



# Towards explainable automatic spoken language assessment

Marian Marchal<sup>1</sup>, Joey Stuiver<sup>1</sup>, Nafsika Lachana<sup>1,2</sup>, Max van der Velde<sup>1,3</sup>, Josine Verhagen<sup>1</sup>, Joost Kruis<sup>1</sup>

<sup>1</sup>CitoLab, Cito Institute for Educational Measurement, the Netherlands

<sup>2</sup>Information and Computing Sciences, Utrecht University, the Netherlands

<sup>3</sup>Cognition, Data and Education, University of Twente, the Netherlands

marian.marchal@cito.nl; joey.stuiver@cito.nl; nafsika.lachana@cito.nl;  
max.vandervelde@cito.nl; josine.verhagen@cito.nl; joost.kruis@cito.nl

## Abstract

Automatic spoken language assessment can significantly reduce human scoring efforts. A crucial aspect of human scoring is its analytic nature, which makes scores explainable and allows for an evaluation of reliability and validity. However, automatic approaches to spoken language assessment often function as black-box models, lacking transparency in their scoring processes. This research aims to address this limitation by investigating the extent to which we can develop a more explainable automatic scoring model. In this study, we explore how several subconstructs of spoken language can be assessed using features automatically extracted from speech data. Additionally, we examine the predictive relationship between these features and the holistic scores provided by human judges. These findings can inform the development of explainable automated spoken language assessment systems, bridging the gap between efficiency and transparency.

**Index Terms:** explainability, spoken language assessment

## 1. Introduction

Spoken language competence is an important aspect of language proficiency. This competence is assessed in many different environments, including formative assessments in the classroom as well as summative testing for entrance exams for universities. Evaluating spoken language proficiency is a difficult and time-consuming procedure. Automated spoken language assessment could alleviate these efforts. However, it is important that scoring remains reliable and valid [1, 2]. Reliability refers to the consistency of a measure: whether a person with a similar ability is scored similarly independent of the human scorer or test. Validity requires scoring to reflect the true ability of the test taker: test-takers should receive a score for the right reasons. Crucially, validity can only be evaluated if the scoring is transparent, which is not always the case in automated assessment. We therefore examine an explainable approach to automated spoken language assessment.

### 1.1. Spoken language proficiency and assessment

Spoken language proficiency consists of many different components [1]. To communicate successfully, a speaker needs to use correct vocabulary and be able to construct grammatically correct utterances without too many disfluencies. The Common European Framework of Reference (CEFR) [3] distinguishes vocabulary range, grammatical accuracy, vocabulary control, phonological control, coherence and cohesion and fluency as descriptors directly relevant to spoken language proficiency.

Due to the multi-faceted nature of spoken language proficiency, human raters usually assess proficiency using analytic

rubrics, with separate criteria for various sub-constructs [1]. Raters are generally asked to provide separate scores for various aspects, such as fluency, coherence or intelligibility. Such analytic scoring ensures that raters take into account different aspects of spoken language proficiency, although raters may find it difficult to assess each aspect in isolation [1].

Automated spoken language assessment systems differ in a number of ways. On the one hand, these systems vary with respect to the assessed component(s) of spoken language proficiency. Some proposed systems focus on a specific aspect of spoken language proficiency, such as pronunciation or fluency [4, 5, 6]. Other systems focus on predicting holistic scores, combining various aspects of spoken language proficiency [7, 8, 9]. On the other hand, there are different approaches for training automated spoken language assessment systems. One option is to extract features from either the audio signal or (human or machine generated) transcripts [7] or both [9, 8]. These features can target different aspects of language proficiency (e.g. type-token ratio as a measure of vocabulary range), although some features might be influenced by multiple aspects of proficiency (e.g. recording duration, which may be influenced by grammatical and lexical proficiency). Feature extraction might leave some aspects that contribute to spoken language proficiency unnoticed. More recently, systems have been developed that are trained directly on embeddings obtained from speech or language models [10, 7, 5]. Although such approaches may outperform feature-extraction approaches, they are much less explainable, making it difficult to evaluate their validity.

### 1.2. Explainable automated spoken language assessment

Ensuring transparency in assessment significantly enhances the trust and fairness perceived by both test-takers and educators. It is therefore crucial that automated assessment systems are explainable. Nevertheless, automated scoring systems, particularly those utilizing large neural networks, often face challenges in terms of transparency due to their complex and opaque nature, making it difficult to discern the rationale behind specific scores. In other words, it is difficult to ensure that these systems are valid. While these ‘black-box’ models often show a greater capability to capture intricate interactions and potentially improve score prediction accuracy, it remains worthwhile to explore the development of a more transparent and understandable form of automated spoken language assessment. This study investigates the feasibility of leveraging modern techniques to extract features from speech data that are indicative of sub-constructs in spoken language assessment. By predicting holistic scores from these features, we aim to strike a balance between achieving high predictive accuracy and maintaining a

level of explainability that assists stakeholders in comprehending the assessment process.

## 2. The present study

To investigate an explainable approach for automated assessment, we present a case study focused on developing a system for automated spoken language assessment. This system is designed based on human assessment guidelines from the Dutch Integration Exams, which emphasize six key aspects indicative of spoken language proficiency: word use, grammar, fluency, coherence, and pronunciation.<sup>1</sup> The system is designed to extract features that are presumed to reflect these aspects. For each feature, we assess its contribution to predicting an holistic spoken language score. This evaluation enables us to explore the viability of implementing such an approach in high-stakes testing settings for automated spoken language assessment.

## 3. Feature extraction

As described above, the theoretical scoring model consists of several constructs. From the description of these constructs, we can derive multiple aspects that contribute to scores on this construct. For each of these aspects, we discuss how they are operationalized with several features. An overview of constructs, aspects and features can be found in Table 1.

**WORD USE** Scoring high on word use requires speakers to use a large variety of words and to use them correctly. Lexical precision refers to whether the test taker uses the words accurately. This measure is difficult to extract automatically and is not included in some other automated scoring systems [8, 9]. Here, we operationalize lexical precision as LLM surprisal, since words that are not lexically correct would also not be predicted in that context by a language model. In addition, LLM surprisal has been shown to be a strong predictor of the processing difficulty for the speaker [12, 13]. Lexical diversity provides an indication of the size of the test takers' vocabulary and their ability to use a wide variety of words. Several (length-independent) measures of this construct have been proposed [14, 15]. [15] show that abundance is the best predictor of human ratings of lexical diversity, especially for L2 speech. In addition, the assessment form makes explicit that word use should not be restricted to words from the assignment. We therefore expect lower marks when more overlapping word (form)s are used. Finally, lexical complexity assesses whether test takers are able to use more complex words. Frequency has been shown to be a strong predictor of lexical complexity [16].

**GRAMMAR** Similarly to word use, grammatical proficiency is demonstrated by precise, diverse and complex grammar use. With respect to precision, we take into account the number of grammatical errors. Grammatical diversity assesses whether the test taker is able to use different syntactic structures. This would result in different types of dependency relations, such as adjective-noun and verb-object (cf. [17]), which is referred to as dependency diversity. Sentences can differ in their structural complexity. More complex sentences are generally longer and contain more complex grammatical structures. This is operationalized with extracting average sentence and dependency length.

**FLUENCY** Fluency concerns the ability to produce speech at an appropriate rate without disturbing interruptions. Speech

<sup>1</sup>A sixth aspect, which evaluates whether the response is appropriate, is not included here, since inappropriate responses have been removed manually from the dataset described below.

rate has been shown to be the best objective predictor of human ratings of speech fluency [18, 19]. Here, we operationalize speech rate as syllable duration, following [20]. Pauses signal sentence planning processes, especially in the case of within-clause pauses. In addition, the assessment form explicitly states that scorers should assess the number of pauses. We therefore incorporate both the duration of silent pauses [8] and number of silent pauses [21].<sup>2</sup> Furthermore, since pauses can also be filled (e.g. *uhm*), we also include the number of hesitations and disfluencies, as annotated in the transcript [22]. Reformulations are also explicitly mentioned in the scoring model. For practical reasons, we only take into account the number of repetitions.

**COHERENCE** The assessment form operationalizes coherence explicitly as the use of (simple) connectives. In addition, speakers can establish coherence by using referential links between sentences. One measure of coherence is therefore whether referring expressions, such as pronouns, are used.

**PRONUNCIATION** The assessment form asks raters to judge pronunciation based on whether the speech is comprehensible. This is reflected in the ASR confidence score.<sup>3</sup> In addition, proficient speakers are expected to make fewer pronunciation errors.

## 4. Analysis

### 4.1. Data

To evaluate how well these features predict human ratings of spoken language proficiency, we use the Speak and Improve (S&I) Corpus [22] as released for the S&I Challenge [7]. The corpus consists of audio recordings by L2 learners of English with various L1 backgrounds, collected on the S&I platform [23]. L2 learners completed five tasks, including a short interview (part 1), giving their opinion (part 3), describing a graphic (part 4) and a discussion (part 5). For more details on the tasks we refer to [22].

For each part, a human rater has provided a holistic score on a scale of 1-6, corresponding to CEFR levels A1 to C1/C2.<sup>4</sup> Transcriptions of the recordings have been generated with the Open AI Whisper small model [24], as part of the Speak and Improve challenge. The corpus consists of two additional manual annotation layers: The first layer provides a faithful transcription of the audio and includes annotations for disfluencies, hesitations and pronunciation errors. The second layer contains a fluent version of the learner's speech. The audio recording, automatic transcriptions and the manual annotations in the first layer serve as input for the current system. We adopt the same dev/train/eval split as provided for the S&I Challenge.

<sup>2</sup>In line with [21], we set the threshold for a silent pause to 300ms

<sup>3</sup>Note that intelligibility is influenced by multiple factors, including goodness of pronunciation, background noise and word predictability. In future work, we aim to tease these aspects further apart.

<sup>4</sup>Note that the assessment model used by the human raters to score these responses slightly differs from the assessment model for the Dutch Integration Exam described above. More specifically, the subconstructs are categorized differently in the two assessment forms (e.g. language resource comprises both word use and grammar) and the LinguaSkill criteria contains a separate category for stress, rhythm and intonation in the former. Nevertheless, both assessment models are based on the CEFR framework and assess different aspects of speech proficiency, including pronunciation, fluency, lexical and grammatical competence and discourse management.

Construct	Aspect	Feature	Description
Word use	Lexical precision	Surprisal	Average negative log probability of content words according to a pre-trained transformer model (BERT)
	Lexical diversity	Abundance	Total number of different lemmas
		MATTR	Moving average type-token ration using a 25-word window.
Lexical complexity	Lexical complexity	Lemma overlap	Number of lemmas that do not overlap with those in the prompt
		Frequency	Average $\log_{10}$ word form frequency of content words based on SUBT-LEX [11]
Grammar	Gramm precision	Grammatical errors	Number of grammatical errors detected by languagetool
	Gramm diversity	Dependency diversity	Total number of unique dependency relations
	Gramm complexity	Sentence length	Average length of sentences in words
Fluency	Speech rate	Dependency length	Average length of dependency relations
		Syllable duration	Average duration of a syllable
		Interruptions	Silence duration
	Silences	Silences	Number of silent pauses.
		<i>Hesitations</i>	<i>Number of hesitations as annotated in the transcript</i>
		<i>Disfluencies</i>	<i>Number of disfluencies annotated in the transcript</i>
Coherence	Connective usage	Repetitions	Number of repeated unigrams and bigrams.
		Connectives	Number of tokens with the dependency tag <code>mark</code> or <code>cc</code>
	Coreference	Connective diversity	Number of different types of connectives
Pronunciation	Intelligibility	Pronouns	Number of tokens with the POS tag <code>PRN</code>
		Demonstratives	Number of demonstrative pronouns (e.g. <i>this</i> )
	Pronunciation	ASR confidence	Average confidence score provided by the ASR system
		<i>Pronunciation errors</i>	<i>Number of pronunciation errors as annotated in the transcript</i>

Table 1: Description of features used to predict spoken language proficiency in our system. Manual features are in italics.

## 4.2. System

Our system first extracts the features described above from the audio recordings, automatic transcriptions - generated by Open AI’s Whisper small model [24] - and human annotations. LLM surprisal is quantified as the negative log probability of the word being predicted by a bi-directional transformer (`bert-base-uncased`) [25]. For grammatical error detection, we use an open-source off-the-shelf language tool, that is available in many different languages.<sup>5</sup> For pos-tagging and dependency parsing, we used Spacy’s `en_core_web_sm` model. Further specifications can be found in Table 1.

We train an ordinary least squares regression model on the train split of the corpus and predict scores on the eval split. No further variable selection was conducted. Scores are predicted separately for each part, but are averaged to calculate the overall score, in line with the S&I Challenge [7].<sup>6</sup> We compare two models: the first one is trained on all features that can be extracted from the audio (and ASR transcripts) automatically without any human annotation. The second model is trained on all features that can be extracted from the corpus, including the ones that require human annotation. This manual data is available for all responses in the test set, but only for 1,460 responses in the train split, resulting in a considerable decrease in training size compared to the first model. Nevertheless, these features might be important to model fluency and pronunciation. In addition to the main effects of the 21 features described above, we include the interaction between mean length of the sentence and the number of grammatical errors, as well as all fluency

<sup>5</sup><https://pypi.org/project/language-tool-python/>. Repeated words or capitalization errors are excluded.

<sup>6</sup>Although the relation between features and the holistic score might in many cases not be linear, this approach is highly explainable and provides insights into the contribution of each feature. We plan to explore models allowing for non-linear relations and interactions in the future.

and pronunciation measures in both models. These interactions were included since disfluencies and grammatical errors might be less problematic in longer sentences. Finally, task is included as a predictor, because some tasks might be easier than others.

## 5. Results

### 5.1. Feature contribution

To explore how each feature contributes to predicting the score, we scale the predictors to obtain standardized coefficients. These are plotted in Figure 1. ASR confidence is found to be the strongest predictor for both models, which positively predicts human scores. Dependency diversity and abundance also strongly and positively predict human scores, closely followed by MATTR. In addition, the number of pronunciation errors and average frequency of content words negatively predict human scores: producing more pronunciation errors and more frequent content words leads to lower scores. Thus, we find evidence that measures for various aspects of spoken language proficiency are necessary to predict holistic scoring. However, not all features predict scores in the expected direction. The number of grammatical errors positively predicts score, suggesting that a higher number of grammatical errors yields a higher score. Including this measure decreases the validity of the scoring system. We point out that such counter-intuitive relationships can only be detected in an explainable model. Furthermore, the direction of this effect is only positive in the model that includes all features. Possibly, the number of grammatical errors does not explain any additional variance when these measures of fluency are taken into account.

### 5.2. Score prediction

To evaluate how well our model generalizes to new data, we predict scores for the test split of the Speak and Improve corpus. Predictions of the overall score on the test set were evaluated in

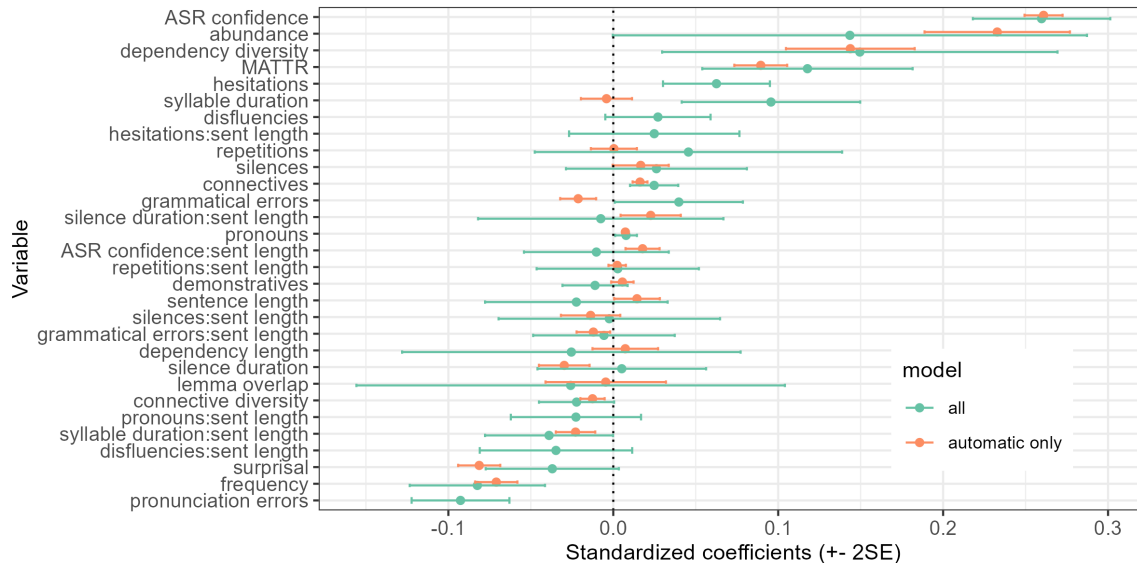


Figure 1: Standardized coefficients for each feature

the same way as for the Speak and Improve Challenge, using root mean squared error (RMSE), Pearson’s correlation coefficient (PCC) and Spearman’s rank coefficient (SRC). In addition, we calculated the proportion of predicted scores that were within 1 (LESS1) or half (LESS5) from the true score. The performance of our model, as well as the baseline model of the S&I challenge described in [7] is provided in Table 2.

On all measures, the model that includes all features slightly outperforms the model including only the automatic features. Both versions of our system are considerably better than the baseline, which uses a much higher number of parameters. Furthermore, despite its simple structure, our model performs on par with models from the closed track with respect to RMSE (1st: 0.368, 2nd: 0.411; 3rd: 0.425).

### 5.3. Error analysis

We observed two main causes for the model to overestimate scores. Firstly, the model achieves high ASR confidence, even in the case of heavily accented speech. Therefore, accented speech does not always result in a lower mark in the system, whereas human raters do seem to assign lower scores to speakers with an accent. Secondly, the model does not always accurately extract the number of grammatical errors. For example, no error was detected in *I am interested to improve my English speaking and listening*. At the same time, the model sometimes underestimated scores, compared to human raters. For example, in some cases the ASR confidence was considerably lower than average due to background noise or false starts, despite being intelligible. In addition, our system predicts lower scores when the audio contains many disfluencies and hesitations, but there are contexts in which these may sound natural to human raters.

## 6. Discussion and conclusion

The present study set out to examine how spoken language proficiency can be assessed automatically while maintaining explainability. Transparency is essential for evaluating scoring validity. Our system automatically extracts features derived from human assessment forms. We also evaluated how these features

	RMSE	LESS1	LESS5	PCC	SRC
S&I baseline	.440	.966	.733	.746	.750
all	.414	.993	.830	.77	.78
automatic only	.422	.980	.810	.76	.77

Table 2: Performance of the models.

predict human marks on a train and test set. Our system explains a significant amount of variance in human marks. This shows that black-box approaches do not necessarily perform better than theoretically informed models with less parameters. Nevertheless, it is currently insufficient for high-stakes testing, requiring further improvements in future work. This also raises questions with respect to the role of automated assessment in these scenarios. Although replacing human raters with an automated scoring system would save a considerable amount of time and effort, to improve reliability, automated scoring can also be combined with human assessment [26] or be used for learner preparation and feedback [27].

For automated scoring to be valid, test takers should receive scores based on meaningful relations between features and proficiency. Our case study shows that a data-driven approach can lead to counterintuitive effects of features on holistic scores. Nevertheless, our work also highlights that such relationships can be examined in detail when the system is explainable. The features in the present study were hypothesized to reflect various sub-constructs of spoken language proficiency. In the present study, we only evaluated how the features predicted holistic scores. We aim to model the relationships between the features and scores on the separate constructs (e.g. ASR confidence~pronunciation) to verify their validity in future work. In addition, valid scoring requires that the extracted features are accurate. Our error analysis shows that this is not always the case (e.g. some grammatical errors are missed). In future work, we plan to improve the operationalization of a number of features and evaluate the accuracy of each feature in more detail.

## 7. References

- [1] N. H. De Jong, "Assessing second language speaking proficiency," *Annual Review of Linguistics*, vol. 9, no. 1, pp. 541–560, 2023.
- [2] J. Xu, E. Jones, V. Laxton, and E. Galaczi, "Assessing L2 English speaking using automated scoring technology: examining automarker reliability," *Assessment in Education: Principles, Policy & Practice*, vol. 28, no. 4, pp. 411–436, 2021.
- [3] Council of Europe, *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Strasbourg: Council of Europe Publishing, 2020, available at [www.coe.int/lang-cefr](http://www.coe.int/lang-cefr).
- [4] B. Lin and L. Wang, "Deep feature transfer learning for automatic pronunciation assessment," in *Interspeech*, vol. 2021, 2021, pp. 4438–4442.
- [5] J. Liu, A. Wumaier, C. Fan, and S. Guo, "Automatic fluency assessment method for spontaneous speech without reference text," *Electronics*, vol. 12, no. 8, p. 1775, 2023.
- [6] Y. Shen, A. Yasukagawa, D. Saito, N. Minematsu, and K. Saito, "Optimized prediction of fluency of L2 English based on interpretable network using quantity of phonation and quality of pronunciation," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 698–704.
- [7] M. Qian, K. Knill, S. Banno, S. Tang, P. Karanasou, M. J. Gales, and D. Nicholls, "Speak & improve challenge 2025: Tasks and baseline systems," *arXiv preprint arXiv:2412.11985*, 2024.
- [8] L. Chen, K. Zechner, S.-Y. Yoon, K. Evanini, X. Wang, A. Loukina, J. Tao, L. Davis, C. M. Lee, M. Ma *et al.*, "Automated scoring of nonnative speech using the speechrater sm v. 5.0 engine," *ETS Research Report Series*, vol. 2018, no. 1, pp. 1–31, 2018.
- [9] Y. Wang, M. J. Gales, K. M. Knill, K. Kyriakopoulos, A. Malinin, R. C. van Dalen, and M. Rashid, "Towards automatic assessment of spontaneous spoken English," *Speech Communication*, vol. 104, pp. 47–56, 2018.
- [10] S. Bannò and M. Matassoni, "Proficiency assessment of L2 spoken English using wav2vec 2.0," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1088–1095.
- [11] M. Brysbaert and B. New, "Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English," *Behavior research methods*, vol. 41, no. 4, pp. 977–990, 2009.
- [12] E. G. Wilcox, T. Pimentel, C. Meister, R. Cotterell, and R. P. Levy, "Testing the predictions of surprisal theory in 11 languages," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1451–1470, 2023.
- [13] A. G. de Varda, M. Marelli, and S. Amenta, "Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data," *Behavior Research Methods*, vol. 56, no. 5, pp. 5190–5213, 2024.
- [14] Y. Bestgen, "Measuring lexical diversity in texts: The twofold length problem," *Language Learning*, vol. 74, no. 3, pp. 638–671, 2024.
- [15] K. Kyle, S. A. Crossley, and S. Jarvis, "Assessing the validity of lexical diversity indices using direct judgements," *Language Assessment Quarterly*, vol. 18, no. 2, pp. 154–170, 2021.
- [16] R. Wilkens, A. D. Vecchia, M. Z. Boito, M. Padró, and A. Villavicencio, "Size does not matter. Frequency does. A study of features for measuring lexical complexity," in *Advances in Artificial Intelligence–IBERAMIA 2014: 14th Ibero-American Conference on AI, Santiago de Chile, Chile, November 24–27, 2014, Proceedings 14*. Springer, 2014, pp. 129–140.
- [17] M.-C. De Marneffe, C. D. Manning, J. Nivre, and D. Zeman, "Universal dependencies," *Computational linguistics*, vol. 47, no. 2, pp. 255–308, 2021.
- [18] C. Cucchiari, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech," *the Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2862–2873, 2002.
- [19] S. Suzuki, J. Kormos, and T. Uchihara, "The relationship between utterance and perceived fluency: A meta-analysis of correlational studies," *The Modern Language Journal*, vol. 105, no. 2, pp. 435–463, 2021.
- [20] N. H. De Jong, R. Groenhout, R. Schoonen, and J. H. Hulstijn, "Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior," *Applied Psycholinguistics*, vol. 36, no. 2, pp. 223–243, 2015.
- [21] N. H. De Jong and H. R. Bosker, "Choosing a threshold for silent pauses to measure second language fluency," in *The 6th workshop on disfluency in spontaneous speech (diss)*, 2013, pp. 17–20.
- [22] K. Knill, D. Nicholls, M. J. Gales, M. Qian, and P. Stroinski, "Speak & improve corpus 2025: an L2 English speech corpus for language assessment and feedback," *arXiv preprint arXiv:2412.11986*, 2024.
- [23] D. Nicholls, K. Knill, M. J. Gales, A. Ragni, and P. Ricketts, "Speak & improve: L2 English speaking practice tool," in *Proceedings of Interspeech 2023*. Sheffield, 2023, pp. 3669–3670.
- [24] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.
- [25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, 2019, pp. 4171–4186.
- [26] L. Davis and S. Papageorgiou, "Complementary strengths? evaluation of a hybrid human-machine scoring approach for a test of oral academic English," *Assessment in Education: Principles, Policy & Practice*, vol. 28, no. 4, pp. 437–455, 2021.
- [27] L. Gu, L. Davis, J. Tao, and K. Zechner, "Using spoken language technology for generating feedback to prepare for the TOEFL iBT® test: A user perception study," *Assessment in Education: Principles, Policy & Practice*, vol. 28, no. 1, pp. 58–76, 2021.