Leveraging Pseudo Segment Labels for Robust Automated Speaking Assessment in Read-Aloud Tasks

Anonymous ROCLING submission

Abstract

Automated speaking assessment (ASA) has become a crucial component in computer-assisted language learning, providing scalable, objective, and timely feedback to second-language learners. While early ASA systems relied on hand-crafted features and shallow classifiers, recent advances in self-supervised learning (SSL) have enabled richer representations for both text and speech, improving assessment accuracy. Despite these advances, challenges remain in evaluating long speech responses, due to limited labeled data, class imbalance, and the importance of pronunciation clarity and fluency, especially for read-aloud tasks. In this work, we propose a segment-based ASA framework leveraging WhisperX to split long responses into shorter fragments, generate pseudo-labels from holistic scores, and aggregate segment-level predictions to obtain final proficiency scores. Experiments on the GEPT corpus demonstrate that our framework outperforms baseline holistic models, generalizes robustly to unseen prompts and speakers, and provides diagnostic insights at both segment and response levels.

Keywords: Automated Speaking Assessment, WhisperX, Pseudo Labels

1 Introduction

With the rapid advances in computing technology and the growing population of second-language (L2) learners, automated speaking assessment (ASA) has attracted increasing attention and become an essential component in computer-assisted language learning (CALL). ASA systems are designed to provide timely and reliable feedback on learners' speaking performance, enabling them to improve their oral proficiency in an autonomous and low-stress environment. In addition, ASA offers scalable, objective, and consistent evaluations, thereby alleviating the workload of language instructors and facilitating large-scale language learning applications.

Early ASA research primarily relied on shallow classifiers and hand-crafted features that captured different aspects of speaking competence, such as delivery (e.g., pronunciation, fluency, intonation), content (e.g., appropriateness, relevance), and language use (e.g., grammar, vocabulary) (Cucchiarini et al., 1998; Chen et al., 2010; Coutinho et al., 2016; Chen et al., 2018; Qian et al., 2019; Wu et al., 2022). More recently, the emergence of self-supervised learning (SSL) paradigms has opened up new opportunities for ASA. Text-based SSL models, such as BERT and its derivatives (Devlin et al., 2019), provide contextualized embeddings that have been successfully adopted in various language assessment tasks, including sentence-level evaluation (Arase et al., 2022), essay scoring (Nadeem et al., 2019; Wu et al., 2023), and spoken monologue assessment (Craighead et al., 2020). In parallel, the rapid development of speech-based SSL models, such as wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021), has further strengthened ASA systems by offering rich acoustic representations (Bannò and Matassoni, 2023; McKnight et al., 2023; Lo et al., 2024).

Despite these advances, automated speaking assessment still faces persistent challenges in handling long speech responses. A representative example is the read-aloud task, where learners are evaluated primarily on pronunciation clarity and fluency. While text-based models can capture lexical accuracy, they are inherently limited in assessing these speech-specific aspects. Moreover, the development of reliable ASA systems is hindered by the scarcity of large-scale annotated data, as existing datasets are often limited in size and imbalanced across proficiency levels. The computational cost of processing extended speech recordings further compounds these difficulties. Consequently, the lack of sufficient labeled resources restricts model robustness and limits the ability to deliver fine-grained and diagnostic feedback.

In this work, we explore an ASA framework designed to address both the scarcity of labeled data and the challenges of long speech recordings. Specifically, we leverage WhisperX (Bain et al., 2023) to process long audio responses and obtain time-aligned segments, each of which is subsequently evaluated with segment-level scoring. To compensate for the lack of labeled resources, we weakly associate each segment with the holistic proficiency score of the full response, thereby generating pseudo-labels for training. This strategy not only increases the number of training instances, especially for underrepresented proficiency levels, but also highlights weaker segments where learner performance diverges from holistic expectations. Finally, segment-level predictions are aggregated (e.g., by mean or median) to reconstruct the overall proficiency score, offering a straightforward and interpretable mapping from local to global assessment.

Experiments on the GEPT corpus demonstrate that our framework consistently outperforms baseline holistic models and generalizes robustly to unseen prompts and speakers. We also investigate whether partial scoring of only the first or last 30 seconds of speech can approximate holistic judgments, revealing systematic differences that highlight both strengths and limitations of segment-level scoring.

In summary, our contributions are threefold:

1. We introduce a segment-based ASA framework for long read-aloud tasks that mitigates data scarcity through pseudo-labeled fragments;

2. We examine aggregation strategies for mapping segment-level predictions to holistic scores; and

3. We provide a comprehensive analysis of ASR quality and response-length effects on ASA performance.

These results offer new insights for designing ASA systems that are both data-efficient and diagnostically informative.

2 Related Work

2.1 volution of Automated Speaking Assessment Systems

Research on automated speaking assessment (ASA) has evolved from traditional feature engineering to the adoption of deep neural architectures. arly approaches relied on shallow classifiers with hand-crafted features targeting specific dimensions of proficiency, such as pronunciation, fluency, prosody, grammar, and vocabulary (Cucchiarini et al., 1998; Chen et al., 2010; Coutinho et al., 2016). While such systems demonstrated the feasibility of automatic scoring, their performance was often constrained by the limited representational power of manually designed features.

The advent of self-supervised learning (SSL) has substantially advanced ASA. On the text side, models such as BERT (Devlin et al., 2019) provide contextualized embeddings that have been successfully applied to various assessment tasks, including essay scoring (Nadeem et al., 2019), readability estimation (Arase et al., 2022), and spoken monologue evaluation (Craighead et al., 2020). These approaches leverage the semantic and syntactic richness of pre-trained language models, enabling more robust prediction of learner proficiency. In parallel, speech-based SSL models, such as wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021), have emerged as powerful tools for capturing acoustic and phonetic information. Recent studies demonstrate their effectiveness in proficiency prediction and related tasks (Bannò and Matassoni, 2023; McKnight et al., 2023; Lo et al., 2024), showing that such representations can encode both linguistic and paralinguistic aspects critical to ASA.

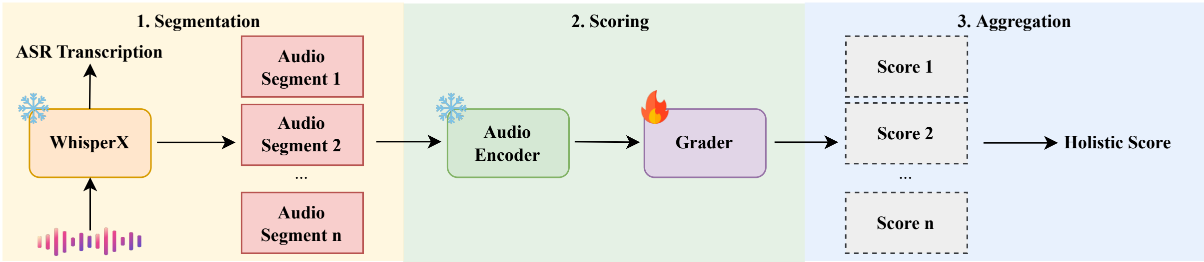


Figure 1: Proposed ASA framework: long read-aloud responses are segmented, each segment is scored independently, and the results are aggregated into a holistic proficiency score.

2.2 Handling Long Audio Inputs by WhisperX

WhisperX (Bain et al., 2023) is a system designed to efficiently transcribe long-form audio with word-level timestamps. It utilizes Voice Activity Detection (VAD) to segment audio into approximately 30-second chunks, which are then transcribed in parallel by Whisper and aligned with phoneme recognition models to produce accurate word-level timestamps. This approach enables batched inference, resulting in a twelve-fold speedup without sacrificing transcription quality. The segmentation process reduces issues like hallucinations and repetition, and the forced alignment ensures time-accurate transcriptions, making WhisperX suitable for applications such as subtitling and diarization.

3 Methodology

In this section, we describe the overall pipeline of our proposed Automated Speaking Assessment (ASA) framework, as illustrated in Figure 1. The system processes long audio responses by dividing them into manageable fragments, scoring each fragment independently, and subsequently aggregating these scores into a single holistic proficiency score.

3.1 Segmentation

Each spoken response in our dataset lasts approximately 90 seconds, which poses challenges for both ASR accuracy and downstream scoring. To address this, we employ WhisperX to obtain word-level timestamps. These timestamps allow us to segment each recording into shorter, coherent units of speech, hereafter referred to as *segments*. Each segment contains a contiguous portion of the learner's response, providing a finer-grained basis for subsequent scoring.

3.2 Pseudo-label Assumption

Since human raters typically provide only one holistic score per response, no ground-truth labels exist at the segment level. To overcome this limitation, we adopt a weak supervision strategy by assigning the holistic score of the full response to each of its segments as a pseudo-label. While this assumption may introduce label noise—because individual segments may not fully reflect overall proficiency—it substantially increases the number of training instances and enables finer-grained analysis of learner performance. This trade-off is particularly valuable under our limited-data setting.

3.3 Segment-Level Scoring

Each audio segment is then processed independently. We employ Whisper large as the acoustic encoder, and the resulting features, along with the ASR transcription, are fed into the grader module. The grader outputs a proficiency score for each segment under the pseudo-label supervision described above. This design not only expands the effective training set, especially for underrepresented proficiency levels, but also provides localized insights into which parts of a response may deviate from the expected holistic proficiency.

3.4 Aggregation Strategies

Finally, the system aggregates segment-level predictions into a holistic proficiency score for the entire response. We consider multiple strategies, including simple averaging and median pooling, to examine which approach best captures the relationship between localized performance and the overall judgment. Moreover, variations among segment scores

	1	2	3	4	5
Train	0	52	505	787	96
Valid	0	9	61	97	13
Known Content	0	6	67	99	8
Unknown Content	0	1	157	392	40

Table 1: Number of speakers for each holistic score in the GEPT dataset.

can highlight weaker portions of a response, offering diagnostic information beyond the final holistic score.

4 Experiments and Results

4.1 Dataset

This study utilizes a private corpus collected from the read-aloud task in the General English Proficiency Test (GEPT), an important large-scale English assessment in Taiwan. In this task, participants were instructed to read aloud two given paragraphs within one and a half minutes. The corpus consists of responses to eight different paragraph sets, with each set corresponding to a distinct passage.

Each response was independently scored by two professional raters on a five-point scale, where 1 represents the lowest performance and 5 the highest. The final score was obtained by averaging the two ratings. To evaluate model generalization, we define responses from unseen paragraph sets as the unknown content test set, while responses from previously seen sets are regarded as known content. The remaining data was further split into training, development, and test subsets following an 80/10/10 ratio.

The overall score distribution across training, validation, and test partitions is summarized in Table 1. This partitioning strategy ensures that the dataset supports evaluation under both familiar and novel content conditions, which is critical for assessing model robustness in practical applications.

4.2 Experimental Setup

We employed Whisper-large-v2[1] as our acoustic encoder in our framework. Model configurations were initialized using pretrained models from the HuggingFace Transformers library (Wolf et al., 2020). Training was conducted on a single NVIDIA 3090 GPU using Adam optimizer with a weight decay of 1e-5. The learning rate was set to 2e-4, and training was conducted for 15 epochs with a batch size of 25.

[1]https://huggingface.co/openai/whisper-large-v2

	Strategy	Known Content Accuracy (%)	Unknown Content Accuracy (%)
BERT	-	62.78	74.24
Whisper	First only	73.89	75.93
	Last only	78.33	76.10
Proposed	Mean	74.44	76.77
	Median	**82.22**	**78.47**

Table 2: Experiment results on the GEPT test dataset. "Known Content Accuracy" denotes the accuracy on the known content test set, while "Unknown Content Accuracy" represents the accuracy on the unknown content test set.

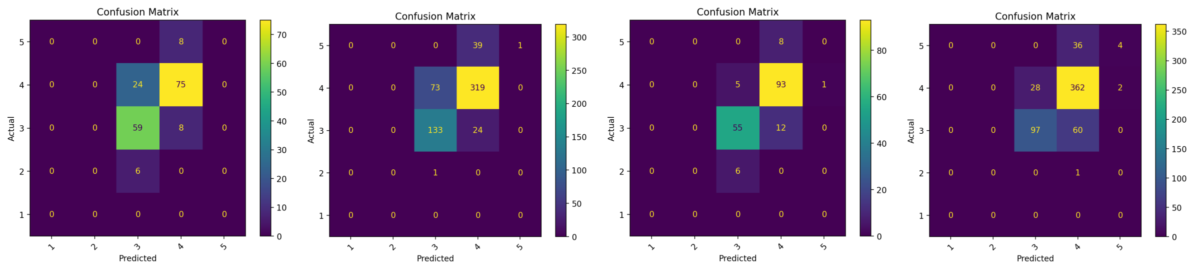
Baseline As a baseline, we employed a text-based SSL model, BERT[2]. The BERT encoder was frozen, while the same grading module used with the Whisper encoder was fine-tuned on top of it. For inference, the read-aloud audio was first transcribed by the Whisper-large-v2 model, and the resulting text was then passed to the frozen BERT encoder. Finally, the grading module predicted the holistic score on a 5-point scale.

Evaluation Metrics We evaluated model performance using classification accuracy, defined as the proportion of predicted holistic scores that exactly match the human ratings. Accuracy was chosen for its simplicity and consistency with prior work.

4.3 Results and Discussion

Table **??** summarizes the performance of our models under different configurations. The text-based baseline (BERT with Whisper transcription) achieved the lowest accuracy, highlighting the limitation of relying solely on ASR transcripts for holistic scoring. In contrast, the Whisper-large encoder trained on full-length read-aloud recordings substantially outperformed the baseline, confirming the benefit of leveraging acoustic-prosodic informa-

[2]https://huggingface.co/google-bert/bert-base-uncased



(a) Mean—Seen (b) Mean—Unseen (c) Median—Seen (d) Median—Unseen

Figure 2: Confusion matrices comparing mean and median aggregation strategies for proficiency prediction: (a) mean—seen prompts, (b) mean—unseen prompts, (c) median—seen prompts, and (d) median—unseen prompts.

tion beyond lexical content.

To investigate the effect of input length, we compared models using only the first 30 seconds and the last 30 seconds of each recording. Both truncated models showed a noticeable drop in accuracy compared to the full-length model, suggesting that information distributed throughout the passage contributes to reliable scoring. Interestingly, the last 30-second condition yielded slightly higher accuracy than the first 30-second condition, implying that later portions of the response may contain more representative prosodic or fluency cues.

We further analyzed performance using a segment-based approach with WhisperX alignment. Each recording was divided into segments, and segment-level scores were aggregated using either the mean or the median. Both strategies performed comparably to the full-length Whisper model, but the median aggregation demonstrated greater robustness to outlier segments.

Error patterns revealed by the confusion matrices (Figure 2) further highlight these differences. With the mean strategy, many level-4 responses were misclassified as level 3, and most level-5 responses were reduced to level 4. Due to the limited number of level-2 samples, the model struggled to classify them correctly. In contrast, the median strategy produced more concentrated predictions across both the known and unknown content test sets. Notably, for the unknown content condition, the median strategy yielded more correct classifications for level-5 responses compared to the mean strategy, indicating improved generalization on higher-proficiency learners.

5 Conclusion and Future Work

In this paper, we proposed a segment-based ASA framework for long read-aloud tasks, addressing both the scarcity of labeled data and the challenges of processing extended speech recordings. By leveraging WhisperX for time-aligned segmentation and applying pseudo-labeling at the segment level, our approach increased training diversity and enabled fine-grained analysis of learner performance. Experiments on the GEPT corpus showed that the proposed framework outperformed text-based baselines, generalized well to unseen prompts, and provided interpretable mappings from local to global proficiency scores through aggregation strategies. Further analyses indicated that median aggregation offered greater robustness than mean aggregation, particularly for high-proficiency learners, while ablations on partial input (first vs. last 30 seconds) highlighted the importance of full-length responses for reliable scoring.

Despite these promising results, the framework still faces limitations. It assumes that all spoken content aligns with the target passage, whereas learners may sometimes produce irrelevant or off-topic speech. Future work could integrate the reference passage to capture lexical or semantic mismatches, and extend the framework to open-response scenarios where content is not predetermined. These directions would enhance both the diagnostic capability and the applicability of ASA systems, paving the way toward more adaptive and learner-centered assessment.

References

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. Cefr-based sentence difficulty annotation and assessment. *arXiv preprint arXiv:2210.11766*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Stefano Bannò and Marco Matassoni. 2023. Proficiency assessment of l2 spoken english using wav2vec 2.0. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1088–1095.

Lei Chen, Keelan Evanini, and Xie Sun. 2010. Assessment of non-native speech using vowel space characteristics. In *2010 IEEE Spoken Language Technology Workshop*, pages 139–144.

Lei Chen, Jidong Tao, Shabnam Ghaffarzadegan, and Yao Qian. 2018. End-to-end neural network based automated speech scoring. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6234–6238. IEEE.

Eduardo Coutinho, Florian Hönig, Yue Zhang, Simone Hantke, Anton Batliner, Elmar Nöth, and Björn Schuller. 2016. Assessing the prosody of non-native speakers of English: Measures and feature sets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1328–1332, Portorož, Slovenia. European Language Resources Association (ELRA).

Hannah Craighead, Andrew Caines, Paula Buttery, and Helen Yannakoudakis. 2020. Investigating the effect of auxiliary objectives for the automated grading of learner english speech transcriptions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2258–2269.

Catia Cucchiarini, Helmer Strik, and Louis Boves. 1998. Quantitative assessment of second language learners' fluency: an automatic approach. In *Proc. ICSLP 1998*, pages paper–0752.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Tien-Hong Lo, Fu-An Chao, Tzu-I Wu, Yao-Ting Sung, and Berlin Chen. 2024. An effective automated speaking assessment approach to mitigating data scarcity and imbalanced distribution. *arXiv preprint arXiv:2404.07575*.

Simon W McKnight, Arda Civelekoglu, Mark Gales, Stefano Bannò, Adian Liusie, and Katherine M Knill. 2023. Automatic assessment of conversational speaking tests. In *9th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 99–103.

Farah Nadeem, Huy Nguyen, Yang Liu, and Mari Ostendorf. 2019. Automated essay scoring with discourse-aware neural models. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pages 484–493.

Yao Qian, Patrick Lange, Keelan Evanini, Robert Pugh, Rutuja Ubale, Matthew Mulholland, and Xinhao Wang. 2019. Neural approaches to automated speech scoring of monologue and dialogue responses. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8112–8116.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tzu-I Wu, Tien-Hong Lo, Fu-An Chao, Yao-Ting Sung, and Berlin Chen. 2022. A preliminary study on automated speaking assessment of English as a second language (ESL) students. In *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, pages 174–183, Taipei, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Tzu-I Wu, Tien-Hong Lo, Fu-An Chao, Yao-Ting Sung, and Berlin Chen. 2023. Effective neural modeling leveraging readability features for automated essay scoring. In *Proceedings of the 9th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 81–85.