A Benchmark Corpus for the Diagnostic Assessment of Content in L2 English Speech

Kosuke Doi[1], Justin Vasselli[2], Taro Watanabe[2]
[1]Seikei University, [2]Nara Institute of Science and Technology
kosuke-doi@st.seikei.ac.jp, {vasselli.justin_ray.vk4, taro}@is.naist.jp

Abstract

When evaluating second language (L2) learners' speech, human raters pay significant attention to its content, and diagnostic feedback on content helps improve learners' speaking ability. Since human scoring and feedback are time-consuming and costly, automatic models aiming to provide such feedback have been developed, specifically models that detect whether certain content, i.e., key points, is included in learner's speech. However, previous studies target only integrated test items where learners speak based on listened or read materials, and the data used are not publicly available. In this study, we construct a speech corpus for key point detection. We extend the target to test items where learners speak based on their own experiences and opinions, which show greater content diversity than integrated test items, using an approach that annotates content along with its connections. Analysis of the constructed data demonstrated that the annotated elements are associated with the speech content scores. We also found that large language models are generally successful at locating content element spans, although their predicted spans are often broader than human-annotated ones. The corpus and annotation guidelines are available at https://language.sakura.ne.jp/icnale/download.html.

Keywords: speech scoring, content assessment, topic development

1. Introduction

The speaking ability of foreign language learners is evaluated across multiple dimensions, such as pronunciation, fluency, grammar, vocabulary, and content. Analytical scoring assigns separate scores to multiple aspects, offering more pedagogically useful feedback than holistic scoring, which provides only a single overall score (Knoch et al., 2021). Since human scoring of learners' speech is time-consuming and costly, researchers have developed automatic scoring models (Müller et al., 2009; Craighead et al., 2020; Bannò et al., 2025). However, scores alone, whether holistic or analytical, provide limited information about the specific strengths and weaknesses of learners.

To address this limitation, researchers have explored automated feedback generation (Nagata, 2019; Chao and Chen, 2025). In language testing research, speech content has been shown to play a major role in human raters' scoring decisions (Sato, 2012). The number of information units in a speech has been used as a criterion in rubrics for evaluating content (Kuiken and Vedder, 2018) and is associated with L2 proficiency. Building on this insight, Wang et al. (2020); Wang and Hamill (2021) developed models that automatically detect whether specific content required in a test question, i.e., key points, is included in a learner's response. Because high-proficiency responses are expected to address these key points appropriately, key point detection can be used for both scoring and feedback. However, previous studies focused on one type of test question: integrated test items, where learners are asked to speak based on the information they have listened to or read. For example, one test item may ask the learner to first listen to a conversation between professors, then provide an oral summary of the problem discussed and recommend one of the proposed solutions (see Wang and Hamill, 2021). Furthermore, the datasets used to develop the models have not been made publicly available.

Therefore, we construct a speech corpus for the diagnostic assessment of content, specifically a benchmark dataset for the key point detection task. We add content-based annotations to transcripts from an existing corpus of second language (L2) English, i.e., the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2023b).

We extend the annotation method proposed by Wang et al. (2020) to other types of test questions, e.g., independent and interactive test items. In these tasks, learners speak from their own experiences or opinions, such as discussing whether or not they agree with college students having part-time jobs, based on their own experiences and knowledge (see Ishikawa, 2023b). In these test questions, speech content varies across learners, unlike in integrated test items where listening or reading materials constrain the content. This variability makes it more challenging to predefine key points. To address this content diversity, we adopt a topic development-based approach similar to that

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1869–1877
11–16 May 2026.

used in the TOEFL iBT speaking rubric[1]. We annotate two types of content elements: **(1) Position**, the learners' response to the test question, and **(2) Support**, content that provides evidence such as reasoning or examples (Crossley et al., 2022). We show that these elements are associated with the analytical scores for content. We also demonstrate that large language models (LLMs) are generally successful at identifying the spans of these elements in transcripts, but they often predict wider spans than human annotators.

2. Background and Related Work

2.1. Content Scoring

Content is one of the core evaluation aspects in speaking assessment (Sato, 2012; Brown et al., 2005) and is included in the rubrics of various tests such as TOEFL iBT[2] and Pearson Test of English[3]. In recent years, automated speech scoring has been studied as a way to reduce the cost of human scoring (see Zechner and Evanini, 2020), and some models specifically focus on scoring content (Xie et al., 2012; Yoon and Lee, 2019; Voskoboinik et al., 2023). Content scoring is sometimes approached through its relationship with discourse coherence, as in Crossley et al. (2022), and several studies have explored the automatic assessment of speech coherence (Wang et al., 2017, 2019; Wang and Evanini, 2020).

However, automatic scoring models focus solely on predicting scores and therefore cannot provide specific feedback to learners. In the context of speech content evaluation, feedback models have been developed by detecting whether specific key points are included in the speech (Wang et al., 2020; Wang and Hamill, 2021). However, these models are limited to only a single type of test item, i.e., integrated test items, and the datasets used for developing the models have not been made publicly available.

2.2. ICNALE

ICNALE (Ishikawa et al., 2013; Ishikawa, 2023b) is a corpus of spoken and written English produced by college-level learners from ten Asian countries and regions, together with data from native speakers of English. The data were collected under controlled conditions to ensure fair comparison, such as identical time limits for response production. The same topics were used for both writing and speaking tasks: (a) the importance of part-time jobs for college students and (b) the complete ban on smoking in restaurants. The spoken component consists of monologue and dialogue modules, with both audio and transcript data available.

The ICNALE also includes a module, the ICNALE Global Rating Archives (GRA) (Ishikawa, 2023a), which provides rubric-based evaluation of 140 essays and 140 spoken dialogues by 80 raters with diverse linguistic and professional backgrounds. Each response was assigned a holistic score and ten analytic scores, four of which pertained to the content aspect. Data annotated by multiple raters can be used to investigate rater characteristics, including rater reliability. Moreover, responses that receive high scores from raters with diverse backgrounds can be regarded as typical examples of high-quality responses. These features enable the analysis of L2 performance quality and rater variation.

Furthermore, the ICNALE GRA has also been used as evaluation data for automated scoring (Bannò and Matassoni, 2022). It contains a balanced collection of learner data from 10 countries, covering proficiency levels from beginner to advanced, making it suitable for use as a benchmark dataset.

3. Dataset Construction

We constructed a dataset by adding annotations of claims and supporting evidence to the speech transcripts in the ICNALE GRA. While existing datasets, including those for essays, are typically annotated with content scores (Voskoboinik et al., 2023; Persing and Ng, 2013, 2014), this study annotates content elements along with their spans.

3.1. Materials

We used transcripts of all 140 speeches in the ICNALE GRA, which originally came from the role-play tasks of the ICNALE Spoken Dialogues module. In the role-play, learners, acting as college students, were asked to persuade a supervisor who opposed part-time jobs to allow them to continue working. Although the scores in the ICNALE GRA were assigned based on the initial 90 seconds of each audio recording, we annotated all learner utterances in the transcript. Annotation was applied only to utterances produced by the learner; the interviewer's utterances were not annotated.

3.2. Annotations

In independent and interactive test items, including role-plays, it is not feasible to predefine key points based on specific content because learners draw on their own experiences and opinions, resulting in diverse responses. This contrasts with the

[1] https://www.ets.org/pdfs/toefl/toefl-ibt-speaking-rubrics.pdf
[2] https://www.ets.org/toefl.html
[3] https://www.pearsonpte.com/

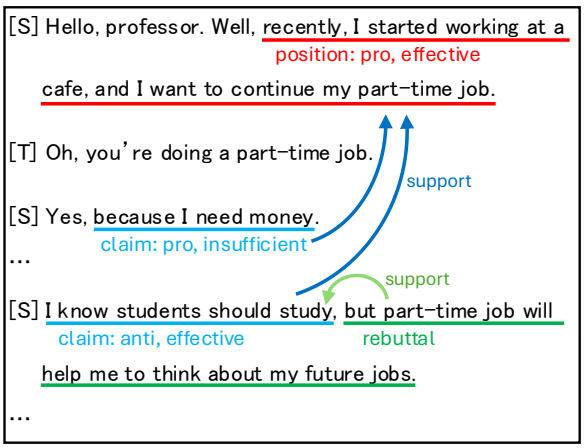

Figure 1: An annotation example. [S] and [T] represent utterances by a student and teacher, i.e., interviewee and interviewer, respectively.

method of Wang et al. (2020), which relies on predefined key points tied to specific source materials. To address this gap, we developed an annotation methodology[4] inspired by topic development-based scoring in the TOEFL iBT speaking rubric. We conducted multiple rounds of guideline revision and trial annotation, incorporating feedback from two researchers with experience in teaching English as a foreign language. Based on the guidelines, we annotated all 140 speech transcripts in the ICNALE GRA for the following content elements.

Position and Support We annotated each learner's speech for two types of discourse-level content elements: **Position**, which represents the speaker's stance toward the test question, and **support**, which reflects how the speaker develops and elaborates on that stance. An annotation example is shown in Figure 1, where the learner first states that they want to continue their part-time job (establishing a position) and then provides additional details, such as a need for money and to think about their future job (providing supporting arguments).

The topic development dimension covers both content and coherence, awarding higher scores to responses that include relevant ideas and appropriate supporting details (Wang and Evanini, 2020). We defined these content elements as key points, as follows:

Position: Represents the speaker's opinion or response to the test question. One of five tags is assigned based on the overall content: `pro` indicates agreement with the statement in the test question or alignment with the test setting; `anti` indicates disagreement or contradiction; `conditional` denotes that the position changes depending on conditions; `unclear` applies when the position cannot be determined; and `irrelevant` refers to speech unrelated to the test question. If a segment best represents the position, its span is also annotated. For the `unclear` and `irrelevant` tags, the span is marked as `n/a`. Furthermore, in some speeches, learners provide only additional details, such as reasons, without explicitly stating their position. When the position can be inferred from the overall speech, we assign the position tag with the span set to `n/a`. When the span is `n/a`, the clarity tag described later is assigned as `not stated`.

Support: These elements reinforce or elaborate on other content elements: `claim` supports the position; `evidence` and `rebuttal` elaborate or refute a `claim`, respectively; and `conclusion` supports the position by restating an opinion as a summary. `claim` and `conclusion` elements are further annotated with `pro` or `anti` flags, based on whether they align with the statement in the test question.

Clarity Inspired by Wang et al. (2020), we evaluated the quality of each position, `claim`, and `conclusion` element based on its clarity and specificity. In the example in Figure 1, the learner clearly states a desire to continue the part-time job, adding the details of working at a café. In contrast, the reason given (needing money) is vague and lacks persuasiveness. We defined the quality of content elements based on the following four categories: `effective` indicates that the statement is clear and specific; `insufficient` refers to a statement that is clear but less convincing, and also includes those that are partially unclear or confusing; `implied` is a dialogue-specific category indicating that a learner refers to an interviewer's statement indirectly, such as replying "Yes" to "Do you want to continue your part-time job?"; and `not stated` represents that the statement is missing.

4. Analysis and Experiments

4.1. Relationship between Annotated Tags and Scores

We investigated whether the types and frequencies of content elements used by learners differed according to the scores assigned in the ICNALE GRA. We divided the data into three groups based on the sum of the four content-related analytical scores in the ICNALE GRA: `low` for scores below the mean minus one standard deviation, `high` for

[4]The annotated data and the annotation guidelines are available at https://language.sakura.ne.jp/icnale/download.html

Score	# data	Tag					Clarity			
		pro	anti	conditional	unclear	irrelevant	effective	insufficient	implied	not stated
low	28	0.93	0.00	0.00	0.07	0.00	0.64	0.04	0.04	0.29
mid	87	0.93	0.01	0.03	0.02	0.00	0.66	0.09	0.03	0.22
high	25	1.00	0.00	0.00	0.00	0.00	0.52	0.04	0.08	0.36

Table 1: Relationship between content scores and position tags and clarity. # data represents the number of data for each score range, while the values for tag and clarity are proportions.

Score	claim	evidence	rebuttal	conclusion
low	5.71	2.36	0.32	0.21
mid	6.33	2.38	0.70	0.34
high	6.48	3.28	0.76	0.20

Table 2: Average number of supporting elements across different score ranges.

scores above the mean plus one standard deviation, and `mid` for those in between.

Position Table 1 shows the relationship between the scores and the position annotations. The predominance of the `pro` label indicates that most learners followed the role-play setting, though some low- and mid-level responses were tagged as `unclear`, or that their stance on continuing part-time work could not be determined from their responses. The `mid` group also includes responses annotated as `anti` or `conditional`. Although learners may take opposing or balanced positions in monologue tasks, they are expected to follow the given scenario in role-plays, which explains why position tags other than `pro` appear only in low- and mid-level learners' responses.

The clarity annotation for position tags suggests that high-scoring responses did not necessarily state a position explicitly. In approximately 20–30% of responses across all score ranges, the speaker's stance was not clearly stated, providing valuable insights for learner feedback.

Support Table 2 presents the average number of supporting elements per speech. High-scoring responses tended to contain more supporting elements, although the differences were not statistically significant across all element types. We further calculated the average number of elements supporting each `claim` and found that `claim`s in high-scoring speeches were supported by a greater number of elements, i.e., 1.29, 1.97, and 2.68 for the `low`, `mid`, and `high` groups, respectively. A one-way ANOVA followed by Tukey's test confirmed that these differences were significant between all the groups ($p < .05$). In the `high` group, about 50% of `claim`s were supported by `evidence` or `rebuttal`, compared with about 30% in the `mid` and `low` groups. Furthermore, the number of `claim`s rated as `effective` in clarity was significantly higher in the `high` group than in the `low` group, with means of 2.07, 3.14, and 4.08 for the `low`, `mid`, and `high` groups, respectively.

System

You are an experienced English teacher, who can score speech transcripts. Your task is to annotate transcripts following guidelines, which is given later.

User

(Annotation Guidelines here)

Task setting for the transcripts is based on the prompt below:

Task setting (role-play)
The interviewee plays the role of a college student wishing to continue his/her part-time job. The interviewee is told to persuade his/her supervisor, who firmly believes that students should not have part-time jobs, to allow him/her to continue working.

Learner speech is marked as [S], and interviewer speech as [T].

Please segment position, claim, evidence, and so on into the smallest possible units.

Output in the following JSON format:

```
{"position": {
    "tag": "Select a position tag based on the overall content.",
    "clarity": "Select a clarity tag.",
    "text": "The section that best expresses the student's
             position on the topic, if it exists. Do not omit any
             tokens in the transcript, except speaker tags." },
  "support": [
    {"text": "The section expressing a claim, evidence, rebuttal,
             or conclusion. Do not omit any tokens in the
             transcript, except speaker tags.",
     "tag": "Select a support tag.",
     "clarity": "Select a clarity tag." },
     ...
     List all instances of 'support' in the transcript.
     ...
  ]}
```

(Transcripts here)

Figure 2: Prompt example for key point detection task.

These results suggest that performance in speech content cannot be determined solely by the presence or number of content elements; rather, the connections among those elements and their clarity also play critical roles in distinguishing higher-scoring responses.

LLM	position		claim			evidence			rebuttal			conclusion		
	span (acc)	tag (acc)	pre	rec	f1	pre	rec	f1	pre	rec	f1	pre	rec	f1
GPT-5 (EM)	0.093	0.923	0.135	0.102	0.116	0.140	0.189	0.161	0.013	0.056	0.021	0.072	0.195	0.105
(PM)	0.557	0.962	0.609	0.430	0.504	0.454	0.631	0.528	0.093	0.393	0.150	0.198	0.537	0.289
Gemini (EM)	0.121	1.000	0.090	0.050	0.065	0.068	0.051	0.058	0.004	0.011	0.006	0.036	0.049	0.042
(PM)	0.429	0.967	0.476	0.276	0.350	0.293	0.237	0.262	0.072	0.191	0.105	0.182	0.220	0.199

Table 3: Results of the key point detection task by LLMs. EM and PM represents exact and partial match, respectively. acc, pre, rec, and f1 are accuracy, precision recall, and F1 scores, respectively. Span detection performance was relatively low, with EM scores around 0.1 or lower and PM scores up to about 0.6.

4.2. Key Point Detection by LLMs

LLMs have achieved high performance across a wide range of tasks, even in zero-shot settings (OpenAI, 2024; Google, 2023; Kojima et al., 2022). Using the dataset constructed in this study, we conducted a key point detection experiment with LLMs under a zero-shot setting to demonstrate the usefulness of the dataset.

Models We used GPT-5[5] and Gemini 2.5 Flash[6] for our experiments. We selected these strong API-accessible models because educational institutions and educators typically lack access to local computing resources capable of running comparably large models.

Prompt A prompt example is shown in Figure 2. Each model received a system prompt instructing it to act as an experienced English teacher capable of scoring and annotating speech transcripts. The prompt included the annotation guidelines and a description of the role-play scenario. We also instructed the LLMs to segment content elements into the smallest possible units[7] and to generate outputs in JSON format.

Evaluation We evaluated the LLMs' predicted spans by measuring their exact or partial overlap with human annotations. We calculated precision as the proportion of LLM-predicted spans found in the reference annotations, and recall as the proportion of reference spans correctly identified by the models; the F1 score was then calculated from these values. However, since position was annotated only once per speech, its span agreement was evaluated using accuracy. Agreement on the position tags and clarity tags was also evaluated using accuracy.

[5] https://openai.com/index/gpt-5-system-card/

[6] https://deepmind.google/models/gemini/flash/

[7] In preliminary experiments, LLMs tended to identify whole sentences rather than segments as content element spans, and adding this instruction slightly improved performance.

LLM	Example
GPT-5	[S] **I want to past-time job** because I, uh, college money. [T] Oh, you need some money for your study, you mean? But ...
Gemini	[T] ... I want you to stop working immediately and then to spend your time for study, not for work. [S] Hmm-um-huh-hmm. **Yes**, but uh in my uh-uh-uh my working space, uh, I have friends and there are not enough people to uh run the shop. So ...

Table 4: Examples where LLM-predicted spans are longer than the references. The underlined segments represent the spans predicted by the LLMs, and the bold segments represent the reference spans. The transcripts retain their original spellings.

Results Table 3 presents the results of the key point detection task. Span agreement scores were generally low under the exact-match condition and improved slightly under the partial match, but even the highest accuracy or F1 scores remained around 0.5, indicating that the LLMs struggled to detect content elements accurately. In contrast, position tags were predicted with high accuracy (above 0.9). The prediction of clarity labels for position, `claim`, and `conclusion` elements was also relatively high, with accuracy scores exceeding 0.6.

To further analyze span detection, we disregarded the element labels and examined the overlap between the spans predicted by the LLMs and those in the reference annotations. Under the exact-match condition, F1 scores remained low, 0.22 for GPT-5 and 0.11 for Gemini, whereas partial match yielded scores of 0.86 and 0.66, respectively.

Initial observations of examples such as those in Table 4, where both models predicted spans that included multiple content elements, suggested

LLM	EM	PO	tgt ⊂ ref	ref ⊂ tgt	no overlap
GPT-5	0.207	0.101	0.057	**0.466**	0.169
Gemini	0.121	0.048	0.011	**0.495**	0.325

Table 5: Results of span-overlap analysis. EM and PO represents exact and partial overlap, respectively.

that the LLMs tended to predict longer spans than the human references, which we suspected might inflate partial-match scores. The average span lengths were 11.43, 14.74, and 19.74 words for the reference, GPT-5, and Gemini, respectively, confirming that LLMs tend to predict longer spans than the references.

To better understand how the predicted spans aligned with reference annotations, we conducted a span-overlap analysis, categorizing each prediction according to its relationship to reference spans. Each case was labeled as one of five types: 1) exact match (identical boundaries); 2) partial overlap (some shared tokens, but neither span fully included the other); 3) tgt ⊂ ref (the predicted span was fully contained within the reference); 4) ref ⊂ tgt (the reference span was fully contained within the prediction); and 5) no overlap. The results showed that in nearly half of all cases, the reference span was fully contained within the prediction, as in summarized in Table 5. These findings indicate that while the LLMs are generally successful at identifying the approximate location of the spans, they often extend the boundaries of their predictions beyond those selected by human annotators. However, we used LLMs in a zero-shot setting, which may not have fully exploited their capabilities. Using in-context learning or other prompting techniques (Mao et al., 2025) may enable more accurate identification of key points.

5. Conclusion

In this study, we constructed an L2 English speech corpus with content annotations, which serves as benchmark data for the key point detection task. We analyzed the annotated content elements in relation to human-assigned content scores and found that content performance in speech is related to claims elaborated by more evidence or rebuttal elements and greater clarity. Finally, we demonstrated that while LLMs showed a promising ability to locate content elements in speech transcripts, they still struggled to classify their discourse roles accurately.

6. Ethics Statement

License of this dataset The dataset constructed in this study was developed by adding new annotations to the ICNALE GRA. The original corpus is distributed under its own license[8], and it is prohibited to reproduce and/or redistribute a part or the whole of the ICNALE data. The newly created annotations will be released on the condition that (1) the annotation data is not published on the open web, and (2) the annotation data is not input into generative AI services that use input data for model training, to prevent future data leakage.

Annotations We outsourced the annotations for the ICNALE GRA. We provided annotators with a set of guidelines that explained the purpose and background of this study, detailed annotation procedures, and annotation examples taken from actual learners' data that was not a part of the annotation target. Following the ICNALE terms of use, annotations were applied only to the learner utterances. The annotators were fairly compensated for their work at rates that reflected the time and cognitive effort required for the task.

Anonymization of personal information We used the ICNALE GRA, which is publicly available. The speech data in the ICNALE GRA originally came from the ICNALE Spoken Dialogues Module, whose data is anonymized to prevent the identification of individuals prior to release (Ishikawa, 2018). Our annotations also contain no personally identifiable information.

7. Limitations

Audio data Although the audio data are available in the ICNALE, we constructed our dataset using their transcripts. In automatic speech scoring, end-to-end models that take audio data as input (Bannò and Matassoni, 2023; Lo et al., 2024) have been proposed, and speech foundation models that are trained on a massive amount of speech data and can be adopted to downstream speech tasks are available (e.g., Chen et al., 2024). We reserve the alignment of annotations with the audio data for future work.

Data size We constructed a dataset based on the ICNALE GRA, which includes 140 speech samples. Although the dataset is relatively small, it was intentionally constructed as a carefully curated benchmark with high-quality manual annotations. We demonstrated its usefulness for the key point

[8] https://language.sakura.ne.jp/icnale/index.html#7

detection task, while larger-scale datasets are generally required for model training. However, since manual data creation is costly, one possible direction would be LLM-based data construction, as our experiments with zero-shot settings showed promising results in predicting the spans of content elements.

Automated key point detection models This study's contribution is the construction of a dataset for the key point detection task. We did not aim to develop automatic models, and we used existing LLMs in a zero-shot setting for our experiments. We refined the prompts to encourage instruction following but did not apply advanced prompting techniques or optimization methods (e.g., Mao et al., 2025), which might have affected the LLMs' key point detection performance, especially in exact match. We did not test the models proposed by Wang et al. (2020); Wang and Hamill (2021) either, since they are BERT-based systems that require training and no checkpoints are publicly available. The development of automatic models, as well as the data construction for model training, is left for future research.

In addition, we used LLMs accessible via API, assuming that educational institutions and educators may not have access to sufficient computing resources. However, this raises concerns regarding privacy and security. Developing automated key point detection models using smaller open-source models or open-weight models accessible via APIs with strict data protection policies would be advantageous for real-world applications.

8. Acknowledgements

A part of this work was supported by JST SPRING Grant Number JPMJSP2140.

9. Bibliographical References

Stefano Bannò, Rao Ma, Mengjie Qian, Siyuan Tang, Kate Knill, and Mark Gales. 2025. Natural Language-based Assessment of L2 Oral Proficiency using LLMs. In *Proceedings of the 10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 189–193.

Stefano Bannò and Marco Matassoni. 2022. Cross-corpora experiments of automatic proficiency assessment and error detection for spoken English. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 82–91, Seattle, Washington. Association for Computational Linguistics.

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.

Annie Brown, Noriko Iwashita, and Tim McNamara. 2005. An examination of rater orientations and test-taker performance on english-for-academic-purposes speaking tasks. *ETS Research Report Series*, 2005(1):1–157.

Fu-An Chao and Berlin Chen. 2025. Towards efficient and multifaceted computer-assisted pronunciation training leveraging hierarchical selective state space model and decoupled cross-entropy loss. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1947–1961, Albuquerque, New Mexico. Association for Computational Linguistics.

William Chen, Wangyou Zhang, Yifan Peng, Xinjian Li, Jinchuan Tian, Jiatong Shi, Xuankai Chang, Soumi Maiti, Karen Livescu, and Shinji Watanabe. 2024. Towards robust speech representation learning for thousands of languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10205–10224, Miami, Florida, USA. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Scott A. Crossley, Perpetual Baffour, Yu Tian, Aigner Picou, Meg Benner, and Ulrich Boser. 2022. The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (persuade) corpus 1.0. *Assessing Writing*, 54:100667.

Yuning Ding, Omid Kashefi, Swapna Somasundaran, and Andrea Horbach. 2024. When argumentation meets cohesion: Enhancing automatic feedback in student writing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17513–17524, Torino, Italia. ELRA and ICCL.

Gemini Team Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv*, arXiv:2312.11805.

Shin'ichiro Ishikawa. 2018. Design of the icnale spoken dialogue : For studies of l2 oral production in dialogues. *Learner Corpus Studies in Asia and the World*, 3. (in Japanese).

Shin'ichiro Ishikawa. 2023a. Aim of the icnale gra project : Global collaboration to collect ratings of asian learners' l2 english essays and speeches from an elf perspective. *Learner Corpus Studies in Asia and the World*, 5:121–144.

Shin'ichiro Ishikawa. 2023b. *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English*. Routledge.

Ute Knoch, Judith Fairbairn, and Jin Yan. 2021. *Scoring second language spoken and written performance: issues, options and directions*. Equinox.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Folkert Kuiken and Ineke Vedder. 2018. Assessing functional adequacy of l2 performance in a task-based approach. In Naoko Taguchi and YouJin Kim, editors, *Task-Based Approaches to Teaching and Assessing Pragmatics*, pages 265–285. John Benjamins.

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. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1352–1362, Mexico City, Mexico. Association for Computational Linguistics.

Yuetian Mao, Junjie He, and Chunyang Chen. 2025. *From Prompts to Templates: A Systematic Prompt Template Analysis for Real-world LLMapps*, page 75–86. Association for Computing Machinery, New York, NY, USA.

Pieter Müller, Febe de Wet, Christa van der Walt, and Thomas Niesler. 2009. Automatically assessing the oral proficiency of proficient l2 speakers. In *Proceedings of the Speech and Language Technology in Education (SLaTE 2009)*, pages 29–32.

Ryo Nagata. 2019. Toward a task of feedback comment generation for writing learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3206–3215, Hong Kong, China. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card. *arXiv*, arxiv:2410.21276.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543.

Takanori Sato. 2012. The contribution of test-takers' speech content to scores on an english oral proficiency test. *Language Testing*, 29(2):223–241.

Ekaterina Voskoboinik, Yaroslav Getman, Ragheb Al-Ghezi, Mikko Kurimo, and Tamas Grosz. 2023. Automated assessment of task completion in spontaneous speech for Finnish and Finland Swedish language learners. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, pages 102–110, Tórshavn, Faroe Islands. LiU Electronic Press.

Xinhao Wang and Keelan Evanini. 2020. Features measuring content and discourse coherence. In Klaus Zechner and Keelan Evanini, editors, *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*, pages 138–156. Routledge.

Xinhao Wang, Keelan Evanini, Klaus Zechner, and Matthew Mulholland. 2017. Modeling discourse coherence for the automated scoring of spontaneous spoken responses. In *7th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2017)*, pages 132–137.

Xinhao Wang, Binod Gyawali, James V. Bruno, Hillary R. Molloy, Keelan Evanini, and Klaus Zechner. 2019. Using Rhetorical Structure Theory to assess discourse coherence for non-native spontaneous speech. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 153–162, Minneapolis, MN. Association for Computational Linguistics.

Xinhao Wang and Christopher Hamill. 2021. Automatic generation of diagnostic content feedback

in spoken language learning and assessment. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 579–586.

Xinhao Wang, Klaus Zechner, and Christopher Hamill. 2020. Targeted content feedback in spoken language learning and assessment. In *Interspeech 2020*, pages 3850–3854.

Shasha Xie, Keelan Evanini, and Klaus Zechner. 2012. Exploring content features for automated speech scoring. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 103–111, Montréal, Canada. Association for Computational Linguistics.

Su-Youn Yoon and Chong Min Lee. 2019. Content modeling for automated oral proficiency scoring system. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 394–401, Florence, Italy. Association for Computational Linguistics.

Klaus Zechner and Keelan Evanini, editors. 2020. *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*. Routledge.

10. Language Resource References

Ishikawa, Shin'ichiro and others. 2013. *The International Corpus Network of Asian Learners of English (ICNALE)*. Distributed at the ICNALE project page. PID https://language.sakura.ne.jp/icnale/.