CBF-AFA: Chunk-Based Multi-SSL Fusion for Automatic Fluency Assessment

Papa Séga WADE
Orange Innovation, Châtillon, France
IMT-Atlantique, Lab-STICC
UMR CNRS 6285 Brest, France
papasega.wade@orange.com

Mihai ANDRIES
IMT-Atlantique, Lab-STICC
UMR CNRS 6285, F-29238
Brest, France
mihai.andries@imt-atlantique.fr

Ioannis KANELLOS
IMT-Atlantique, Lab-STICC
UMR CNRS 6285, F-29238
Brest, France
ioannis.kanellos@imt-atlantique.fr

Thierry MOUDENC
Orange Innovation
Lannion, France
thierry.moudenc@orange.com

***Abstract*—Automatic fluency assessment (AFA) remains challenging, particularly in capturing speech rhythm, pauses, and disfluencies in non-native speakers. We introduce a chunk-based approach integrating Self-supervised learning (SSL) models—Wav2Vec2, HuBERT, and WavLM—selected for their complementary strengths in phonetic, prosodic, and noisy speech modeling, with a hierarchical CNN-BiLSTM framework. Speech is segmented into breath-group chunks using Silero voice activity detection (Silero-VAD), enabling fine-grained temporal analysis while mitigating over-segmentation artifacts. SSL embeddings are fused via a learnable weighted mechanism, balancing acoustic and linguistic features, and enriched with chunk-level fluency markers (e.g., speech rate, pause durations, n-gram repetitions). The CNN-BiLSTM captures local and long-term dependencies across chunks. Evaluated on Avalinguo and Speechocean762, our approach improves F1-score by 2.8 and Pearson correlation by 6.2 points over single SSL baselines on Speechocean762, with gains of 4.2 F1-score and 4.0 Pearson points on Avalinguo, surpassing Pyannote.audio-based segmentation baselines. These findings highlight chunk-based multi-SSL fusion for robust fluency evaluation, though future work should explore generalization to dialects with irregular prosody.**

***Index Terms*—Fluency, Wav2vec2, HuBert, WavLM, speech chunking, self-supervised learning, Silero-VAD, breath-group**

I. Introduction

Linguistic fluency is a multidimensional construct encompassing speech rate, pauses, syntactic complexity, and coherence, influenced by both linguistic (grammar, vocabulary) and non-linguistic (hesitations, pauses) factors. Assessments typically rely on subjective ratings or objective acoustic metrics, each with inherent limitations.

Early AFA methods used handcrafted acoustic features (e.g., speech rate, syllable counts, pauses) combined with statistical and deep learning (DL) models, such as (SVMs, LSTMs) [1],[2],[3],[4]. However these approaches required extensive feature engineering and large annotated datasets. While DL reduced manual feature extraction, data scarcity remains a key challenge.

SSL models Wav2Vec2 [5], HuBERT [6], and WavLM [7] learn speech representations directly from raw audio, each excelling in distinct aspects: Wav2Vec2 captures phonetic content, HuBERT models prosodic structure, and WavLM enhances robustness in noisy speech [7]. These models complement each other by providing distinct fluency cues [7]–[9]. However, most AFA studies rely on full-utterance or frame-level analysis, which may overlook fine-grained disfluencies [10].

We introduce a chunk-based AFA framework that integrates Silero Voice Activity Detection (Silero-VAD) [11] for breath-group segmentation, multi-SSL fusion (Wav2Vec2, HuBERT, WavLM), and explicit fluency markers (speech rate, pause frequency). A CNN-BiLSTM architecture captures both local and global dependencies across speech chunks. Our key hypotheses are: (1) Chunking enables more fine-grained disfluency analysis than full-utterance processing, (2) Multi-SSL fusion enhances representational diversity, (3) Explicit fluency features complement SSL embeddings.

Our main contributions are:

- A novel breath-group chunking approach for AFA,
- A learnable SSL fusion mechanism,
- An end-to-end CNN-BiLSTM framework integrating SSL embeddings and fluency markers.

We validate our approach on Avalinguo [12] (1,388 conversational speech samples) and Speechocean762 [13] (5,000 scripted utterances), covering a range of fluency levels. The remainder of this paper is organized as follows: related works are discussed in Section II, methodology in Section III, experiments and results in Section IV, and conclusion in Section V.

II. Related Work

AFA has evolved from relying on handcrafted acoustic features—such as speech rate and pause duration—to modern DL and SSL approaches [2], [4]. Traditional systems often used Random Forests or Support Vector Machines trained on manually engineered features. However, they remained constrained by data scarcity and feature-selection bias. DL architectures, such as Bidirectional LSTMs, later improved

temporal modeling [1], but their effectiveness remained contingent on large annotated datasets.

Recent SSL methods, such as Wav2Vec2 [5], HuBERT [6], and WavLM [7], learn representations directly from raw audio, reducing reliance on transcripts. These have enabled ASR-free (Automatic Speech Recognition) fluency scoring [14] and pronunciation evaluation [15]. However, fluency encompasses more than segmental aspects; factors such as speech flow and pause patterns play a crucial role in perceived fluency [10]. Studies leveraging SSL embeddings have shown promising results on non-native speech corpora such as Avalinguo and Speechocean762 [1], [7], [12]. However, many still process entire audio or simple frame-level splits, limiting fine-grained disfluency analysis.

Parallel research on speaker diarization and related tasks has highlighted the benefits of segmenting long audio into smaller chunks using frameworks like `pyannote.audio` [16] or Silero-VAD modules such as Silero. However, the full potential of chunk-based segmentation for AFA remains underexplored. Our work addresses this gap by segmenting audio into breath-group chunks, integrating multiple SSL embeddings (Wav2Vec2, HuBERT, WavLM) via a learnable weighted fusion, and incorporating fluency markers (e.g., speech rate, pause frequency) within a hierarchical CNN-BiLSTM model. This approach combines fine-grained analysis of local disfluencies with robust global modeling, introducing a novel framework for AFA.

III. Automated Fluency Assessment methodology

Our proposed end-to-end methodology for AFA combines chunk-based segmentation of audio, self-supervised learning (SSL) embeddings fused via a learnable weighted-sum, and additional fluency-related features (e.g., speech rate, pauses) within a CNN-BiLSTM classification architecture. The approach comprises the following steps:

1) Silero-VAD and Breath-Group Chunking
2) Fine-Tuning SSL & Feature Extraction
3) Weighted Fusion of SSL Embeddings
4) Fluency Feature Computation
5) CNN-BiLSTM Classification

We detail each step below.

A. Step 1: Silero-VAD and Breath-Group Chunking

We first apply a Silero-VAD [11] module to identify segments of active speech and discard irrelevant portions (e.g., long silences, background noise). Although utterances are pre-segmented, this additional VAD step enables finer prosodic chunking by identifying intra-utterance breath pauses. Specifically, we employ Silero-VAD, which generates start and end timestamps

$$\mathcal{T} = \{(t_1^{\text{start}}, t_1^{\text{end}}), \ldots, (t_N^{\text{start}}, t_N^{\text{end}})\} \tag{1}$$

for detected speech regions in the input signal $\mathbf{x}$ as $\mathcal{T} = \text{Silero-VAD}(\mathbf{x})$. Each VAD-segmented region is then further subdivided into *breath-group chunks*. A breath group typically ends where the speaker takes a noticeable pause or breath, helping us capture natural prosodic boundaries. Practically, this can be done by detecting within-segment silences of duration above a threshold δ. Let $\mathbf{x}_\ell$ denote the ℓ-th speech region from Silero-VAD; breath-group chunking yields as $\mathbf{c}_i = \mathbf{x}_\ell[\text{start}_i : \text{end}_i]$ with $i = 1, \ldots, M$, where start_i and end_i mark the breath-group boundaries within that Silero-VAD segment. This multi-stage segmentation ensures that each chunk is both speech-only (thanks to Silero-VAD) and aligned with a natural prosodic boundary (breath group).

B. Step2: Fine-Tuning SSL & Feature Extraction

1) Fine-Tuned SSL Models for Fluency Embedding

To capture robust and nuanced acoustic features, we leverage three pre-trained SSL models — *wav2vec2-large-960h-lv60-self*, *hubert-large-ls960-ft* , and *wavlm-large* — each fine-tuned on our fluency dataset using a Connectionist Temporal Classification (CTC) loss. This fine-tuning, on the training data, aligns audio sequences with fluency labels without requiring manual alignment, enabling the models to learn task-specific features while preserving their general speech representation capabilities.

- **Wav2Vec2**: Learns contextualized representations directly from raw audio through a contrastive learning objective, excelling in capturing fine-grained phonetic details and temporal dynamics.
- **HuBERT**: Utilizes masked prediction of clustered acoustic units, emphasizing prosodic and structural patterns in speech, which are critical for modeling fluency.
- **WavLM**: Enhances robustness by jointly modeling speech content and speaker variations, making it particularly effective in diverse or noisy speaking conditions.

The embeddings from these models are fused to create a comprehensive representation that encapsulates phonetic, prosodic, and rhythmic aspects of non-native speech.

2) SSL Feature Extraction

For each chunk $\mathbf{c}_i \in \mathbb{R}^L$, we extract high-level acoustic embeddings from three SSL models. Formally, each model Φ_m (where $m \in \{\text{wav2vec}, \text{hubert}, \text{wavlm}\}$) outputs a frame-level representation:

$$\mathbf{F}_m^{(i)} = \Phi_m(\mathbf{c}_i) \in \mathbb{R}^{T_i \times d_m} \tag{2}$$

where T_i is the number of frames for chunk i and d_m is the model's embedding dimension. We then apply a mean-pooling operation over time to obtain a fixed-dimensional vector.

C. Step 3: Weighted Fusion of SSL Embeddings

To optimally exploit the complementary strengths of Wav2Vec2, HuBERT, and WavLM, we combine their chunk-level embeddings via a learnable weighted-sum mechanism (referred to as SSLFusion):

$$\mathbf{f}_{\text{fused}}^{(i)} = \alpha_1 \, \mathbf{f}_{\text{wav2vec}}^{(i)} + \alpha_2 \, \mathbf{f}_{\text{hubert}}^{(i)} + \alpha_3 \, \mathbf{f}_{\text{wavlm}}^{(i)} \tag{3}$$

where $\alpha_1, \alpha_2, \alpha_3 \geq 0$ and $\sum_{j=1}^{3} \alpha_j = 1$. These fusion weights $\{\alpha_j\}$ are learned jointly with the downstream classification objective, allowing the model to automatically determine the relative importance of each SSL source. The fused embedding has dimension $d = \max(d_{\text{wav2vec}}, d_{\text{hubert}}, d_{\text{wavlm}})$; $d = 1024$.

D. Step 4: Fluency Feature Computation

In addition to the fused SSL embedding, we compute chunk-level fluency markers that capture disfluencies and temporal aspects of speech. Common metrics include:

- SpeechRate$_i$: number of words per second in chunk i.
- PauseDuration$_i$: average silence length surrounding chunk i.
- ArticulationRate$_i$: syllables per second of active speech i.
- NgramRepetition$_i$: measure of repeated n-grams from transcriptions.

Let $\mathbf{m}_i \in \mathbb{R}^k$ gather these fluency markers for chunk i. We then concatenate $\mathbf{m}_i$ with the fused SSL embedding $\mathbf{f}_{\text{fused}}^{(i)}$:

$$\mathbf{h}_i = \text{Concat}\big(\mathbf{f}_{\text{fused}}^{(i)}, \mathbf{m}_i\big) \in \mathbb{R}^{d_{\text{fused}}+k} \tag{4}$$

This enriched chunk representation $\mathbf{h}_i$ encodes both acoustic/linguistic cues from SSL and explicit fluency metrics.

E. Step 5: CNN-BiLSTM Classification

Each utterance, segmented into M breath-group chunks, is processed by a CNN-BiLSTM model to classify fluency levels (`Low`, `Medium` or `Intermediate`, `High`). This architecture conceptually resembles dual-path modeling in DPRNN [17].

A 1D convolutional layer with 128 filters ($kernel = 3, stride = 1$) extracts local acoustic patterns from each chunk, capturing diverse temporal features. Its outputs are passed to a 2-layer Bidirectional LSTM (BiLSTM) ($hidden\_size = 256$), which models dependencies across chunks by processing the sequence in both forward and backward directions. The final BiLSTM outputs are aggregated via mean-pooling, passed through a fully connected layer, and classified using a softmax function. Dropout ($p = 0.3$) is applied to prevent overfitting with a learning rate of 1e-4.

IV. Experiment

Our experimental setup aims to validate the effectiveness of chunk-based segmentation and multi-SSL fusion for AFA. We first describe the datasets and preprocessing steps, followed by a study on the optimal segmentation threshold. Finally, we present the results of our comparative experiments.

A. Datasets

We evaluate our approach on two publicly available non-native English corpora, summarized in Table I. Each utterance is labeled for fluency based on human annotations.

1) Avalinguo Audio Dataset [12]

Avalinguo consists of 1424 audio recordings from non-native English speakers, labeled with *Low*, *Intermediate*, or *High* fluency. After excluding 36 files with minimal speech, 1388 recordings remain, sampled at 16 kHz and spanning approximately 2 hours of conversational speech. The speech regions range from 0.5 s to 4.9 s depending on the number of words spoken. We apply 5-fold cross-validation to ensure robust evaluation. A normalized version, including transcriptions and statistical analyses, is available online.[1]

2) Speechocean762 [13]

Speechocean762 contains 5000 English utterances from 250 non-native speakers. Each speaker reads 20 sentences, and each utterance is rated by five human experts on a 0–10 fluency scale. Following a total duration of about 6 hours (2.88 h for training and 2.69 h for testing). For consistency, we consolidate the original 0–10 ratings into three categories: *Low_fluency* (0–5), *Medium_fluency* (6–7), and *High_fluency* (8–10). The processed dataset, along with our fluency annotations, is publicly accessible.[2]

Table I
Overview of the Avalinguo and Speechocean762 datasets.

Dataset	Subset	#Segments	Duration
Speechocean762	Train	2500	2.88 h
	Test	2500	2.69 h
Avalinguo	Train/Test	1388	$\approx$ 2 h

To ensure consistent input across models, all audio files are down sampled to 16 kHz and normalized for amplitude variations.

B. Optimal chunking threshold analysis

The segmentation threshold $\delta = 300$ ms was empirically determined as optimal. F1-score and Pearson Correlation Coefficient (PCC) comparisons across 200, 250, 300, and 350 ms show that 300 ms yields the highest performance on both datasets. Shorter thresholds ($\leq 250ms$) over-fragment speech, while longer ones ($\geq 350ms$) introduce noise.

Results confirm that 300 ms best preserves fluency cues, aligning with natural prosodic transitions while optimizing classification accuracy and correlation with human ratings.

C. Baseline Comparisons and Ablation Study

To assess the effectiveness of our chunk-based multi-SSL framework, we compare it against two baselines: (i) a frame-level segmentation method based on Pyannote.audio, and (ii) individual SSL models without chunking or fusion. An ablation study is also conducted to quantify the impact of chunking, fusion, and fluency-specific features.

1) Baseline Comparisons

Table II summarizes the results obtained for both datasets.

Observations:

1) Pyannote.audio baseline provides reasonable performance, particularly on Avalinguo, but is outperformed by all SSL-based models.
2) Among the individual SSL models, WavLM achieves the best results, suggesting it captures fluency-relevant acoustic features more effectively.
3) Our Fusion + Silero-VAD Chunking approach achieves the highest performance across both datasets, demonstrating the benefit of:
   - Breath-group chunking, which enables better segmentation of fluency patterns.
   - Multi-SSL fusion, which improves the model's representational power.

[1] https://huggingface.co/datasets/papasega/Avalinguo-Audio-Dataset-splitted

[2] https://huggingface.co/datasets/papasega/speechocean762_fluency

Table II
COMPARISON OF BASELINE MODELS AND OUR PROPOSED APPROACH ACROSS SPEECHOCEAN762 AND AVALINGUO DATASETS.

Model	Speechocean762		Avalinguo	
	F1-score	PCC	F1-score	PCC
Pyannote.audio Baseline	0.759	0.676	0.857	0.825
Best Single SSL Model (wavLM)	0.802	0.749	0.931	0.926
Kim, E. et al. (Interspeech 2022)	–	0.78	–	–
Panda, A. et al. (EUSIPCO 2023)	0.740	–	0.950	–
Liu, W. et al. (ICASSP 2023)	–	0.795	–	–
Fusion + Silero-VAD Chunking (Ours)	**0.825**	**0.796**	**0.969**	**0.963**

2) Ablation Study

To evaluate the contribution of each component, we conduct an ablation study by removing key elements from the full model. WavLM outperforms other SSL models in both datasets. Fluency speech rate correlates best with WavLM, confirming its effectiveness in temporal modeling. Fusion of all three SSL models improves performance, leveraging complementary fluency cues. Silero-VAD-based chunking further boosts results, capturing speech rhythm and pauses. Multi-SSL fusion and chunking are crucial for accurate AFA.

3) Voice Quality Analysis and Fluency Patterns

Beyond SSL embeddings, we analyze voice quality features, fundamental frequency (F0), shimmer, and harmonic-to-noise ratio (HNR), to examine their correlation with fluency. Figure 1 shows that, in an audio sample, high shimmer (32.03%) and negative HNR (-2.01 dB) suggest vocal instability and spectral noise, common in non-native or hesitant speech. F0 variations further highlight prosodic irregularities.

These findings support the idea that voice quality metrics complement SSL-based fluency assessment, providing additional insights into speech rhythm and stability.

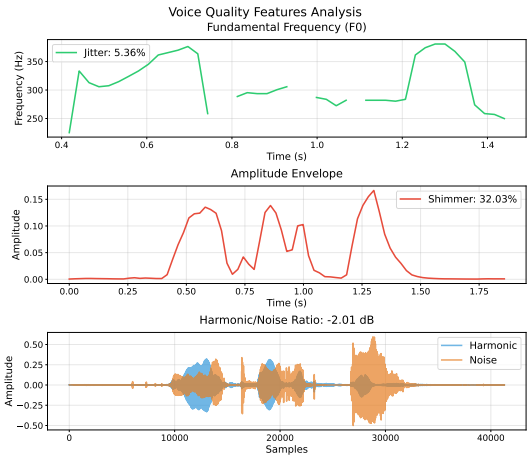

Figure 1. Voice Quality Analysis (F0, shimmer, HNR).

4) t-SNE Visualization of SSL Embeddings

To further analyze the effectiveness of SSL embeddings and fusion, we visualize their t-SNE projections for both datasets in Figures 2 and 3. These plots illustrate how well fluency levels are separated in the feature space.

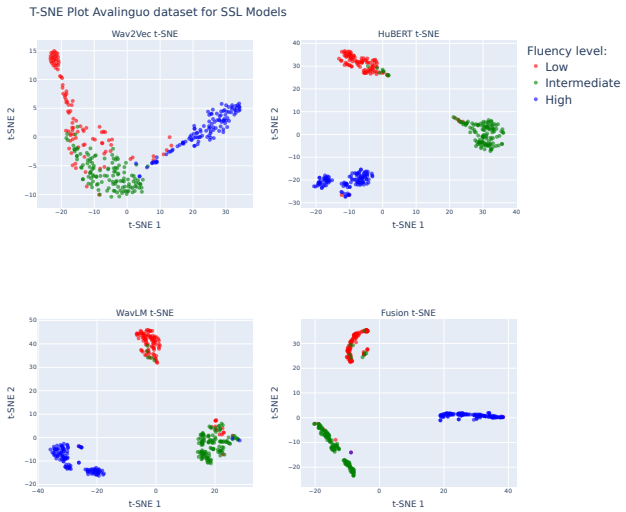

Figure 2. t-SNE Visualization of SSL embeddings (Avalinguo)

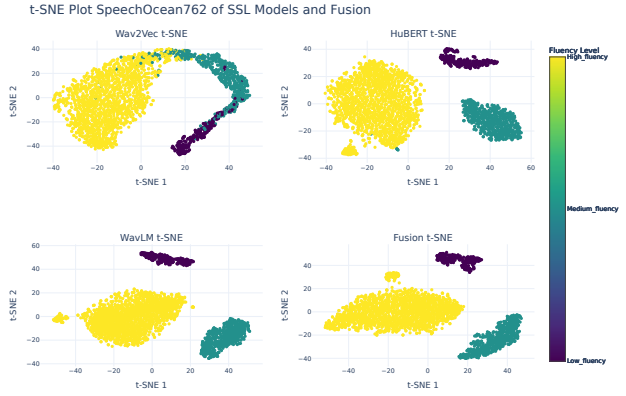

Figure 3. t-SNE Visualization of SSL embeddings (Speechocean762)

Key Findings from t-SNE Analysis:

1) Fusion-based embeddings exhibit more coherent clustering compared to Wav2Vec2, and provide slight improvements in separability over HuBERT and WavLM.
2) Fluency classes become more distinct with our approach, suggesting that chunk-based embeddings help disentangle fluency-related acoustic properties.

These findings further support the use of fusion and chunking,

as they lead to better feature space organization, which is crucial for fluency classification.

5) Results and Discussion

Our model exceeds baselines, with multi-SSL fusion enhancing phonetic and prosodic fluency representation. Silero-VAD-based chunking enhances segmentation, reducing noise while preserving speech rhythm.

Ablation results confirm that no single SSL model is sufficient; fusion consistently improves performance, with WavLM performing best among individual models. Analysis of fluency-related speech rate further supports the effectiveness of segmentation-aware models.

t-SNE visualizations reveal improved class separability with fusion, while voice quality analysis highlights instability patterns in non-fluent speech, suggesting that acoustic markers could enrich fluency assessment.

Our approach provides a scalable and adaptable fluency evaluation framework.

V. Conclusion

We introduced a chunk-based, multi-SSL fluency assessment framework, leveraging Wav2Vec2, HuBERT, and WavLM fusion with Silero-VAD segmentation. Our approach outperforms baselines and demonstrates that chunking enhances fluency modeling by preserving speech rhythm and disfluency markers.

Beyond SSL embeddings, our study suggests that voice quality metrics (F0, shimmer, HNR) can enrich fluency evaluation, bridging acoustic and linguistic fluency perspectives.

Future work could explore the adaptability of our approach to multilingual fluency assessment, particularly for low-resource languages such as Wolof, where fluency modeling remains underexplored. Other promising directions include integration with speech recognition models and attention-based fusion mechanisms for more dynamic weighting of SSL embeddings across linguistic contexts. Our approach assumes breath groups align with fluency, but this may not hold for speakers with irregular breathing patterns (e.g., due to anxiety). Future work could explore adaptive thresholding.

References

[1] Panda, A. et al., "Oral fluency classification for speech assessment," in *Proc. EUSIPCO '23*, 2023, pp. 231–235. DOI: 10.23919/EUSIPCO58844.2023.10289791.

[2] Deng, H. et al., "Automatic fluency evaluation of spontaneous speech using disfluency-based features," in *Proc. ICASSP*, 2020, pp. 9239–9243.

[3] Deng, H. et al., "Comparison of static and time-sequential features in automatic fluency detection of spontaneous speech," in *Proc. O-COCOSDA*, 2021, pp. 158–163.

[4] Zhang, H. et al., "Multilingual speech evaluation: Case studies on english, malay and tamil," in *Interspeech*, 2021.

[5] Baevski, A. et al., "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *ANIPS*, vol. 33, pp. 12 449–12 460, 2020.

[6] Hsu, W-N. et al., "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM*, vol. 29, pp. 3451–3460, 2021.

[7] Chen, S. et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of STSP*, vol. 16, no. 6, pp. 1505–1518, 2022.

[8] R. C. M. C. Shekar, M. Yang, K. Hirschi, S. Looney, O. Kang, and J. H. L. Hansen, "Assessment of non-native speech intelligibility using wav2vec2-based mispronunciation detection and multi-level goodness of pronunciation transformer," in *Interspeech 2023*, 2023, pp. 984–988. DOI: 10.21437/Interspeech.2023-2371.

[9] L. Qu, T. Li, C. Weber, T. Pekarek Rosin, F. Ren, and S. Wermter, "Disentangling prosody representations with unsupervised speech reconstruction," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 32, pp. 39–54, Oct. 2023. DOI: 10.1109/TASLP.2023.3320864.

[10] De Fino, V. et al., "Prediction of l2 speech proficiency based on multi-level linguistic features," in *Proc. Interspeech '22*, 2022, pp. 4043–4047.

[11] S. Team, *Silero vad: Pre-trained enterprise-grade voice activity detector*, 2021. [Online]. Available: https://github.com/snakers4/silero-vad.

[12] A. P. Grijalva, *Avalinguo audio set*, https://github.com/agrija9/Avalinguo-Audio-Set, 2018.

[13] J. Zhang, Z. Zhang, Y. Wang, *et al.*, "Speechocean762: An open-source non-native english speech corpus for pronunciation assessment," in *Interspeech 2021*, 2021, pp. 3710–3714. DOI: 10.21437/Interspeech.2021-1259.

[14] Liu, W. et al., "An asr-free fluency scoring approach with self-supervised learning," in *Proc. ICASSP 2023*, IEEE, 2023, pp. 1–5.

[15] Kim, E. et al., "Automatic Pronunciation Assessment using Self-Supervised Speech Representation Learning," in *Proc. Interspeech*, 2022, pp. 1411–1415. DOI: 10.21437/Interspeech.2022-10245.

[16] H. Bredin, R. Yin, J. M. Coria, *et al.*, "Pyannote.audio: Neural building blocks for speaker diarization," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7124–7128.

[17] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 46–50. DOI: 10.1109/ICASSP40776.2020.9054266.