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

Yue-Yang He, Berlin Chen
National Taiwan Normal University, Taipei, Taiwan
{yueyanghe, berlin}@ntnu.edu.tw

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 weak 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, Weak 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), HuBERT (Hsu et al., 2021), and Whisper (Radford et al., 2023), has further strengthened ASA systems by offering rich acoustic representations (Bannò and Matassoni, 2023; McKnight et al., 2023; Wu and Chen, 2024; 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

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 146–152

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 weak 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 alleviates the scarcity of sentence-level annotations by exploiting weak labels derived from holistic scores;
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 Evolution of Automated Speaking Assessment Systems

Research on automated speaking assessment (ASA) has evolved from traditional feature engineering to the adoption of deep neural architectures. Early 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), HuBERT (Hsu et al., 2021), and Whisper (Radford et al., 2023), 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.

However, most existing ASA systems treat each spoken response as a single, monolithic input, which becomes increasingly problematic when applied to long read-aloud tasks. Long-form speech raises both computational and temporal costs during training and inference, and more importantly, such systems typically produce only a holistic score without revealing which specific portions of the response contributed to the learner's performance. As a result, localized feedback is largely absent, and the literature contains relatively little work explicitly targeting the unique challenges of long-form ASA.

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 Weak-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 weak-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 processed independently to enable segment-level assessment. The Whisper encoder is adopted as the acoustic backbone, and its representations are fed into a grader module trained with weak segment-level supervision derived from holistic scores. This architecture effectively enlarges the usable training distribution, especially for low-resource proficiency levels, while providing localized diagnostic feedback that would otherwise be lost under holistic-only scoring.

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

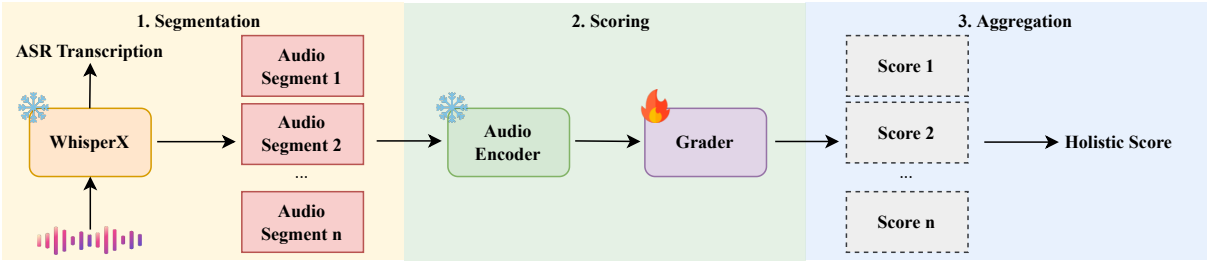


Figure 1: Proposed ASA framework: long read-aloud responses are segmented, each segment is scored independently, and the results are aggregated into a holistic proficiency score.

	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.

4 Experiments and Results

4.1 Dataset

This study utilizes a private corpus collected from the reading aloud task[1] 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 two 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[2] 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.

Baseline As baselines, we employed both a text-based SSL model and a speech-based SSL model, namely BERT[3] and wav2vec 2.0[4]. For the text-based baseline, the read-aloud audio was first transcribed by Whisper-large-v2, and the resulting text embeddings were extracted using a frozen BERT encoder; the same grading module used in our proposed framework was fine-tuned on top of it to predict holistic proficiency scores. For the speech-based baseline, we adopted wav2vec 2.0 as a frozen acoustic encoder and fine-tuned only the grading module on top of its representations.

Evaluation Metrics We evaluated model performance using three metrics: accuracy (ACC), weighted F1 score (F1), and Pearson correlation coefficient (PCC). ACC is defined as the proportion of predictions that exactly match the human-assigned holistic score. The

[1] https://www.gept.org.tw/Exam_Intro/t02_introduction.asp

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

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

[4] https://huggingface.co/facebook/wav2vec2-base

	Strategies	Known Content			Unknown Content		
		ACC↑	F1↑	PCC↑	ACC↑	F1↑	PCC↑
BERT	-	61.67	52.20	0.462	70.50	61.98	0.295
W2V	-	58.33	52.28	0.217	68.01	60.63	0.217
Whisper	First only	73.89	70.64	0.577	75.93	72.57	0.496
	Last only	78.33	74.97	0.679	76.10	72.72	0.499
Proposed	Mean	74.44	71.57	0.722	76.77	74.54	**0.623**
	Median	**82.22**	**79.04**	**0.748**	**78.47**	**76.01**	0.562

Table 2: Experimental results on the GEPT test dataset. "Known Content" denotes test samples with seen content, while "Unknown Content" denotes test samples with unseen content.

weighted F1 score accounts for label imbalance across proficiency levels, providing a more reliable estimate of performance on underrepresented categories. PCC further measures the monotonic relationship between predicted and reference scores, reflecting how well the model preserves the human-assigned ranking structure. These metrics jointly capture both discrete correctness (ACC) and ordinal consistency (F1, PCC), and are consistent with common practice in automated speaking assessment.

4.3 Results and Discussion

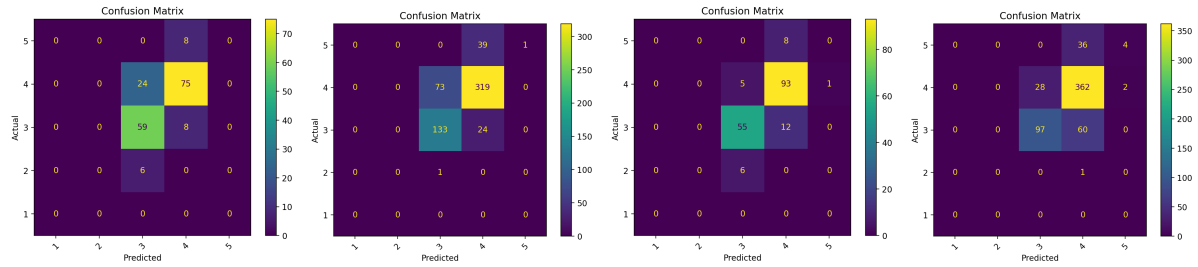
Baseline Performance Analysis. Table 2 summarizes the performance of our models under different configurations. The text-based baseline (BERT with Whisper transcription) achieved acceptable accuracy, highlighting the limitation of relying solely on ASR transcripts for holistic scoring. Interestingly, the speech-based SSL model (wav2vec 2.0) produced performance comparable to BERT in accuracy and weighted F1, but its PCC was substantially lower, particularly on the known-content set. This indicates that although both baselines can correctly classify a similar proportion of samples at the categorical level, the wav2vec-based model struggles to preserve the ordinal relationship among proficiency levels, likely due to its predictions being more distributionally concentrated and less sensitive to fine-grained prosodic variation relevant for human scoring. In contrast, BERT implicitly benefits from lexical cues captured via ASR, which may preserve a closer monotonic alignment with human-assigned proficiency levels.

Effect of Full-Length Training. The Whisper-based grader trained on full-length read-aloud recordings substantially outperformed both baselines across all three metrics, confirming the effectiveness of leveraging acoustic-prosodic information beyond lexical content. The performance gain in PCC further suggests that holistic fluency and speech quality are better reflected in continuous acoustic patterns than in discrete lexical sequences extracted from ASR transcriptions.

Temporal Coverage Analysis. To investigate the effect of temporal coverage, we compared models using only the first 30 seconds and the last 30 seconds of each recording. Both truncated variants yielded a noticeable drop across all metrics relative to the full-length model, suggesting that proficiency-related cues are distributed throughout the entire utterance rather than being concentrated at the onset. Notably, the last-30-second condition slightly outperformed the first-30-second condition, implying that later segments of the response may contain more stable or representative prosodic evidence of proficiency, potentially due to speakers settling into a more consistent speaking rhythm after the initial articulation phase.

Segment-Based Aggregation and Error Patterns. We further analyzed performance using a segment-based aggregation 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 achieved performance comparable to the full-length Whisper model, while the median aggregation proved more robust to local inconsistencies and noisy or disfluent segments. This suggests that outlier-prone stretches of speech disproportionately affect global predictions when treated as a single unit, and that segment-wise aggregation can stabilize scoring by emphasizing the speaker's typical performance rather than transient fluctuations.

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 pro-



(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.

duced 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 introduced a segment-based ASA framework for long read-aloud scoring, which addresses the data sparsity and temporal modeling challenges of full-length utterances. Using WhisperX for time-aligned segmentation and weak segment-level labeling, the framework improves supervision granularity and stabilizes the learning of proficiency-relevant speech cues. Experiments on the GEPT corpus showed consistent gains over text-only and speech-only baselines, and revealed that segmentation combined with median aggregation enhances robustness against disfluent or noisy segments. The analysis further highlights that full-length coverage remains essential for reliable scoring, as proficiency cues accumulate beyond early articulation.

Despite these promising results, the framework still assumes that all spoken content aligns with the target passage, whereas learners may occasionally insert off-topic or paraphrastic segments. Since WhisperX already provides high-resolution temporal alignment, future work could exploit this timing information to detect lexical or prosodic deviations from the reference passage, enabling segment-wise content validation rather than treating misalignment as uniform noise. This direction would further extend the framework from holistic scoring toward diagnostic assessment, and could generalize to open-response scenarios where content is not predetermined. Ultimately, incorporating alignment-based semantic verification would improve both the interpretability and applicability of ASA systems in real-world learner-centered settings.

Acknowledgments

This work was supported by the Language Training and Testing Center (LTTC), Taiwan. Any findings and implications in the paper do not necessarily reflect those of the sponsor.

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. *Proceedings of Interspeech, 2023*.

Stefano Bannò and Marco Matassoni. 2023. Proficiency assessment of l2 spoken english using wav2vec 2.0. In *Proceedings of SLT, 2022*, pages 1088–1095.

Lei Chen, Keelan Evanini, and Xie Sun. 2010. Assessment of non-native speech using vowel space characteristics. In *Proceedings of SLT, 2010*, pages 139–144.

Lei Chen, Jidong Tao, Shabnam Ghaffarzadegan, and Yao Qian. 2018. End-to-end neural network based automated speech scoring. In *Proceedings of ICASSP, 2018*, 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 *Proceedings of 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 *Proceedings of SLaTE, 2023*, 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 *Proceedings of ICASSP, 2019*, pages 8112–8116.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

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.

Chung-Wen Wu and Berlin Chen. 2024. Optimizing Automatic Speech Assessment: W-RankSim Regularization and Hybrid Feature Fusion Strategies. In *Proceedings of Interspeech, 2024*, pages 4004–4008.

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 SLaTE, 2023*, pages 81–85.