Holistic Trimodal Assessment of Fluency: Audio, Transcription, and Fluency Markers

Papa Séga Wade[1,2] Mihai Andries[1] Ioannis Kanellos[1] Thierry Moudenc[2]
[1] IMT Atlantique, Brest, France [2] Orange Innovation, Lannion, France

May 29, 2024

Abstract Fluency contributes to the clarity of speech, though it is not the sole measure. Models for automatic fluency assessment (AFA) used in language education technology often do not provide a comprehensive analysis. This study introduces a trimodal system using audio analysis, transcription timestamps, and fluency markers, such as speech rate, lexical diversity, lexical density or n-gram repetitions, to comprehensively evaluate language fluency. Leveraging the strengths of self-supervised learning (SSL) models like wav2vec 2.0, HuBERT, and WavLM, we extract rich audio embeddings that capture the nuances of spoken language. For textual analysis, we employ a BERT-based model to classify and extract features from transcriptions, while fluency marker features are computed using a combination of Random Forest, KNN, and CNN-LSTM models, with a final layer of SVM stacking. These extracted features are then integrated and classified using a BiLSTM network, providing an extended evaluation of fluency that surpasses traditional methods. Our approach offers a more exhaustive assessment by integrating multiple data modalities. For the validation of the proposed system, we used data from the Speechocean762 corpus, which includes 5,000 utterances from 250 speakers, half of whom are children. To ensure comprehensive evaluation, we enriched our analysis with the Avalinguo dataset, featuring spontaneous speech. Combining these datasets tests, the system's adaptability to both read and spontaneous speech demonstrates robustness and generalization ability. Speech rate was found to be the most relevant feature, significantly contributing to overall fluency assessment.

Keywords Fluency, Language Learning, Fluency Markers, Speech Patterns, Speech Rate, Lexical Diversity, Lexical Density, Self-Supervised Models, Embeddings.

References

1. Segalowitz, N. *Second language fluency and its underlying cognitive and social determinants*, IRALLT, vol. 54, no. 2, pp. 79–95, 2016.
2. Zhang, J. et al. *Speechocean762: An open-source non-native English speech corpus for pronunciation assessment*, arXiv preprint arXiv:2104.01378, 2021.
3. Grijalva, A. P. *Avalinguo audio set*, `https://github.com/agrija9/Avalinguo-Audio-Set`, 2018.
4. Liu, J. et al. *Multimodal automatic speech fluency evaluation method for Putonghua Proficiency Test propositional speaking section*, in *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 260–264, 2022.

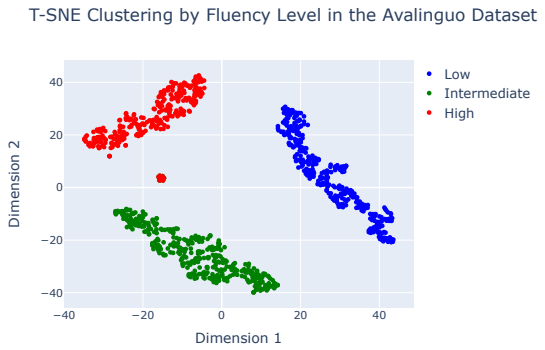

Figure 1: T-SNE Clustering by Fluency Level in the Avalinguo Dataset

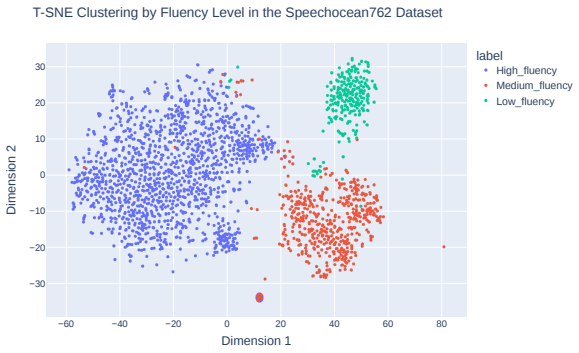

Figure 2: T-SNE Clustering by Fluency Level in the Speechocean762 Dataset

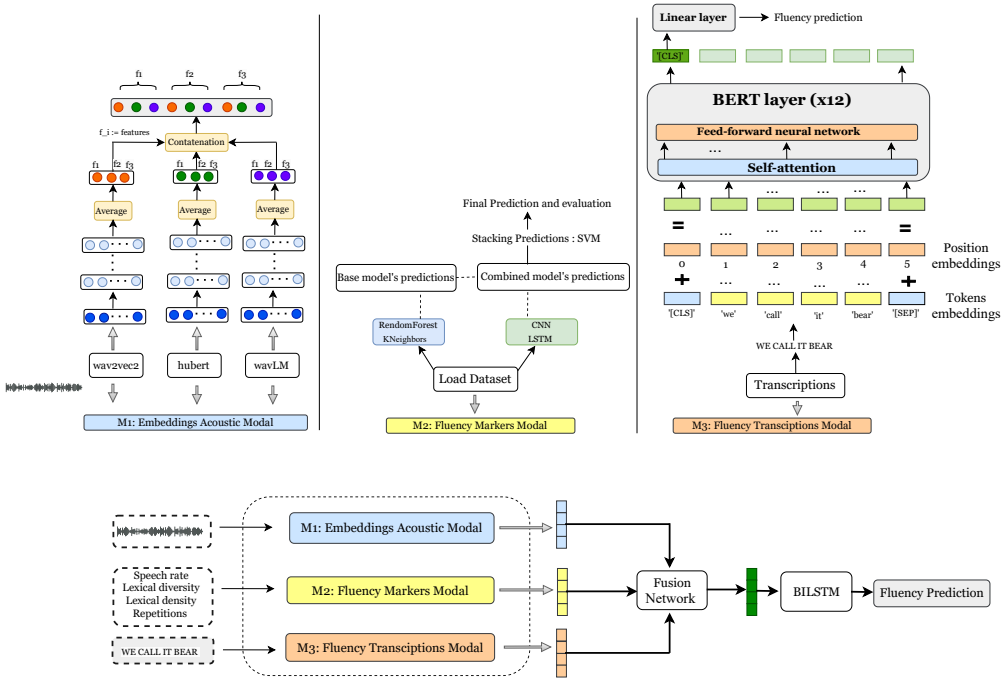

Figure 3: Trimodal Fluency system

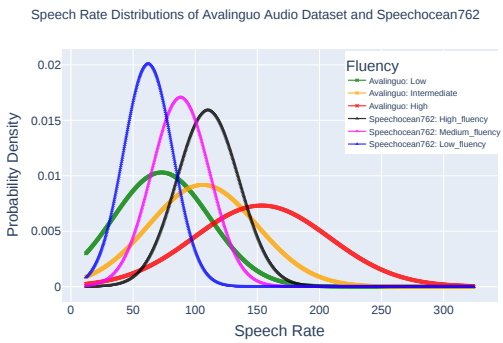

Figure 4: Speech rate distribution of Avalinguo and Speechocean762 dataset

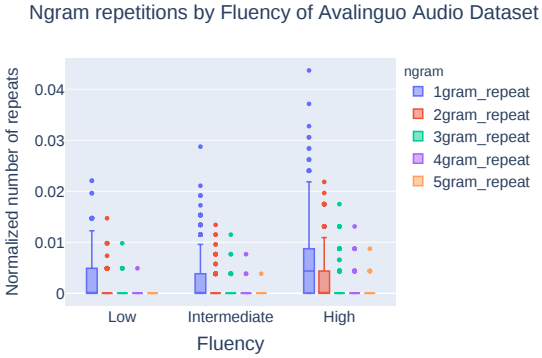

Figure 5: Ngram repetitions of Avalinguo Dataset