Computers and Education: Artificial Intelligence 8 (2025) 100386

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Beyond accuracy: Multimodal modeling of structured speaking skill indices in young adolescents

Candy Olivia Mawalim [a,*,1], Chee Wee Leong [b], Guy Sivan [c], Hung-Hsuan Huang [d], Shogo Okada [a]

[a] Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan
[b] Educational Testing Service, Princeton, NJ, USA
[c] Vericant.com, Beijing, China
[d] The University of Fukuchiyama, Fukuchiyama, Kyoto, Japan

ARTICLE INFO

Keywords:
Speaking skills
Multimodal
Interpretability
Interview

ABSTRACT

This study introduces a novel method for explainable speaking skill assessment that utilizes a unique dataset featuring video recordings of conversational interviews for high-stakes outcomes (i.e., admission to high schools and universities). Unlike traditional automated speaking assessments that prioritize accuracy at the expense of interpretability, our approach employs a new multimodal dataset that integrates acoustic and linguistic features, visual cues, turn-taking patterns, and expert-derived scores quantifying various speaking skill aspects observed during interviews with young adolescents. This dataset is distinguished by its open-ended question format, which allows for varied responses from interviewees, providing a rich basis for analysis. The experimental results demonstrate that fusing interpretable features, including prosody, action units, and turn-taking, significantly enhances the accuracy of spoken English skill prediction, achieving an overall accuracy of 83% when a machine learning model based on the light gradient boosting algorithm is used. Furthermore, this research underscores the significant influence of external factors, such as interviewer behavior and the interview setting, particularly on the coherence aspect of spoken English proficiency. This focus on an innovative dataset and interpretable assessment tools offers a more nuanced understanding of speaking skills in high-stakes contexts than that offered by previous studies.

1. Introduction

The ability to communicate effectively in English is crucial for both academic success and professional advancement in today's globalized world. Spoken English proficiency is a particularly important aspect of communication, as it allows individuals to participate in real-time conversations and exchange ideas fluently. Traditionally, spoken English proficiency has been assessed through human-administered tests, which are both time-consuming and expensive. However, the development of automated spoken English assessment systems has the potential to revolutionize the way in which we evaluate language skills (Wang et al., 2018).

An earlier study by Cheng et al. (Cheng et al., 2014) explored automatic scoring methods for spoken responses in the Arizona English Language Learner Assessment and reported high correlations between machine and human scores across various age groups ranging from 4 to 11 years. Additionally, they specified predefined criteria to evaluate speech quality objectively for automatic scoring. The criteria include pronunciation, fluency, and prosody to encompass aspects such as vocabulary choice (Yoon et al., 2012) and sentence structure (Bernstein et al., 2010a). These criteria can help reduce bias and ensure consistency in scoring.

The automatic assessment of communication skills across various interaction settings has been a well-researched area. In monolog scenarios, studies have concentrated on analyzing public speaking skills through features such as eye contact, gesture usage, and voice control (Wortwein et al., 2015; Chen et al., 2014). With respect to dyadic interactions, research has focused on job interviews, where speech content, prosody, and body language features have been extracted to predict qualities such as hireability and interpersonal skills (Nguyen et al., 2014; Chen

* Corresponding author.
E-mail address: candylim@jaist.ac.jp (C.O. Mawalim).
[1] Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, Japan.

https://doi.org/10.1016/j.caeai.2025.100386
Received 17 September 2024; Received in revised form 3 February 2025; Accepted 7 March 2025

Available online 20 March 2025

et al., 2017; Ohba et al., 2022). Additionally, research has examined social skills in group interactions, particularly leadership, by analyzing features such as speaking turns and nonverbal cues (Sanchez-Cortes et al., 2013; Okada et al., 2016). While monolog-based speaking assessments are prevalent, they lack real-world relevance, particularly in online learning environments where teacher–student (dyadic or group) interaction is crucial. However, existing research on automated speaking skills assessment in such interactive settings remains limited and focuses primarily on improving accuracy using single modalities such as text or audio (Townshend et al., 1998; Cheng et al., 2014; Yoon et al., 2012; Bernstein et al., 2010a).

While existing research on automated speaking skills assessment in interactive settings has shown promising results in improving accuracy, it often relies on complex models that are difficult to interpret. This overreliance on black-box models presents significant limitations. First, it hinders our understanding of the intricate interplay of factors that contribute to effective spoken communication, such as fluency, vocabulary, grammar, and pronunciation. Second, the lack of model transparency severely limits the ability to provide meaningful and actionable feedback to learners and instructors. To address these limitations, in this paper, we propose an approach that leverages a multioutput learning framework (Xu et al., 2019). By simultaneously predicting multiple aspects of spoken performance (e.g., fluency, accuracy, complexity) while considering multimodal cues (e.g., audio, video, text), our approach aims to achieve a balance between prediction accuracy and model interpretability. This framework allows for the explicit modeling of the relationships between different aspects of speaking proficiency, thereby enhancing our understanding of the underlying factors and enabling the generation of more informative feedback for learners and educators.

Furthermore, we analyze multimodal cues across the entirety of speaking assessment criteria within open-ended interview settings. Unlike previous studies, which often rely on fixed-question formats (Saeki et al., 2021), our approach allows for a more naturalistic and comprehensive assessment of spoken communication skills. This approach enables the capture of a broader range of abilities, including sociolinguistic competence and the ability to engage in spontaneous and creative communication, which are more naturally exhibited during open-ended interactions (Davis & Norris, 2024).

The expected outputs of our study focus on the following three research questions (RQs):

- (RQ_1) How can spoken English scores be accurately predicted while maintaining model interpretability?
- (RQ_2) Which cues contribute most significantly to accurately predicting speaking skills?
- (RQ_3) How do external factors influence interviewee performance?

Unlike previous works, we address our research questions using a novel dataset of structured Vericant interview sessions[2] that are specifically designed for high school students applying to U.S. universities. This dataset offers a reliable assessment of critical spoken communication skills in a *high-stakes, open-ended interview scenario*. By leveraging open-ended questions, our findings hold greater generalizability by capturing a broader range of speaking abilities. Furthermore, we delve into the interpretability of the assessment process by analyzing how information from multimodal cues and external factors contributes to the evaluation of speaking skills.

2. Related work

This section reviews the literature related to automatic spoken English assessment, challenges in the discourse context, and interpretability.

[2] https://www.vericant.com/see/.

2.1. Automatic spoken English assessment

Automated spoken English assessment has evolved significantly, driven by the increasing demand for efficient and objective language proficiency evaluation. This progression has led to a shift from traditional methods to sophisticated AI-driven approaches.

Automated speaking tests have demonstrated strong validity, accurately reflecting a person's spoken communication ability. Studies, such as Bernstein's research on 'facility-in-L2' tests (Bernstein et al., 2010b), have shown high correlations between automated assessments and traditional oral proficiency interviews, indicating their reliability across diverse languages, including English. These tests effectively measure both receptive and productive language skills by requiring meaningful language use.

To optimize assessment accuracy, a hybrid approach that combines both human and automated scoring has emerged (Yoon & Zechner, 2017). This approach leverages automated systems to score the majority of responses, whereas human raters focus on complex cases identified by filtering systems. This collaborative method has demonstrated significant improvements in scoring performance and validity.

The advent of deep learning and advanced speech recognition systems has revolutionized automated spoken English assessment (Wang et al., 2018). These systems effectively handle the variability and disfluency characteristics of nonnative speech, providing scores comparable to those of human examiners. Techniques such as Gaussian process grading and interpolation with human grades further enhance the accuracy and reliability of these automated assessments.

2.2. Challenges in discourse context

The rise of online education has significantly increased the demand for automated assessment tools that can evaluate speaking skills in interactive settings, such as dialogs (Eskenazi, 2009). However, analyzing conversations presents unique challenges. First, the inherent variability in speaking patterns and responses during dialogs increases the complexity of assessment (Oliveri & Tannenbaum, 2019). Second, the limited availability of large, dialog-based datasets often restricts research to using single modalities (text or audio), hindering a comprehensive understanding of communication within interactive environments (Saeki et al., 2021).

These challenges underscore the need for innovative approaches that can capture the dynamic nature of conversational interactions. Recent research has emphasized the importance of examining speaking proficiency within its natural discourse context, particularly for young learners, as exemplified by the work of Firth and Wagner (Firth & Wagner, 2017). This approach necessitates the development of assessment tools that analyze not only the content of speech but also the way in which it unfolds in interactive settings, including turn-taking, back-and-forth exchanges, and the use of language to negotiate meaning.

To address these challenges, researchers have developed automated systems for assessing conversational speaking skills, which involve interactions with an interlocutor (McKnight et al., 2023). These systems utilize advanced models to process audio and text data, providing accurate assessments of conversational proficiency. By leveraging these advancements and addressing the limitations of current approaches, researchers can develop more robust and effective automated assessment tools for evaluating conversational skills in the context of online education and beyond.

2.3. From speaking accuracy to interpretability

The Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2018) provides a comprehensive framework for assessing language proficiency across Europe and beyond. The CEFR provides a framework for assessing language proficiency across six levels

(A1 to C2). It outlines key components of language ability, namely, lexical range, grammatical accuracy, fluency, pronunciation, interactional competence, and coherence.

- **Lexical Range:** Refers to the breadth and depth of the vocabulary used.
- **Grammatical Accuracy:** Assesses the correct use of grammatical structures.
- **Fluency:** Measures the smoothness and ease of speech.
- **Pronunciation:** Evaluates the clarity and accuracy of pronunciation.
- **Interactional Competence:** Focuses on effective communication and interaction.
- **Coherence:** Assesses the logical organization and connection of ideas.

Despite the comprehensiveness of the CEFR, many automated spoken English assessments have historically focused on limited aspects, such as grammar correctness or fluency, and often rely on simplistic metrics such as Likert scales. This limited scope hinders the ability of these systems to provide nuanced feedback aligned with the full spectrum of speaking skills outlined in the CEFR.

Recent efforts have been aimed at improving the interpretability of automated assessments. Gretter et al. (Gretter et al., 2019) demonstrated this by defining automatic scores on the basis of low-level proficiency indicators, such as lexical richness, syntax correctness, pronunciation quality, discourse fluency, and semantic relevance to the prompt, i.e., indicators that are typically aligned with human expert evaluations of language proficiency.

Saeki et al. (2021) (Saeki et al., 2021) contributed by creating a large dialog-based interview dataset for assessing English proficiency in Japanese learners. Their approach aligns with the CEFR framework. They further proposed a neural network model that integrates audio, text, and visual cues for dialog-based assessment. However, the model's reliance on predefined questions and predefined features (such as a language model posterior for grammatical accuracy) raises concerns about potential feature dependence, which may limit its generalizability and impact overall assessment accuracy. Furthermore, the dataset exhibited a moderate level of interrater reliability (Krippendorff's α 0.56–0.75, average 0.65), indicating some variability in human expert annotations. This variability can introduce noise into the dataset and pose challenges for training and evaluating robust automated assessment systems.

Another significant limitation of existing works lies in their lack of interpretability. This lack of transparency hinders our understanding of the spoken communication process, which limits our ability to provide meaningful feedback to learners and instructors. To address this challenge, we propose a multioutput learning framework for automatic comprehensive speaking skill assessment (Xu et al., 2019). By simultaneously predicting multiple aspects of speaking performance (e.g., fluency, accuracy, complexity) while considering multimodal cues (e.g., audio, video, text), our approach aims to improve both prediction accuracy and model interpretability. This framework allows for a deeper understanding of the intricate relationships between different aspects of speaking proficiency, ultimately enabling the generation of more informative and actionable feedback for learners and instructors.

In summary, our review of previous research identifies limitations in automated assessment of spoken proficiency, particularly within discourse contexts. Existing systems often rely on restricted assessment labels, hindering their ability to comprehensively capture the nuances of spoken English in interactive environments. This study addresses these shortcomings by proposing a novel approach that engages in the following tasks:

- **Models key aspects of spoken English skills:** Beyond traditional metrics, we focus on a broader range of skills, including fluency, accuracy, complexity, and discourse management.
- **Incorporates multimodal analysis and a multioutput learning framework:** We leverage audio, video, and text data to gain a more holistic and comprehensive understanding of multiple aspects of speaking performance.
- **Utilizes a diverse dataset:** We employ a multimodal dataset that includes both remote and in-person settings, allowing for a more robust and generalizable model.

This multifaceted approach aims to advance the field of automated speaking proficiency assessment by providing more comprehensive, interpretable, and reliable evaluations.

3. Multimodal spoken English evaluation dataset

To facilitate the development and evaluation of automatic comprehensive speaking skill assessment systems, we introduce a novel multimodal spoken English evaluation (SEE) dataset. This dataset comprises a diverse collection of spoken English interactions, capturing a wide range of interview content and proficiency levels. The dataset includes synchronized audio, video, and text transcripts, providing multimodal information for analysis. A detailed description of the dataset collection process, interview content, and SEE assessment annotation procedures is presented in the following section.

3.1. Data collection

The multimodal dataset comprises video recordings of interviews conducted between expert communication evaluators (interviewers) and high school students preparing for U.S. university applications (interviewees). The interviewees, whose native languages are non-English, were evenly distributed across gender (female: 210, male: 236) and age (9–16). To capture a variety of communication contexts, the interviews were conducted in both remote (102 clips) and in-person (344 clips) settings. The data collection received ethical approval from the authors' institutional committees.

In addition to the video data, the dataset also includes the extracted audio and a transcript generated by AWS Amazon diarization[3] tool. Amazon diarization is a feature within the AWS Amazon Transcribe service that automatically identifies and separates different speakers in an audio recording. The resulting transcript from the diarization tool has been redacted to remove any personally identifiable information (PII). Importantly, the transcript includes timestamps that mark the beginning and end of both the interviewers' and the interviewees' turns, allowing for a detailed analysis of the conversation flow.

3.2. Interview content

The interviews provide a well-rounded assessment of personal qualities, communication styles, and career goals. The following is a breakdown of the key areas explored:

- **Leisure Activities:** The interviews delve into the interviewees' hobbies and how they spend their free time, revealing insights into their interests and personalities.
- **Image Description:** By presenting the interviewees with an image and asking them to describe it, the interviewer assessed their observational skills, critical thinking abilities, and ability to interpret visual cues.
- **Social Dynamics:** Through discussions about their friendships, the interviewees showcased their communication style, ability to collaborate, and approach to conflict resolution.

[3] https://docs.aws.amazon.com/transcribe/latest/dg/diarization-output-batch.html.

- **Personality Traits:** The interviewees responded to a prompt that revealed their awareness of the complexities of emotional resilience, including both its advantages and potential drawbacks.
- **Career Aspirations:** The interviews provide a clear picture of the interviewees' career goals, including their motivations and long-term vision for their professional life.

The abovementioned explored areas serve as a sample; the specific content may vary depending on the interview flow and the interviewees' responses.

3.3. SEE assessment criteria and annotation procedures

The SEE score[2] goes beyond simply measuring grammar and vocabulary knowledge and instead focuses on how well someone can use language for social purposes. The SEE score is a measure of a person's spoken English proficiency during a Vericant interview. Unlike traditional tests that focus on grammar and vocabulary, the SEE score evaluates how well a person can communicate in a real-life conversation.

The SEE score ranges from 1 (minimal proficiency) to 6 (native speaker level). Applicants receive a detailed report explaining their score and how it reflects their strengths and weaknesses in different communication areas. This score provides valuable information to both schools and applicants. Schools can use it to assess an applicant's ability to participate effectively in an English-speaking academic environment. For applicants, the SEE score offers clear feedback on their spoken English skills and helps them identify areas for improvement.

The SEE score considers the following five key aspects:

- **Range:** Using appropriate vocabulary for complex ideas,
- **Accuracy:** Forming grammatically correct sentences,
- **Fluency:** Speaking smoothly and naturally,
- **Interaction:** Actively participating in the conversation, and
- **Coherence:** Clearly conveying ideas with logical flow.

To establish the ground truth for the SEE scores, two expert human annotators independently evaluated each interview video clip using a predefined scoring rubric that considered all the subvariables listed in Table 1. The annotators were native English speakers from a native English-speaking country, and they were proficient in both grammar and syntax. They demonstrated proficiency in English, as evidenced by receiving an SEE score of 6 during the interview process.

To ensure consistency in the scoring process, we calculated Krippendorff's α (Hayes & Krippendorff, 2007), which is a measure of inter-rater reliability. The results, which are presented in the Krippendorff's α column of Table 1, demonstrate excellent agreement between the annotators (averaging $\alpha = 0.8$ across all subvariables). The final SEE score, which is calculated by averaging the raters' assessments on a scale ranging from 1 to 6 (with decimals), reflects a speaker's overall communication ability. A score of 1 indicates very limited conversation skills, whereas a score of 6 signifies the proficiency of a fluent and articulate speaker.

The dataset consists of 267 samples with high SEE scores and 179 samples with low SEE scores. These categories were defined using a threshold SEE score of 4:

$$y = \begin{cases} 0 & \text{SEE-score} < 4 \\ 1 & \text{otherwise} \end{cases} \quad (1)$$

where y is the target label for the binary classification task. The threshold of 4 for binary classification aligns with SEE score definitions (4 or above: proficient; 3: intermediate; 2 or below: beginner). The distribution of target labels for the regression task is illustrated in Fig. 1.

Table 1
Description of SEE assessment criteria.

Aspect	ID	Description	Krippendorff's α
Range	R1	Range of topics	0.801
	R2	Range of vocabulary	0.777
	R3	Circumlocution	0.842
	R4	Precision of language	0.823
Accuracy	A1	Sentence structure	0.832
	A2	Subject-verb agreement	0.735
	A3	Pronouns	0.659
	A4	Tenses	0.836
	A5	Conjugation and prepositions	0.817
Fluency	F1	Accent	0.726
	F2	Tempo and pausing	0.822
	F3	Intonation and fluidity	0.826
	F4	Free speech	0.842
Interaction	I1	Participation	0.630
	I2	Conversational ease	0.837
	I3	Clarifications	0.836
	I4	Conversational cues	0.830
Coherence	C1	Conversational planning	0.848
	C2	Details	0.839
	C3	Rambling	0.842

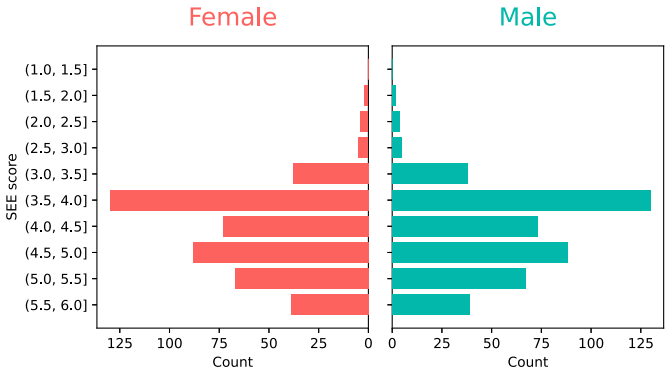


Fig. 1. SEE score distribution according to sex: female (210 clips) and male (236 clips).

4. Method

Fig. 2 shows our approach to predicting spoken English evaluation scores by leveraging a combination of multimodal features extracted from interview video data. A multioutput machine learning algorithm was trained on an expert-rated SEE dataset to learn complex patterns associated with different proficiency levels and subvariables in the SEE assessment criteria.

4.1. Feature extraction

A multimodal feature set was employed for SEE score prediction, encompassing acoustic, linguistic, visual, and turn-taking attributes. Table 2 provides a detailed summary of these feature modalities.

4.1.1. Acoustic feature

[WavLM] For the acoustic feature, we employed the waveform language model (WavLM) as described by Chen et al. (Chen et al., 2022), which represents a recent advancement in speech processing. Given the varied durations of the input videos in the multimodal dataset, ranging from 10 to 20 minutes, we extracted the output of the embedding layer of WavLM, comprising 1,024 units. This extraction was performed using each utterance as the input. To standardize the feature length across all videos, we unified the length through a combination of clipping and zero padding, resulting in a final length of 66,560 units. For longer wav

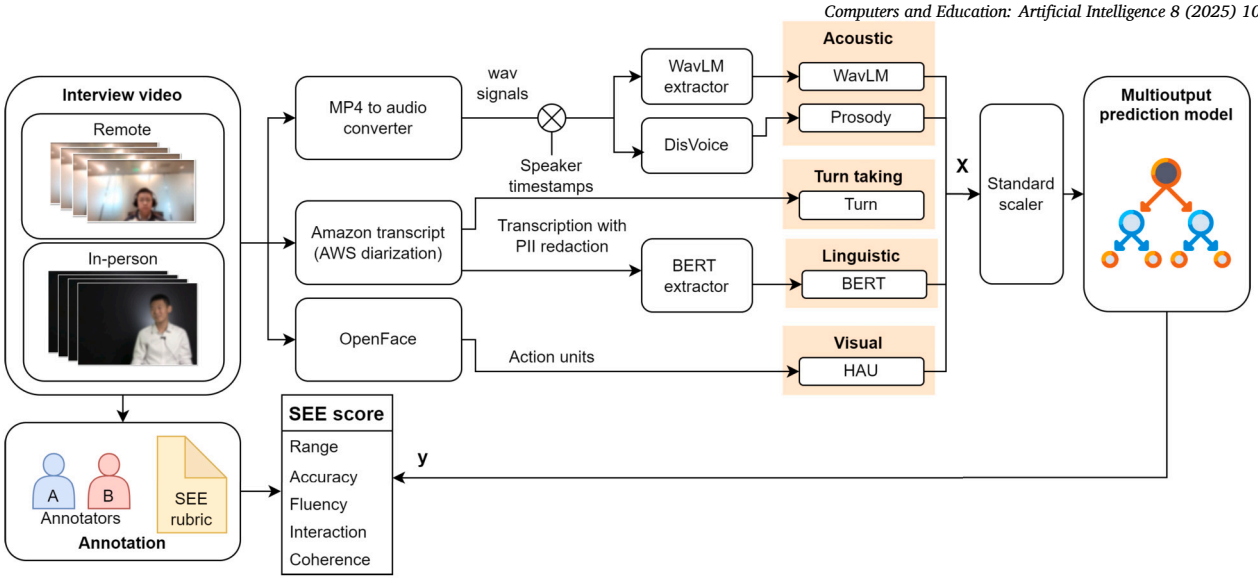


Fig. 2. Overview of the proposed automatic SEE prediction. (Note: The images of learners are for illustrative purposes only and do not necessarily represent the actual demographic breakdown of the research participants. The participants' faces were intentionally blurred to protect their privacy and comply with ethical guidelines.)

Table 2
Summary of extracted multimodal features.

Modality	Feature name	#Feature	Description (feature index)
Acoustic	WavLM (Chen et al., 2022)	1024	Waveform language model
	prosody (Dehak et al., 2007; Vásquez-Correa et al., 2018)	103	F0-based features (1-30) Energy-based features (31-78) Duration-based features (79-103)
Linguistic	BERT (Lee et al., 2024; Li & Li, 2023)	1024	Bidirectional encoder representations from transformers
Visual	HAU	18	Histogram of AUs extracted using OpenFace (Amos et al., 2016)
Turn-taking	speaker turn's utterances	6	Count, sum, mean, standard deviation, skewness, kurtosis

files, we centered the clip around the audio by determining the midpoint, which was calculated as half of the difference between the total length of the original wav file and the target length of 66,560 units. For shorter wav files, we simply zero-padded the insufficient units.

[Prosody] To analyze spoken English proficiency comprehensively, we extracted 103 prosody features from the continuous speech segments in the interview recordings. These features can capture how a speaker utilizes pitch (fundamental frequency or F0), energy (loudness), and duration to deliver spoken language. We employed DisVoice,[4] which is a well-established tool, to extract these features efficiently. The specific feature set we employed was informed by prior research on prosodic analysis for speaker verification (Dehak et al., 2007) and the evaluation of speech disorders (Vásquez-Correa et al., 2018).

For pitch, we computed statistics such as average, standard deviation, maximum, minimum, skewness, and kurtosis, not only for the overall F0 contour but also for specific segments such as the first and last voiced parts. Similarly, we analyzed energy features for both voiced and unvoiced segments, providing insights into the speaker's volume control and emphasis patterns. Additionally, we extracted various duration-based features, including the rate of voiced speech, average and variability in the length of voiced and unvoiced segments, and the ratios between pause durations and voiced/unvoiced speech durations. These comprehensive prosody features offer valuable information about the speaker's fluency, intonation, and overall delivery style, which can be crucial for understanding spoken English proficiency.

4.1.2. *Linguistic feature*

[BERT] We leveraged state-of-the-art (SOTA) text features for each chunk using bidirectional encoder representations from transformers (BERT). BERT is a pretrained deep learning model that excels at capturing the meaning and context of textual information. We used the pretrained model of BERT[5] (Lee et al., 2024; Li & Li, 2023). In March 2024, this BERT model achieved outstanding results on the massive text embedding benchmark leaderboard, outperforming commercially available options such as OpenAI's text embedding-3-large and even matching the performance of much larger models such as the echo-mistral-7b, which is 20 times larger.

4.1.3. *Visual feature*

[Histogram of Action Units (HAU)] We utilized OpenFace (Amos et al., 2016) to extract the AUs from a given video clip. Each AU corresponds to a specific muscle movement in the face, and the histogram visually represents the frequency of occurrence for each AU throughout the interview. By analyzing the histogram of each AU, we can identify patterns in the speaker's emotional state and expressivity. For example, a high frequency of AU 1 (raised eyebrows) might suggest moments of surprise or confusion, whereas frequent AU 12 (lip corner pull) could indicate smile or positive engagement. Examining these patterns in conjunction with the spoken content and other features can offer a more comprehensive understanding of the speaker's communication style and their potential impact on the SEE score.

[4] https://github.com/jcvasquezc/DisVoice/tree/master.

[5] https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1.

4.1.4. Turn-taking feature

To capture the conversational dynamics between the interviewer and interviewee, we extracted various statistical properties from their turn-taking patterns. These properties were derived from the start and end timestamps of each speaker's utterances within conversation chunks (utterances). The features included count (total number of utterances), mean, standard deviation, skewness (distribution asymmetry), and kurtosis (peakedness of the distribution) of the utterance duration. By analyzing these features, we aim to better understand how factors such as conversation flow, speaking pace, and turn-taking balance might influence the overall SEE score.

4.2. Machine learning model

We propose a multioutput light gradient boosting model (LightGBM) for SEE score prediction. Multioutput learning has emerged as a valuable tool for enhancing the interpretability of AI models (Xu et al., 2019). This approach involves simultaneously training a model on a set of interconnected labels, such as accuracy, fluency, and coherence in speaking skills. By analyzing how the model utilizes information from various modalities (audio, text, visuals) for each individual task, we can obtain insights into the specific cues that contribute to each assessed speaking skill. This heightened explainability is particularly valuable in spoken English assessment, as it empowers educators and learners to pinpoint areas of strength and weakness with greater precision. Ultimately, this approach leads to more targeted instruction and facilitates the development of personalized learning experiences.

The objective function of the multioutput LightGBM can be expressed as follows:

$$\min L(Y, F(X)), \tag{2}$$

where:

$$F(X) = \Sigma_{m=1}^{M} F_m(X) \tag{3}$$

where $F_m(X)$ is the prediction for the m-th target variable, and N and M are the number of samples and target variables, respectively. X is the feature matrix of size $N \times P$. Y is the target matrix of size $N \times M$. F is the ensemble of decision trees. Finally, L is the loss function.

Compared with traditional deep learning models, we subsequently chose the LightGBM since it offers several advantages, including faster training times, lower memory usage, and potentially superior accuracy in certain tasks (Ke et al., 2017). The original LightGBM is an algorithm that combines a gradient boosting decision tree (GBDT), gradient-based one-sided sampling (GOSS), and exclusive feature bundling (EFB). It has been reported to work well on multiple public datasets and can reduce the training process by more than 20 times with almost similar accuracy.

In our preliminary experiments in which a subset of development data was used to compare various machine learning models, the LightGBM model indeed achieved the highest accuracy in all unimodal settings among the evaluated methods (CNN, CNN-LSTM, and LightGBM). Owing to this strong performance and its ability to provide information on feature importance, we opted to focus on the LightGBM model for our proposed method.

5. Experiment

This section details our experiments designed to validate the effectiveness of our proposed multioutput learning approach for predicting SEE scores. We aim to answer three key research questions addressed in Section 1, with the following proposed approaches:

- **(RQ1):** We investigate not only the inherent difficulty of predicting SEE scores but also how multioutput learning on multiple speaking skill indices can contribute to a more interpretable assessment of spoken English proficiency.
- **(RQ2):** We identify the features that have the strongest correlation with the SEE score and its associated assessment criteria (fluency, range, coherence, etc.).
- **(RQ3):** We examine two external factors that might affect interviewee performance. The first factor is the interviewer's features. The second factor is the interview setting, i.e., remote or in person.

5.1. Experiment settings

As mentioned in Section 4, we utilized a multioutput LightGBM for SEE score prediction on a multimodal SEE dataset (Section 3.1). All features were extracted on the basis of the utterance obtained from diarization transcription. We carried out a hyperparameter tuning process using a grid search algorithm as part of the cross-validation procedure to achieve optimal model performance. This process involved adjusting key parameters, such as the learning rate (set to 0.1) and the maximum number of leaves (set to 30). A gradient-boosting decision tree algorithm was employed, utilizing binary log-loss as the loss function for the classification task (identifying high versus low proficiency) and RMSE for the regression task (predicting the exact SEE score). The remaining hyperparameters were set to the default values provided by the LightGBM library[6] to ensure consistency and facilitate model interpretability.

We also leveraged multioutput learning, which is an approach in which a single model tackles multiple related labels simultaneously. The model analyzes not only the final SEE score but also intermediate factors in assessing spoken English proficiency. As mentioned in Section 3.1, the dataset comprises remote and in-person interviews. To account for these settings and prevent bias, we used the standard scaler technique. This technique normalizes each feature by subtracting the mean and dividing by the standard deviation.

5.2. Evaluation

To ensure the model's generalizability and robustness, we conducted extensive experiments using both binary classification and regression. Fivefold cross-validation with three repetitions was employed for evaluation. Furthermore, stratified cross-validation (Diamantidis et al., 2000) was used to ensure that each fold maintained the same class distribution as the entire dataset, which is crucial for imbalanced datasets such as spoken English scores.

To comprehensively assess the classification performance, we employed two metrics, namely, accuracy (overall percentage of correct predictions) and macro F1 (combining precision and recall for a balanced view, which is especially important for imbalanced datasets). We used both metrics to understand the overall performance across all classes and in an imbalanced setting. During training, we utilized binary cross-entropy loss. This common function penalizes the model for incorrect predictions, guiding it toward better classification of high and low proficiency.

The regression experiment follows a similar approach but has two key differences in evaluation. Instead of stratified cross-validation, we used random cross-validation, as the SEE score is a continuous value. Additionally, the loss function for regression uses the RMSE (root mean squared error) to measure the difference between the predicted and actual scores. We assessed regression model performance using Pearson correlation (ρ) and RMSE.

5.3. Results

We carried out three main analyses to answer our research questions. First, we compared the efficacy of unimodal and multimodal approaches. Second, we investigated the features and the sequences that exert the strongest influence on SEE scores. This analysis provided

[6] https://lightgbm.readthedocs.io/en/latest/index.html.

Table 3
Experimental results using unimodal feature. We compared the results of binary classification and regression using the state-of-the-art features, i.e., WavLM and BERT, and also more interpretable conventional features, i.e., prosody, HAU, and speaker turn's utterances ("Turn"). The description of the "Target" column is detailed in Table 1. The corresponding multicomparison pairwise significant tests are shown in Appendix A and Appendix B.

Target	Binary classification										Regression									
	WavLM		Prosody		BERT		HAU		Turn		WavLM		Prosody		BERT		HAU		Turn	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	ρ	RMSE	ρ	RMSE	ρ	RMSE	ρ	RMSE	ρ	RMSE
SEE	81.98	**81.21**	**82.05**	81.17	79.52	77.72	65.10	62.87	81.84	81.00	**0.666**	**0.346**	0.618	0.384	0.608	0.395	0.309	0.948	0.622	0.400
R1	**81.54**	**79.93**	81.46	79.77	76.54	72.93	67.64	63.91	78.10	76.41	**0.641**	**0.441**	0.609	0.474	0.590	0.495	0.386	0.674	0.625	0.477
R2	**76.60**	**76.45**	75.40	75.27	73.10	72.85	59.57	58.82	72.50	72.29	**0.607**	**0.388**	0.554	0.423	0.533	0.442	0.364	0.555	0.470	0.522
R3	**76.09**	**75.68**	74.59	74.15	68.99	68.00	61.67	60.38	72.43	71.99	**0.625**	**0.394**	0.544	0.457	0.509	0.479	0.355	0.595	0.533	0.493
R4	**76.75**	**76.36**	73.91	73.50	71.45	70.63	60.40	59.04	72.27	71.91	**0.604**	**0.422**	0.534	0.474	0.527	0.479	0.333	0.626	0.493	0.540
A1	71.67	**67.99**	69.43	64.82	**71.90**	64.39	65.10	58.94	67.19	63.44	**0.560**	**0.476**	0.474	0.54	0.536	0.494	0.271	0.909	0.429	0.611
A2	69.29	68.83	69.43	68.97	**70.25**	**69.40**	61.66	59.99	68.69	68.26	0.502	0.491	0.448	0.527	**0.517**	**0.480**	0.249	1.000	0.414	0.591
A3	**78.47**	**76.06**	76.53	73.45	74.58	68.87	64.05	58.84	71.68	68.55	**0.535**	**0.463**	0.491	0.494	0.528	0.469	0.265	0.943	0.407	0.588
A4	72.04	**66.16**	71.01	64.49	**73.69**	63.93	65.85	55.78	68.01	62.31	**0.540**	**0.528**	0.502	0.557	0.518	0.549	0.215	0.917	0.411	0.685
A5	69.06	64.63	71.00	**66.32**	**71.74**	63.74	66.00	58.49	66.89	62.79	**0.514**	**0.476**	0.455	0.511	0.486	0.489	0.265	0.907	0.400	0.594
F1	72.94	70.49	73.61	70.87	**74.30**	70.90	64.05	60.80	73.32	**71.70**	**0.548**	**0.451**	0.499	0.485	0.510	0.474	0.279	0.958	0.453	0.548
F2	77.36	77.14	**77.43**	**77.22**	71.82	71.29	60.99	60.31	75.04	74.75	**0.658**	**0.400**	0.627	0.43	0.555	0.486	0.294	0.979	0.611	0.461
F3	76.61	76.37	**77.20**	**77.03**	71.37	70.91	61.66	61.12	73.10	72.85	**0.594**	**0.441**	0.592	0.444	0.563	0.467	0.303	0.979	0.548	0.505
F4	**84.53**	**84.04**	83.86	83.27	75.26	73.59	67.94	66.36	83.85	83.28	0.682	**0.398**	0.678	0.402	0.596	0.484	0.360	0.908	**0.688**	0.412
I1	81.08	**80.84**	**81.09**	80.74	75.71	75.06	64.28	63.70	**81.09**	80.78	0.674	**0.394**	0.651	0.417	0.610	0.454	0.346	0.936	**0.679**	0.402
I2	**81.69**	**81.48**	81.32	81.10	74.74	74.31	63.75	63.31	81.17	80.93	**0.691**	**0.390**	0.662	0.418	0.563	0.508	0.324	0.969	0.688	0.404
I3	**80.12**	**79.61**	78.48	77.93	77.88	76.87	65.62	64.59	77.44	76.78	**0.669**	**0.416**	0.634	0.452	0.629	0.455	0.330	0.948	0.642	0.464
I4	78.92	72.13	**82.81**	**77.65**	76.61	63.29	71.00	59.71	78.55	73.27	**0.544**	**0.384**	0.527	0.395	0.468	0.424	0.279	0.834	0.529	0.412
C1	**77.20**	**76.76**	77.19	76.65	68.00	66.77	61.21	59.48	74.22	73.76	**0.636**	**0.422**	0.591	0.461	0.560	0.487	0.232	1.025	0.581	0.494
C2	**78.55**	**78.40**	78.39	78.22	70.85	70.67	61.59	61.24	77.65	77.48	**0.662**	**0.386**	0.614	0.428	0.585	0.458	0.248	1.049	0.630	0.434
C3	**73.69**	**73.41**	71.67	71.46	66.74	66.31	58.90	58.27	71.23	71.00	**0.554**	**0.439**	0.517	0.461	0.472	0.487	0.223	1.030	0.505	0.496

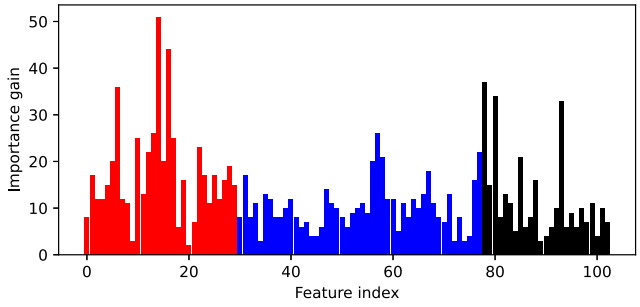
Table 4
Experimental results using multimodal features of interpretable conventional features. The description of the "Target" column is detailed in Table 1. The corresponding multicomparison pairwise significant tests are shown in Appendix A and Appendix B.

Target	Binary classification								Regression							
	Prosody-HAU		Prosody-Turn		HAU-Turn		Prosody-HAU-Turn		Prosody-HAU		Prosody-Turn		HAU-Turn		Prosody-HAU-Turn	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	ρ	RMSE	ρ	RMSE	ρ	RMSE	ρ	RMSE
SEE	81.83	80.93	82.66	81.78	81.46	80.63	**83.18**	**82.35**	0.628	0.375	**0.681**	**0.333**	0.658	0.356	0.680	0.335
R1	76.30	75.80	**76.97**	**76.59**	72.28	71.67	76.45	76.06	0.594	0.458	**0.646**	**0.413**	0.611	0.453	0.642	0.416
R2	78.61	78.44	79.21	79.04	76.91	76.75	**79.51**	**79.35**	0.625	0.421	**0.680**	**0.371**	0.631	0.424	0.680	0.373
R3	71.59	71.37	**73.39**	**73.15**	68.84	68.66	71.15	70.86	0.525	0.455	0.574	0.422	0.529	0.471	**0.576**	**0.421**
R4	81.60	79.89	81.76	80.22	79.15	77.42	**82.28**	**80.70**	0.616	0.465	**0.680**	**0.404**	0.652	0.438	0.676	0.409
A1	75.63	75.49	**76.22**	**76.08**	72.73	72.57	76.16	76.01	0.557	0.423	**0.597**	**0.394**	0.538	0.451	0.590	0.400
A2	74.74	74.34	**76.24**	**75.93**	72.95	72.49	75.34	75.01	0.559	0.445	**0.607**	**0.410**	0.564	0.457	0.604	0.413
A3	73.99	73.59	**74.21**	**73.80**	71.45	71.00	74.13	73.74	0.532	0.476	**0.585**	**0.436**	0.514	0.51	0.582	0.437
A4	69.06	64.46	69.73	65.59	67.27	62.90	**70.33**	**66.03**	0.484	0.533	**0.554**	**0.479**	0.460	0.572	0.545	0.487
A5	68.24	67.71	**68.53**	**68.09**	66.60	65.87	68.23	67.70	0.440	0.530	0.491	0.497	0.442	0.555	**0.498**	**0.491**
F1	76.45	73.27	**76.60**	**73.68**	73.17	70.09	76.60	73.65	0.496	0.490	**0.539**	**0.462**	0.467	0.529	0.536	0.464
F2	69.66	62.62	71.15	64.36	69.96	64.30	**71.23**	**64.81**	0.500	0.560	**0.521**	**0.546**	0.482	0.595	0.519	0.546
F3	70.55	66.01	**71.15**	**66.90**	67.94	63.28	69.58	64.55	0.452	0.513	**0.497**	**0.484**	0.46	0.527	0.488	0.490
F4	72.57	69.52	73.01	70.17	72.64	70.50	**73.61**	**71.01**	0.503	0.482	0.500	0.487	0.453	0.538	**0.506**	**0.482**
I1	77.72	77.53	**78.62**	**78.44**	75.94	75.74	78.40	78.20	0.635	0.422	**0.669**	**0.392**	0.644	0.425	0.668	0.394
I2	76.76	76.59	76.76	76.59	76.83	76.61	**77.13**	**76.93**	0.598	0.439	0.617	0.423	0.605	0.439	**0.628**	**0.413**
I3	84.31	83.75	86.18	85.69	83.64	83.06	**86.25**	**85.75**	0.685	0.394	**0.735**	**0.344**	0.712	0.373	0.734	0.345
I4	80.26	79.93	82.20	81.90	80.86	80.61	**82.73**	**82.41**	0.658	0.410	0.705	0.364	0.692	0.38	**0.706**	**0.364**
C1	81.09	80.88	**82.13**	**81.86**	81.09	80.85	81.69	81.41	0.668	0.412	0.714	0.366	0.692	0.394	**0.717**	**0.362**
C2	78.17	77.65	**79.60**	**79.00**	78.93	78.43	78.93	78.35	0.640	0.445	0.701	0.383	0.699	0.389	**0.702**	**0.382**
C3	81.54	75.76	81.39	75.83	77.80	72.00	**81.84**	**76.28**	0.521	0.398	**0.574**	**0.366**	0.568	0.379	0.564	0.374

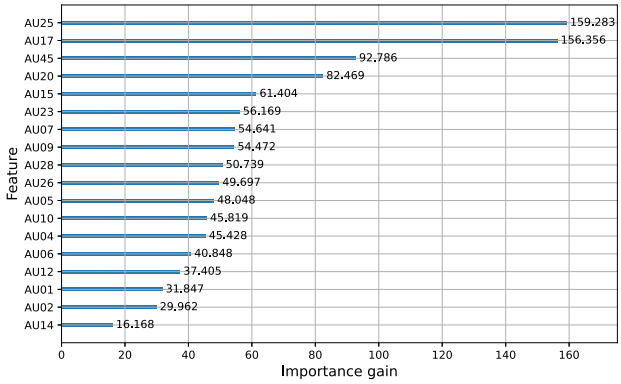
valuable insights into the specific aspects of spoken English proficiency that the model prioritizes during prediction. Finally, the inclusion of interviewer features within the model revealed a measurable impact on the prediction accuracy of the SEE score. This finding suggests that seemingly minor aspects, such as interviewer communication style or questioning technique, may play a significant role in interviewee performance.

5.3.1. Learning multiple speaking skill indices assessment

Table 3 presents the results of our multioutput prediction using unimodal features, whereas Table 4 shows the results achieved with multimodal features. In the multimodal approach, we incorporated explainable features, i.e., prosody, histogram of action units (HAU), and turn-taking features. Although potentially more effective for scoring, WavLM and BERT features were excluded because of their inherent lack



(a) Feature importance of prosody. The prosodic features consist of F0-related features (red), energy-related features (blue), and duration-related features (black).



(b) Feature importance of HAU. The five most important features are AU 25 (lip funneler), AU 17 (chin raiser), AU 45 (blink), AU 20 (lip stretcher), and AU 15 (lip corner depressor).

Fig. 3. Feature importance of prosody and HAU.

of interpretability. While these embedding layer vectors have previously demonstrated high levels of accuracy in various tasks, their black-box nature hinders their suitability for an explainable assessment system (see Tables A.1, A.2, B.1 and B.2).

From Table 3, we report several findings from the unimodal approach. Interestingly, both the classification and regression tasks exhibited minimal variations in the most important features for inferring SEE scores and their subcategories. WavLM emerged as the strongest feature, achieving the highest accuracy, correlation, and lowest error across most target labels. Prosody, turn-taking patterns, BERT outputs, and histograms of action units (HAUs) followed in terms of performance. Notably, the easily extractable turn-taking feature achieved comparable results for predicting the final SEE score. Prosodic features also rival WavLM's performance (no significant difference in SEE prediction) and even excelled in predicting specific subcategories of fluency and interaction. This outcome might be due to WavLM encompassing both acoustic and linguistic features, which are typically emphasized in automatic speech recognition tasks. Similarly, BERT features, while significant, were less dominant than WavLM features were. Since BERT primarily represents the grammatical correctness of sentence structures, it performs better in inferring accuracy. Finally, HAU was found to be the least effective feature for SEE score prediction, with an achieved accuracy of approximately 65%.

An analysis of the results in Table 4 revealed several findings concerning the efficacy of the multimodal approach. Our multimodal approach incorporated prosodic, turn-taking, and HAU features. WavLM was excluded because of its negligible impact on SEE prediction compared with prosodic features and its lack of interpretable descriptors, unlike the other features employed in our analysis. For the binary classification task aiming to predict the final SEE score, the combination of prosody, HAU, and turn-taking features yielded the most accurate results. However, in the regression task predicting the continuous SEE score value, excluding HAU resulted in marginally better performance, although this difference was not statistically significant ($p > 0.05$). In the multimodal approach, prosody and turn-taking features were generally shown to be sufficient for achieving optimal performance. Thus, the inclusion of only HAUs could significantly improve the prediction of the I4 subcategory (conversational cues), as evidenced by a p value lower than 0.05.

An analysis of both unimodal and multimodal results revealed some interesting findings. The multimodal approach that used prosody and turn-taking features achieved the most accurate final SEE score prediction. This was evident in both the classification (accuracy of approximately 83%) and regression tasks (ρ: 0.681, RMSE: 0.333). While the unimodal WavLM model exhibited slightly lower performance, its lack of interpretability limits its usefulness in the context of understanding the SEE assessment.

5.3.2. Feature and sequence importance

To gain a deeper understanding of feature contributions and identify potential redundancies or synergies among modalities, we conducted a unimodal feature importance analysis using the importance gain unit in a gradient boosting method. By examining features within their respective modalities, we can better appreciate their individual significance and potential interactions. Additionally, unimodal models often provide a simpler and more interpretable framework than multimodal models do, facilitating a clearer understanding of the underlying factors influencing spoken English proficiency.

Fig. 3 shows the feature importance according to the gain method in the regression task of estimating the SEE score using the LightGBM model. The gain method evaluates how much a feature influences the decision-making process in a tree. It assesses how well a feature separates data points, aiming for either lower overall error (classification) or higher purity (homogeneity) within tree nodes.

Fig. 3a illustrates the significance of prosody features in predicting SEE scores. Prosody encompasses 103 features categorized into three subgroups: F0-based (indices 1–30), energy-based (indices 31–78), and duration-based (indices 79–103). The figure reveals that each prosody feature contributed to the prediction model's efficacy. Notably, F0-based features, particularly those derived from the mean squared error of linear F0 estimation for voiced segments (indices 13–18), substantially improved the model's performance. Similarly, duration-based features, especially those associated with voiced segments and pauses, also yielded significant gains in the prediction process.

Fig. 3b highlights the importance of each HAU in estimating the SEE score. Action units (AUs) associated with lip movement emerged as the most significant features. Among these, AU 25 (lip part: depressor labii, relaxation of mentalis (AU 17), orbicularis oris) and AU 17 (chin raiser: mentalis) exhibited the greatest influence, with a gain exceeding 150 in feature importance. AU 45 (blink: relaxation of levator palpebrae and contraction of orbicularis oculi, pars palpebralis), which potentially indicates changes in cognition load while speaking (Brych et al., 2021), also significantly increased. The remaining most influential AUs were found to be AU 20 (lip stretcher: risorius) and AU 15 (lip corner depressor: depressor anguli oris (triangularis)), both of which are related to lip movement. Our findings align with nonverbal communication research, which indicates that eye and lip movements are associated with uncertainty (Givens, 2002). In particular, the combination of AU 15 and AU 17 corresponds to the verbal expression of hesitation phrases, such as "I do not know" (Ricci Bitti et al., 2014).

Table 5
Experimental results using additional interviewer features. Δ denotes the improvement of accuracy in comparison to the prediction results obtained by the interviewee features only.

Feature	Target					
	SEE (Δ)	$\text{Range}_{\text{Avg.}}$ (Δ)	$\text{Accuracy}_{\text{Avg.}}$ (Δ)	$\text{Fluency}_{\text{Avg.}}$ (Δ)	$\text{Interaction}_{\text{Avg.}}$ (Δ)	$\text{Coherence}_{\text{Avg.}}$ (Δ)
WavLM	81.990 (0.006)	**76.773 (-0.971)**	73.126 (1.020)	**72.010 (-5.849)**	**80.048 (-0.404)**	81.490 (5.011)
Prosody	**81.760 (-0.290)**	**75.278 (-1.063)**	**71.034 (-0.446)**	**69.788 (-8.238)**	**80.158 (-0.767)**	80.043 (4.293)
BERT	**77.580 (-1.940)**	**68.535 (-3.983)**	**69.418 (-3.016)**	**72.383 (-0.807)**	**71.578 (-4.656)**	74.397 (5.867)
HAU	68.760 (3.665)	64.108 (1.787)	**62.978 (-1.555)**	65.850 (2.190)	66.633 (0.469)	68.513 (7.945)
Turn	82.350 (0.514)	75.933 (2.107)	71.302 (2.810)	**70.028 (-6.301)**	**78.983 (-0.579)**	80.390 (6.022)
Prosody-HAU	**81.760 (-0.070)**	**75.338 (-1.688)**	**71.110 (-1.222)**	**70.010 (-2.298)**	80.030 (0.267)	80.370 (0.103)
Prosody-Turn	83.030 (0.370)	**75.485 (-2.348)**	**72.166 (-0.820)**	**70.160 (-2.818)**	**80.645 (-0.295)**	**80.790 (-0.250)**
HAU-Turn	**81.240 (-0.220)**	74.735 (0.440)	70.756 (0.556)	71.578 (0.650)	79.825 (0.508)	79.697 (0.423)
Prosody-HAU-Turn	83.550 (0.370)	**76.160 (-1.188)**	**72.138 (-0.700)**	**70.125 (-2.630)**	**80.828 (-0.300)**	81.040 (0.220)

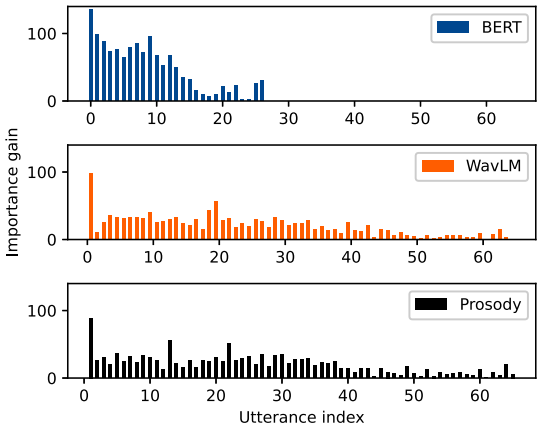


Fig. 4. Sequence importance for estimating SEE scores via BERT, WavLM, and prosodic features.

Fig. 4 explores the importance of the utterance sequence for spoken English proficiency estimation, analyzing three features, namely, BERT, WavLM, and prosody. The results revealed a clear trend, namely, the first utterance held the most significance in predicting spoken proficiency. This importance gradually diminished as the sequence progresses, although a slight peak was observed in the acoustic features (WavLM and prosody) toward the second last of the utterance. These findings align with psychology research suggesting that first impressions play a substantial role in hiring decisions during job interviews (Bernieri, 2000).

5.3.3. The role of external factors (interviewer features and interview setting)

Interviewer behavior can significantly influence an interviewee's performance. Studies have shown that first impressions formed by interviewers can be lasting, potentially impacting how they evaluate a candidate throughout the interview process (Bernieri, 2000). Additionally, interviewer communication style and questioning techniques can influence interviewee anxiety and self-disclosure levels (Jourard & Jaffe, 1970; Little et al., 1976). For example, interviewers who use open-ended questions and active listening may encourage interviewees to speak more freely, potentially leading to a more accurate assessment of their skills and experiences. Conversely, interviewers who display impatience or interrupt frequently might create a stressful environment that hinders their performance.

In addition to assessing interviewee proficiency, this study explored the influence of interviewer behavior. The feature representation of interviewer behavior was found to be equivalent to the features extracted for an interviewee, except for visual cues. Owing to the camera focusing solely on the interviewee, the visual features of the interviewer were not captured. Therefore, we analyzed interviewee behavior through visual cues (as a proxy for interviewer features) while the interviewer spoke segments. The interviewees' reactions obtained through visual cues can provide valuable insights into the dynamics of the interview while the interviewer is speaking. We employed a classification task to predict SEE scores and the average accuracy for aspects such as range, accuracy, fluency, interaction, and coherence.

Table 6
SEE prediction in macro F1 score (%) based on interview setting.

	Gender		
Type	Female	Male	All
Remote	89.08	79.57	84.01
In-Person	75.26	83.64	80.09
All	79.59	83.22	81.73

The impact of incorporating interviewer features on SEE score classification accuracy is presented in Table 5. HAU features from interviewer speech led to a significant 3.665% increase in SEE score classification accuracy. Interestingly, the greatest improvement was observed in the coherence aspect. This finding aligns with the observation that coherence has the strongest correlation ($\rho > 0.9$) with SEE scores, suggesting that interviewer features significantly influence the coherence of the interviewees' responses.

We extended our analysis beyond interviewer features to explore interview setting prediction using a 5-fold cross-validation (Table 6). Without considering gender, the results indicated greater accuracy for models predicting SEE scores in remote sessions than in in-person settings. However, different outputs were obtained when considering gender. In remote sessions, the model predicted SEE scores for females to be nearly 10% higher than those for males. Conversely, the in-person setting exhibited the opposite trend, with male SEE score predictions being 8% higher than female predictions. These findings warrant further investigation into potential gender bias within the model's predictions or differential difficulty in predicting SEE scores on the basis of the interplay between the interview setting and gender.

6. Discussion

This research contributes to the evolving field of educational technology and language assessment by enhancing the interpretability of spoken English evaluation (SEE) scores in practice.

6.1. Technology-enhanced language learning

The use of machine learning models to increase the interpretability of SEE scores aligns with the growing field of technology-enhanced language learning. Moving beyond traditional metrics, this research focuses on a broader range of speaking skills, including fluency, accuracy, coherence, and discourse management. By analyzing data from both remote

Table 7
Illustrative Example: Prompt and AI-generated Feedback for Comprehensive Spoken English Evaluation. Feedback generated using the Gemini 1.5 flash model.

Prompt	Feedback
Generate personalized feedback on a speaker's English **fluency** based on the following scores: **Accent**: 4 **Tempo and Pausing**: 3 **Intonation and Fluidity**: 2 Fluency is defined as the ability to speak smoothly and naturally, encompassing factors such as the following: Smoothness: Minimal hesitations, false starts, and self-corrections. Tempo: A natural and even flow of speech with appropriate pauses and hesitation. Rate: Speaking at a comfortable and appropriate speed. The feedback should be as follows: Constructive and encouraging: Focus on areas of strength and areas for improvement. Personalized: Tailor the feedback to the specific scores provided. Actionable: Provide specific suggestions for how the speaker can improve their fluency.	Your pronunciation is generally clear, but there are some areas where your accent could be more natural. You tend to pause slightly more frequently than native speakers, which can sometimes disrupt the flow of your speech. Working on speaking more smoothly and naturally, and paying attention to intonation patterns, will significantly improve your overall fluency.

and in-person settings, we aim to develop a more robust and generalizable model. This research demonstrates the potential of technology to significantly support and improve spoken language assessment practices, providing valuable information for both educators and learners.

6.2. Learner-centered approach

The proposed development of an automatic system for predicting SEE scores, leveraging model insights and the impact of external factors, directly supports a learner-centered approach to language education. This system has the potential to provide personalized feedback to both interviewees and interviewers, enabling them to identify areas for improvement and enhancing their language learning and teaching practices.

To facilitate this, we propose leveraging a large language model to generate constructive feedback for interviewees. This model receives detailed information about the predicted SEE scores as input. Table 7 presents an example of a prompt constructed on the basis of predicted SEE scores related to fluency and the corresponding constructive and actionable feedback provided to the interviewees. The feedback comments are generated based on the inference results of the proposed interpretable model. Our model successfully minimized the generation of false or misleading information, improving the reliability and accuracy of the LLM-generated feedback.

The findings of this study have broad implications, offering diverse applications for stakeholders across various fields, as follows:

- **Researchers:** The proposed multi-output learning approach provides a framework for developing more transparent and interpretable models in language assessment. This allows researchers to analyze the factors influencing spoken English proficiency, facilitating the development of more valuable assessment tools and pedagogical strategies.
- **Teachers:** By understanding the specific features contributing to SEE scores (e.g., fluency, range, coherence), teachers can tailor their instruction to address individual student needs. The identified impact of interviewer behavior and setting also highlights the importance of creating supportive learning environments.
- **Students:** The development of a feedback system based on the interpretable model and external factor analysis will provide students with actionable insights into their spoken English performance. This personalized feedback can help them identify areas for improvement and develop more effective learning strategies.
- **System Designers:** The findings offer valuable guidance for designing automated spoken English assessment systems that prioritize both accuracy and interpretability. Incorporating multimodal cues and considering the impact of external factors can lead to more robust and reliable assessment tools.

7. Conclusion and future work

This research introduces a novel approach to enhancing the interpretability of the SEE process by leveraging multioutput learning models (RQ1). Our method improves interpretability using multiple indices related to SEE assessment while maintaining accuracy comparable to that of state-of-the-art black-box models. We also successfully identify features that strongly correlate with the overall SEE score and its constituent aspects, such as fluency, range, and coherence (RQ2). Intriguingly, further analysis reveals a quantifiable impact of external factors, namely, interviewer behavior and the interview setting, on interviewees' spoken proficiency (RQ3). Our findings suggest that incorporating interviewer characteristics into future models could be beneficial, particularly for inferring coherence.

In terms of practical implications, our proposed system can be integrated into existing language learning platforms to provide personalized feedback to learners. This feedback highlights areas of strength and weakness in their spoken English. Such information can subsequently empower learners to track their progress and tailor their learning strategies accordingly.

Although this work demonstrates the potential of the proposed approach, we acknowledged some limitations of our research. The proposed approach, while innovative, may present certain challenges in terms of inheritability. The complexity of the multimodal data processing pipeline, involving audio, video, and text streams, may require significant computational resources and expertise. Data cleaning, which is a crucial step, can be particularly challenging, requiring careful handling of issues such as noise in audio recordings, inconsistencies in video quality, and inaccuracies in transcriptions.

Furthermore, while this study primarily evaluates system performance, future research will investigate its real-world impact by examining its effectiveness within actual educational settings, specifically focusing on its influence on student learning outcomes.

Addressing potential biases within the dataset, such as proficiency bias and interview setting bias, is crucial to ensuring the fairness and generalizability of the model. Future research will address bias mitiga-

tion strategies and perform a comprehensive analysis of demographic aspects and interview settings to improve the robustness of the model and ensure equitable results.

Statements on open data

This study received approval from Vericant's compliance committee. The participants provided their written informed consent prior to participation, and their privacy rights were strictly observed. Pending final approval, anonymized data supporting the findings of this study will be made available upon reasonable request to the corresponding author. To ensure responsible data sharing and maintain ethical considerations, further consultation may be required on data access and potential collaborations. Data availability will be subject to the following conditions:

- **Confidentiality and privacy:** All identifying information will be removed from the data before release to ensure the anonymity of the participants.
- **Data use restrictions:** Data may only be used for research purposes consistent with the original study objectives.
- **Data citation:** Researchers using the data must appropriately cite the original publication.
- **Data-sharing agreement:** A data-sharing agreement may be required to outline the terms of data access and use.

CRediT authorship contribution statement

Candy Olivia Mawalim: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation. **Chee Wee Leong:** Writing – review & editing, Validation, Data curation. **Guy Sivan:** Writing – review & editing, Data curation. **Hung-Hsuan Huang:** Writing – review & editing, Validation. **Shogo Okada:** Writing – review & editing, Validation, Supervision, Methodology.

Declaration of competing interest

There are no potential conflicts of interest associated with this study.

Acknowledgement

This work was partially supported by JSPS KAKENHI (22H00536, 23H03506).

Appendix A. Multi-comparison pairwise tests on classification task

Table A.1
Bonferroni-adjusted pairwise tests on unimodal features to determine significant differences between groups.

Target	SEE	R1	R2	R3	R4	A1	A2	A3	A4	A5	F1	F2	F3	F4	I1	I2	I3	I4	C1	C2	C3
Pair	**BERT**																				
WavLM	+	***	**	***	***	+	+	***	+	+	+	+	***	***	***	***	+	+	***	***	***
Prosody	+	***	+	***	+	+	+	+	+	+	+	+	***	***	***	***	+	***	***	***	***
HAU	***	***	***	***	***	+	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
Turn	+	+	+	+	+	**	+	**	***	**	+	+	+	***	***	***	+	+	***	***	**
Pair	**HAU**																				
WavLM	***	***	***	***	***	+	***	***	***	+	***	***	***	***	***	***	***	***	***	***	***
Prosody	***	***	***	***	***	+	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
Turn	***	***	***	***	***	***	***	***	+	+	***	***	***	***	***	***	***	***	***	***	***
Pair	**Prosody**																				
WavLM	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	**	+	+	+
Turn	+	**	+	+	+	+	+	***	+	**	+	+	**	+	+	+	+	**	+	+	+
Pair	**Turn**																				
WavLM	+	**	**	**	***	***	+	***	**	+	+	+	**	+	+	+	+	+	+	+	+

Note: '***': $p < 0.01$, '**': $p < 0.05$, '+': $p \geq 0.05$.

Table A.2
Bonferroni-adjusted pairwise tests on multimodal features to determine significant differences between groups.

Target	SEE	R1	R2	R3	R4	A1	A2	A3	A4	A5	F1	F2	F3	F4	I1	I2	I3	I4	C1	C2	C3
Pair	**HAU-Turn**																				
Prosody-HAU	+	+	**	+	+	+	+	**	+	+	+	+	+	+	+	+	+	**	**	+	+
Prosody-Turn	+	+	**	**	+	+	+	**	+	+	+	+	+	+	+	+	+	**	***	+	**
Prosody-HAU-Turn	+	**	**	+	+	+	+	**	+	+	+	+	+	+	+	+	+	**	**	+	+
Pair	**Prosody-HAU**																				
Prosody-Turn	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Prosody-HAU-Turn	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Pair	**Prosody-HAU-Turn**																				
Prosody-Turn	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Note: '***': $p < 0.01$, '**': $p < 0.05$, '+': $p \geq 0.05$.

Appendix B. Multi-comparison pairwise tests on regression task

Table B.1
Bonferroni-adjusted pairwise tests on unimodal features to determine significant differences between groups.

Target	SEE	R1	R2	R3	R4	A1	A2	A3	A4	A5	F1	F2	F3	F4	I1	I2	I3	I4	C1	C2	C3
Pair	**BERT**																				
WavLM	**	+	***	***	***	+	+	+	+	+	+	***	+	***	***	***	+	**	***	***	***
Prosody	+	+	+	+	+	***	**	+	+	+	+	***	+	***	+	***	+	+	+	+	+
HAU	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
Turn	+	+	**	+	+	***	***	***	***	**	+	**	+	***	***	***	+	**	+	+	+
Pair	**HAU**																				
WavLM	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
Prosody	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
Turn	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
Pair	**Prosody**																				
WavLM	**	+	+	***	**	***	**	+	+	**	+	+	+	+	+	+	+	+	**	+	+
Turn	+	+	***	+	+	+	+	**	***	+	+	+	+	+	+	+	+	+	+	+	+
Pair	**Turn**																				
WavLM	+	+	***	***	***	***	***	***	***	***	***	+	+	+	+	+	+	+	**	+	+

Note: '***': $p < 0.01$, '**': $p < 0.05$, '+': $p \geq 0.05$.

Table B.2
Bonferroni-adjusted pairwise tests on multimodal features to determine significant differences between groups.

Target	SEE	R1	R2	R3	R4	A1	A2	A3	A4	A5	F1	F2	F3	F4	I1	I2	I3	I4	C1	C2	C3
Pair	**HAU-Turn**																				
Prosody-HAU	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	***	+	+	+	+
Prosody-HAU-Turn	+	+	+	+	**	***	**	**	+	+	+	+	+	+	+	+	+	+	+	+	+
Prosody-Turn	+	+	+	+	**	***	+	**	+	+	+	+	+	+	+	+	+	+	+	+	+
Pair	**Prosody-HAU**																				
Prosody-HAU-Turn	**	***	+	+	+	***	+	+	+	+	+	**	+	**	**	**	***	+	**	**	+
Prosody-Turn	***	***	+	+	+	***	+	+	+	+	+	**	+	**	**	**	***	**	**	**	+
Pair	**Prosody-HAU-Turn**																				
Prosody-Turn	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Note: '***': $p < 0.01$, '**': $p < 0.05$, '+': $p \geq 0.05$.

References

Amos, B., Ludwiczuk, B., & Satyanarayanan, M. (2016). OpenFace: A general-purpose face recognition library with mobile applications. Technical Report. CMU-CS-16-118, CMU School of Computer Science.

Bernieri, F. J. (2000). The importance of first impressions in a job interview. In *The annual meeting of the midwestern psychological association*.

Bernstein, J., Cheng, J., & Suzuki, M. (2010a). Fluency and structural complexity as predictors of L2 oral proficiency. In *Proceedings of interspeech, ISCA* (pp. 1241–1244).

Bernstein, J., Moere, A. V., & Cheng, J. (2010b). Validating automated speaking tests. *Language Testing*, *27*, 355–377. https://doi.org/10.1177/0265532210364404.

Brych, M., Murali, S., & Händel, B. (2021). How the motor aspect of speaking influences the blink rate. *PLoS ONE*, *16*, Article e0258322. https://doi.org/10.1371/journal.pone.0258322.

Chen, L., Feng, G., Joe, J., Leong, C. W., Kitchen, C., & Lee, C. M. (2014). Towards automated assessment of public speaking skills using multimodal cues. In *Proceedings of ICMI* (pp. 200–203). New York, NY, USA: ACM.

Chen, L., Zhao, R., Leong, C. W., Lehman, B., Feng, G., & Hoque, M. E. (2017). Automated video interview judgment on a large-sized corpus collected online. In *Proceedings of ACII* (pp. 504–509). USA: IEEE.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., & Wei, F. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, *16*, 1505–1518. https://doi.org/10.1109/JSTSP.2022.3188113.

Cheng, J., Zhao D'Antilio, Y., Chen, X., & Bernstein, J. (2014). Automatic assessment of the speech of young English learners. In J. Tetreault, J. Burstein, & C. Leacock (Eds.), *Proceedings of the ninth workshop on innovative use of NLP for building educational applications* (pp. 12–21). Baltimore, Maryland: ACL. https://aclanthology.org/W14-1802.

Council of Europe (2018). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Davis, L., & Norris, J. M. (2024). *Challenges and innovations in speaking assessment: Innovations in language learning and assessment at ETS, Vol. 7* (1 ed.). Routledge.

Dehak, N., Dumouchel, P., & Kenny, P. (2007). Modeling prosodic features with joint factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*, 2095–2103. https://doi.org/10.1109/TASL.2007.902758.

Diamantidis, N., Karlis, D., & Giakoumakis, E. (2000). Unsupervised stratification of cross-validation for accuracy estimation. *Artificial Intelligence*, *116*, 1–16. https://doi.org/10.1016/S0004-3702(99)00094-6.

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication*, *51*, 832–844.

Firth, A., & Wagner, E. (2017). Speaking proficiency of young language students: A discourse-analytic study. *Language and Education*, *31*, 719–738.

Givens, D. B. (2002). *The nonverbal dictionary of gestures, signs and body language cues*. Spokane, Washington, DC: Center for Nonverbal Studies Press.

Gretter, R., Allgaier, K., Tchistiakova, S., & Falavigna, D. (2019). Automatic assessment of spoken language proficiency of non-native children. In *ICASSP 2019 - 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 7435–7439).

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*, 77–89.

Jourard, S. M., & Jaffe, P. E. (1970). Influence of an interviewer's disclosure on the self-disclosing behavior of interviewees. *Journal of Counseling Psychology*, *17*, 252–257. https://doi.org/10.1037/h0029197.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 3149–3157). Hook, NY, USA: Curran Associates Inc., Red.

Lee, S., Shakir, A., Koenig, D., & Lipp, J. (2024). Open source strikes bread - new fluffy embeddings model. https://www.mixedbread.ai/blog/mxbai-embed-large-v1.

Li, X., & Li, J. (2023). Angle-optimized text embeddings. arXiv preprint. arXiv:2309.12871.

Little, B. R., Arlett, C., & Best, J. A. (1976). The influence of interviewer self-disclosure and verbal reinforcement on personality tests. *Journal of Clinical Psychology*, *32*, 770–775.

McKnight, S. W., Civelekoglu, A., Gales, M., Bannò, S., Liusie, A., & Knill, K. (2023). Automatic assessment of conversational speaking tests (pp. 99–103).

Nguyen, L. S., Frauendorfer, D., Mast, M. S., & Gatica-Perez, D. (2014). Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Transactions on Multimedia*, *16*, 1018–1031. https://doi.org/10.1109/TMM.2014.2307169.

Ohba, T., Mawalim, C. O., Katada, S., Kuroki, H., & Okada, S. (2022). Multimodal analysis for communication skill and self-efficacy level estimation in job interview scenario. In

21th international conference on mobile and ubiquitous multimedia (MUM 2022) (p. 11). New York, NY, USA, Lisbon, Portugal: ACM.

Okada, S., Ohtake, Y., Nakano, Y. I., Hayashi, Y., Huang, H. H., Takase, Y., & Nitta, K. (2016). Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets. In *Proceedings of ACM ICMI* (pp. 169–176). New York, NY, USA: ACM.

Oliveri, M. E., & Tannenbaum, R. J. (2019). *Are we teaching and assessing the English skills needed to succeed in the global workplace?* John Wiley & Sons, Ltd. (pp. 343–354). chapter 19.

Ricci Bitti, P. E., Bonfiglioli, L., Melani, P., Caterina, R., & Garotti, P. (2014). Expression and communication of doubt/uncertainty through facial expression. *Ricerche di Pedagogia e Didattica–Journal of Theories and Research in Education*, *9*. Special Issue. Communicating certainty and uncertainty: Multidisciplinary perspectives on epistemicity in everyday life. Edited by Andrzej Zuczkowski and Letizia Caronia.

Saeki, M., Matsuyama, Y., Kobashikawa, S., Ogawa, T., & Kobayashi, T. (2021). Analysis of multimodal features for speaking proficiency scoring in an interview dialogue. In *2021 IEEE spoken language technology workshop (SLT)* (pp. 629–635). Shenzhen, China: IEEE.

Sanchez-Cortes, D., Aran, O., Jayagopi, D., Mast, M., & Gatica-Perez, D. (2013). Emergent leaders through looking and speaking: From audio-visual data to multimodal recognition. *Journal on Multimodal User Interfaces*, *7*, 39–53. https://doi.org/10.1007/s12193-012-0101-0.

Townshend, B., Bernstein, J., Todic, O., & Warren, E. (1998). Estimation of spoken language proficiency. In *Proc. ETRW on speech technology in language learning (STiLL)* (pp. 179–182). Marholmen, Sweden: ISCA.

Vásquez-Correa, J., Orozco-Arroyave, J., Bocklet, T., & Nöth, E. (2018). Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease. *Journal of Communication Disorders*, *76*, 21–36.

Wang, Y., Gales, M., Knill, K., Kyriakopoulos, K., Malinin, A., van Dalen, R., & Rashid, M. (2018). Towards automatic assessment of spontaneous spoken English. *Speech Communication*, *104*, 47–56. https://doi.org/10.1016/j.specom.2018.09.002.

Wortwein, T., Morency, L. P., & Scherer, S. (2015). Automatic assessment and analysis of public speaking anxiety: A virtual audience case study. In *Proceeding of ACII* (pp. 187–193). Xi'an, China: IEEE Computer Society.

Xu, D., Shi, Y., Tsang, I., Ong, Y., Gong, C., & Shen, X. (2019). Survey on multi-output learning. *IEEE Transactions on Neural Networks and Learning Systems*, *31*, 2409–2429. https://doi.org/10.1109/TNNLS.2019.2945133.

Yoon, S. Y., Bhat, S., & Zechner, K. (2012). Vocabulary profile as a measure of vocabulary sophistication. In *Proceedings of the 7th workshop on innovative use of NLP for building educational applications* (pp. 180–189). Montreal, Canada: NAACL-HLT, ACL.

Yoon, S. Y., & Zechner, K. (2017). Combining human and automated scores for the improved assessment of non-native speech. *Speech Communication*, *93*, 43–52. https://doi.org/10.1016/j.specom.2017.08.001.