, Y. Lin, T. Utsuro, A. Kobayashi, H. Nishizaki, and J. Hoshino, “Automatic fluency evaluation of spontaneous speech using disfluency-based features,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[9] L. Fontan, M. L. Coz, and S. Detey, “Automatically measuring l2 speech fluency without the need of asr: a proof-of-concept study with japanese learners of french,” in *INTERSPEECH*, 2018.

[10] Agriga9-github, “Avalinguo-audio-dataset: Dataset for speaker fluency level classification,” https://github.com/agrija9/Avalinguo-Audio-Set, Accessed: Feb 2023.

[11] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, “speechocean762: An open-source non-native english speech corpus for pronunciation assessment,” in *Proc. Interspeech*, 2021.

[12] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using gaussian mixture speaker models,” *IEEE transactions on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995.

[13] M. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Scholkopf, “Support vector machines,” *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, 1998.

[14] S. Yan, L. Ye, S. Han, T. Han, Y. Li, and E. Alasaarela, “Speech interactive emotion recognition system based on random forest,” in *2020 International Wireless Communications and Mobile Computing (IWCMC)*, 2020, pp. 1458–1462.

[15] A. Tripathi, S. Bhosale, and S. K. Kopparapu, “Automatic speaker independent dysarthric speech intelligibility assessment system,” *Computer Speech & Language*, vol. 69, p. 101213, 2021.

[16] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, “Deep speech: Scaling up end-to-end speech recognition,” *arXiv preprint arXiv:1412.5567*, 2014.

[17] I. Steinwart and A. Christmann, *Support vector machines*. Springer Science & Business Media, 2008.

[18] D. M. Powers, “Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation,” *arXiv preprint arXiv:2010.16061*, 2020.

[19] T. Fushiki, “Estimation of prediction error by using k-fold cross-validation,” *Statistics and Computing*, vol. 21, pp. 137–146, 2011.

[20] S. Yadav and S. Shukla, “Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification,” in *2016 IEEE 6th International conference on advanced computing (IACC)*. IEEE, 2016, pp. 78–83.

[21] A. Preciado-Grijalva and R. F. Brena, “Speaker fluency level classification using machine learning techniques,” *arXiv preprint arXiv:1808.10556*, 2018.

[22] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[23] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.