Research Methods in Applied Linguistics 4 (2025) 100177

Contents lists available at ScienceDirect

Research Methods in Applied Linguistics

journal homepage: www.elsevier.com/locate/rmal

Gauging the validity of machine learning-based temporal feature annotation to measure fluency in speech automatically

Ryuki Matsuura [a,b,*], Shungo Suzuki [a], Kotaro Takizawa [a], Mao Saeki [a], Yoichi Matsuyama [a]

[a] Waseda University, 1-104, Totsukamachi, Shinjuku-ku, Tokyo 1698050, Japan
[b] Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, United States

ARTICLE INFO

Keywords:
Speech annotation
Utterance fluency
Second language speech
Natural language processing
Machine learning

ABSTRACT

Machine learning (ML) techniques allow for automatically annotating various temporal speech features, particularly by the cascade connection of ML-based modules. Although such systems are expected to enhance scalability of second language (L2) speech research, their annotation accuracy is potentially moderated by speaking tasks and proficiency levels due to the mismatch between training and real-world data. Accordingly, we developed and validated an ML-based temporal feature annotation system on L2 English datasets split by speaking tasks (monologic vs. dialogic tasks) and proficiency levels, operationalized as overall fluency levels (low, mid vs. high). We compared the annotations by experts and the system in terms of the agreement between manual and automatic annotations, correlations between manual and automatic measures, and the predictive power for listener-based fluency judgments. Results showed a substantial degree of agreement in the annotations for monologic tasks and a general tendency of strong correlations between manual and automatic measures regardless of tasks and overall fluency levels. Furthermore, automatic measures yielded substantial predictive power of fluency scores in monologic tasks. These findings suggest the substantial applicability of ML-based annotation systems to monologic tasks possibly without biases by holistic levels of fluency.

Introduction

In second language (L2) speech research, objective measurement of oral fluency has been extensively used to evaluate treatment effects (Suzuki & Hanzawa, 2022) and predict proficiency-related scores (Tavakoli et al., 2023). In the context of fluency research, objective fluency measures are conceptualized as *utterance fluency* (UF) and are assumed to encompass three distinctive dimensions: *speed fluency* (SF), *breakdown fluency* (BDF), and *repair fluency* (RF) (Tavakoli & Skehan, 2005). SF refers to the pace and density of information delivery. This dimension is closely related to cognitive processing speed in speech production (S. Suzuki & Kormos, 2023) and has a strong relationship with L2 oral proficiency (Tavakoli et al., 2023). BDF is the dimension of pauses in speech and reflects disruptions in speech production processing. Pauses within a clause tend to indicate a breakdown in linguistic formulation processes due to some lexico-grammatical difficulty and thus occur frequently in lower-proficiency learners' speech (de Jong, 2016). RF is concerned with disfluency phenomena (e.g., repetitions, self-corrections, and false starts) in L2 speech and is associated with speaker's

* Corresponding author at: Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, United States.
E-mail addresses: rmatsuur@andrew.cmu.edu (R. Matsuura), ssuzuki@pcl.cs.waseda.ac.jp (S. Suzuki), kzmw0628@fuji.waseda.jp (K. Takizawa), saeki@pcl.cs.waseda.ac.jp (M. Saeki), matsuyama@pcl.cs.waseda.ac.jp (Y. Matsuyama).

https://doi.org/10.1016/j.rmal.2024.100177
Received 2 September 2024; Received in revised form 18 December 2024; Accepted 21 December 2024
Available online 31 December 2024

self-monitoring processes as well as disruptions in speech processing (Kormos, 2006).

Despite the usefulness of objective fluency measures in L2 speech research, the annotation of temporal speech features (e.g., pauses, disfluency words) is highly labor-intensive, which limits the scalability of fluency studies. To address the labor-intensiveness, temporal feature annotation has been automatized. de Jong and Wempe (2009) and de Jong et al. (2021) created a script for Praat (Boersma & Weenink, 2024) to detect syllables and silent and filled pauses and calculate UF measures by acoustic processing techniques. Moreover, incorporating machine learning (ML) techniques, scholars have developed automatic annotation systems for a more comprehensive set of temporal features, including disfluency words and pause locations (Chen & Yoon, 2011; Coulange et al., 2024; Matsuura et al., 2022). One notable advantage of ML techniques over acoustic methods is, despite the necessity of L2 datasets for fine-tuning, their applicability to various languages (e.g., Al-Ghezi et al., 2023). Furthermore, a recent meta-analysis suggested that automatically calculated UF measures are valid alternatives to manual ones in terms of the predictive power for listener-based fluency ratings (i.e., perceived fluency [PF]) (Suzuki et al., 2021). Therefore, ML-based annotation methods are promising to further accelerate fluency research.

Despite the growing demands of ML-based temporal feature annotation systems, annotation accuracy can vary according to speech elicitation tasks and speakers' proficiency levels due to the gaps between training and real-world data (Knill et al., 2018, 2019; Skidmore & Moore, 2023). However, a system validation has been typically limited to a specific speech task, and an annotation system has been tested on a dataset that is the mixture of utterances produced by learners at different proficiency levels (e.g., Matsuura et al., 2022). These conventional approaches, albeit evaluating system's overall accuracy, may fail to test the robustness of automatic annotation in another research context with different speaking tasks and target proficiency levels. Therefore, the current study further validates an ML-based annotation system in light of its robustness to speech tasks and proficiency levels, comparing manual and automatic annotation, UF measures, and fluency rating prediction in L2 English.

Literature review

Theoretical background of L2 oral fluency

To better understand L2 oral fluency, previous studies have extensively examined the link between temporal speech features (i.e., UF) and listener-based fluency judgments (i.e., PF) (Segalowitz, 2010). Given a consensus that oral fluency has three sub-dimensions (i.e., SF, BDF, and RF; Tavakoli & Skehan, 2005), UF measures should cover these three characteristics for the sake of construct validity. SF is commonly operationalized by an articulation rate. BDF measures are classified into two types: frequency-based and duration-based pause measures. RF is captured by the frequency of repetitions, self-corrections, and false starts. Scholars have also employed composite measures that reflect multiple dimensions of UF (e.g., speech rate). A recent meta-analysis demonstrated that the strength of the relationship between UF measures and PF ratings can vary across those sub-dimensions of fluency (S. Suzuki et al., 2021). Speed and pause frequency measures were strongly associated with PF ratings ($r = .62, .59$, respectively), while there was a moderate link between the ratings and pause duration ($r = .46$). Notably, the relationship between PF ratings and pause measures was moderated by pause locations: mid-clause pauses (MCPs) strongly contributed to PF rating prediction ($r = .72$); whereas end-clause pauses (ECPs) contributed moderately ($r = .48$). RF measures show a weak but significant association with PF ratings when multiple disfluency features are counted into one repair measure ($r = .20$). The strongest association was found between PF ratings and composite measures ($r = |.72 - .76|$). These findings suggest that a whole range of SF, BDF, RF, and composite measures should be included to objectively capture L2 oral fluency.

Automatic annotation of temporal features

To calculate UF measures automatically, the development of reliable temporal feature annotation systems is a prerequisite. Traditionally, automatic annotation has been achieved by acoustic processing techniques. de Jong and Wempe (2009) proposed to acoustically detect syllable nuclei in speech in relation to intensity peaks based on the certain threshold and voicedness obtained from the pitch contour. Combining the silent interval detection function of Praat (Boersma & Weenink, 2024), their Praat script allows for automatically calculating speech rate, which had a strong correlation between the manual and automatic measures in their L1 and L2 combined Dutch spoken corpus ($r = .71$). More recently, scholars have also attempted to detect filled pauses based on formants and pitch of speech (e.g., de Jong et al., 2021; Rose, 2020). de Jong et al. (2021) evaluated their filled pause detection method and found that 83 % of filled pauses were correctly detected in their L2 English spoken corpus. The accuracy score achieved 84 % even in another unseen L2 English corpus, that was not used for training, suggesting the substantial accuracy and stability of their automatic filled pause detection method.

Despite the promising accuracy of detecting syllables and filled pauses, the lack of written transcriptions may hinder the validity of automatic fluency measures. The acoustic processing approaches not only fail to capture pause location information to calculate BDF measures but also cannot compute RF measures. Because the information regarding pause location and RF measures contributes to L2 oral fluency ratings (Suzuki et al., 2021), the automatic methods should also detect clause boundaries and disfluency words integrating written transcriptions. To extract such information from written transcriptions, ML techniques, including natural language processing (NLP), have been incorporated into temporal feature annotation systems. Chen and Yoon (2011) proposed the NLP-based method to detect clause boundaries and the onsets of disfluency word sequences by predicting whether these events occurred before a given word. When they adopted the system to the written transcriptions based on automatic speech recognition (ASR), the detection accuracy in L2 English monologue data achieved the acceptable level for detecting clause boundaries and the onsets of disfluency words ($F1 = .690, .304$, respectively). Although this detection method allows for counting disfluency phenomena and classifying pause locations, it does

not employ the pruning function (i.e., removal of disfluency words). In other words, their system does not standardize the number of syllables by excluding disfluency words (see Suzuki & Révész, 2023). To realize automatic disfluency word pruning, Matsuura et al. (2022) evaluated their disfluency detection method, as well as clause boundary annotations. They fine-tuned a large language model, BERT (Devlin et al., 2019), to detect and remove disfluency words for each transcription (cf. Lu et al., 2022). For each pruned transcription, a dependency parser identifies clause boundaries and subsequently classifies pause locations. They found a substantial agreement of disfluency detection and pause location classification between human and the system in terms of Cohen's kappa (κ = .674, .613, respectively). These studies indicate the potential of the ML-based temporal feature annotation as a valid alternative to manual work.

Challenges in validating temporal feature annotation systems

Temporal feature annotation systems have been validated primarily from two perspectives: the reliability of automatic annotation and the predictive power of automatic measures for proficiency-related scores. The annotation reliability has been measured as the agreement between system-driven annotations and the corresponding manual annotations in terms of Cohen's kappa and other metrics reflecting false positives and negatives (e.g., precision, recall, and F1) (Chen & Yoon, 2011; de Jong et al., 2021; Matsuura et al., 2022). The annotation reliability can also be judged by correlation coefficients between manual and automatic measures (e.g., Chen & Yoon, 2012; de Jong et al., 2021). Moreover, automatic annotation systems have been evaluated in terms of the automatic measures' predictive power on L2 speaking ability to test their adaptability to automated language assessment (de Jong et al., 2021). Employing ML methods, previous studies have constructed automated scoring systems and evaluated prediction accuracy to relevant proficiency-based measures using regression-based indices (e.g., R^2; Chen et al., 2018; de Jong et al., 2021). For example, despite the fact that each study adopted different outcome variables, the prediction accuracy of linear regression models achieved $R^2 = .320$ for L2 oral fluency (de Jong et al., 2021) and $R^2 = .584$ for an ETS's TOEFL iBT Speaking section score (Chen et al., 2018). Following these evaluation methods, we thus decided to validate our ML-based temporal annotation system in terms of human-system annotation agreement, manual and automatic measure correlation, and predictive power for L2 oral fluency scores.

Although the validation methods for temporal feature annotation systems have been established, their robustness to various communicative contexts and populations has been underresearched. Since most annotation systems have been developed in the context of automated assessment research, scholars have been interested in the validity of their systems exclusively in their own contexts and thus have evaluated their systems using an entire dataset elicited from a specific assessment task (e.g., Chen & Yoon, 2012; Matsuura et al., 2022). This evaluation approach, albeit revealing the overall accuracy of an annotation system in a specific L2 speech assessment, has had limitations in evaluating its robustness in different assessment contexts. In addition, given the growing demand for annotation systems in L2 speech research with an experimental design, where multiple formats of speaking tasks should be used for a pretest, treatment, and a posttest to minimize practice effects (e.g., task-repetition; Suzuki & Hanzawa, 2022), the robustness of annotation systems to speaking tasks is an important agenda of validation. Similarly, L2 speech research typically examines target variables across proficiency levels (e.g., Tavakoli et al., 2020, 2023), meaning that the annotation system should be robust to proficiency levels so that the computed measures are independent of systematic biases related to proficiency levels. In line with de Jong et al. (2021) claim that it is essential to test systems considering potential factors, the current study, therefore, focuses on speaking tasks and proficiency levels, both of which can potentially affect annotation accuracy.

Speaking tasks. Regarding the robustness to L2 speaking tasks, to the best of our knowledge, previous studies have examined only disfluency detection accuracy. Skidmore and Moore (2023) developed a BERT-based disfluency detection system fine-tuned with an L2 English corpus collected from three tasks, open dialogue, a role-play, and a picture narrative. Although their system yielded notable accuracy in a test dataset split from the fine-tuning dataset ($F1$ = .916, .922, and .935 for open-dialogue, role-play, and picture narrative, respectively), the accuracy decreased approximately 15% in an unseen test dataset elicited from different tasks ($F1$ = .764, .776, and .735 for different picture narrative, listening-to-speaking, and argumentative tasks, respectively). Their findings confirm that the tendency of lower annotation accuracy in an unseen dataset is common across three different speaking task types. From the perspective of L2 fluency studies, different task designs, such as closed (e.g., picture narrative) and open tasks (e.g., argumentation), result in different cognitive demands (Suzuki & Kormos, 2023), possibly differentiating how oral proficiency and cognitive processes are reflected in temporal speech characteristics including hesitations and pausing behaviors across tasks. Task modality (monologue vs. dialogue) also impacts the triad of UF because in dialogue a speaker can utilize an interlocutor's utterances as speech production resources (i.e., alignment), potentially leading to faster speech and fewer pauses and disfluency than monologue (Tavakoli, 2016).

In addition to automatic annotation accuracy, task design and modality can moderate automated speech scoring accuracy. According to Suzuki et al. (2021) meta-analysis, task designs had a significant moderator effect on a UF-PF link ($Q(1) = 7.91$, $p = .019$), and post-hoc tests showed that controlled production (e.g., read-aloud) had a stronger effect size ($r = .74$) than closed and open tasks (r = .53, .51, respectively). They also found a significant moderator effect of task modality ($Q(1) = 29.14$, $p < .001$) though an effect size of dialogic tasks was not significant ($r = .08$, $p = , 389$) partly due to a small number of primary studies. Peltonen (2021) demonstrated that SF and composite measures had strong correlations with PF ratings in a dialogic speaking task (r = .65, .83, respectively), but pause frequency and duration measures were weakly or moderately correlated with the ratings ($r = |.30 - .50|$). These findings altogether indicate that UF measures differently contribute to PF across speaking tasks. It may thus be plausible that the accuracy of automated speech scoring depends on the accuracy of automatic annotation on which fluency measures are based. Given the potential use of systems in a range of speaking tasks, we therefore decided to evaluate the task robustness of an annotation system using different datasets for training and testing.

Proficiency level. The automatic annotation performance could be degraded when the input speech contains phonological and linguistic errors and excessive disfluency phenomena, both of which can hinder the ML-based predictions in each module of a system. In light of proficiency levels, those speech features might be characteristics of beginner L2 learners. Due to the limited availability of L2 spoken datasets, ML-based systems are typically developed using large L1 datasets (e.g., Matsuura et al., 2022). This method has thus suffered from a discrepancy between training and test datasets in the patterns of disfluency, which can cause poor system performance (Knill et al., 2018, 2019). Knill et al. (2019) showed that L2 English learners at the lower-proficiency levels produced more grammatical errors, which subsequently increased the word error rate (WER) of their ASR system (A1 level: $\boldsymbol{WER} = 39.0\%$; C level: $\boldsymbol{WER} = 21.0\%$). Possibly due to the unique patterns of grammatical errors in beginners' speech, prediction errors are more likely to occur and lead to the disagreement between manual and automatic annotation results.

Another reason for the poor performance caused by proficiency levels is that prediction errors occurred in a module are propagated to subsequent modules in a system. Annotation systems for L2 speech are commonly implemented with individually developed modules in a cascade fashion because training or tuning each module independently is feasible even with a relatively small dataset (Lu et al., 2022). Nevertheless, since each module inherently cannot reach perfect accuracy of predictions, the input from downstream modules inevitably contains some prediction errors from upstream modules. Moreover, ML-based modules are usually developed using error-free datasets (e.g., manual transcriptions), and thus erroneous inputs can be out-of-domain, resulting in more prediction errors when inputs are erroneous speech samples (Lu et al., 2022). Furthermore, the number of prediction errors caused in earlier modules of the system can affect the degree of disruptions in downstream modules (Knill et al., 2018, 2019; Tao et al., 2014). Knill et al. (2018) reported that the increase in the WER of ASR can substitute, delete, or insert words and thus affects the syntactic structure of the given utterance, which could subsequently lower the accuracy of clause boundary annotations. Annotation systems should thus be tested to ensure that annotation accuracy is maintained regardless of the proficiency level as a proxy for the difficulties for modules primarily trained by L1 datasets (de Jong et al., 2021). We therefore investigate the extent to which an annotation system is robust to proficiency level, operationalizing as listener-based fluency judgments.

The current study

Previous studies have demonstrated the potential of ML techniques to automatically annotate complicated features in speech, including disfluency words and pause locations. Connecting ML-based modules in a cascade structure, automatic temporal feature annotation systems can be developed for a diverse application in L2 speech research. Meanwhile, speaking tasks and proficiency levels potentially moderate automatic annotation accuracy due to a mismatch between training and application. Therefore, the current study developed an ML-based annotation system and evaluated its robustness to speech tasks and proficiency levels. To this end, we employed L2 English monologue and dialogue datasets collected through five speaking tasks for the evaluation. In addition, operationalizing proficiency levels as PF ratings, the automatic annotation accuracy was tested on subsets of the datasets divided by the three levels of ratings (low, mid, high). This study addressed the following research questions (RQs):

- RQ1. To what extent can the ML-based system accurately annotate disfluency words and pause locations across different speaking tasks and fluency levels of learners?
- RQ2. To what extent can automatically calculated utterance fluency measures simulate the corresponding measures based on manual annotations across different speaking tasks and fluency levels of learners?
- RQ3. To what extent can the automatically calculated UF measures predict listener-based judgements of fluency and how stable are they across different speaking tasks?

Method

Materials

English monologue corpus

To evaluate the ML-based temporal feature annotation system, we utilized a corpus consisting of 512 English monologues elicited via four different speaking tasks (Suzuki & Kormos, 2023). The monologues were collected from 128 Japanese English learners, and speech duration is around two minutes on average. The following four speaking tasks with different cognitive demands for speech production were adopted: an argumentative, a picture narrative task, a reading-to-speaking (RtoS), and a reading-while-listening-to-speaking (RwLtoS) tasks (see Suzuki & Kormos, 2023).

For all speech samples, the pruned transcription and the manual annotation of clause boundaries and pause locations were available. As a result of the fluency annotation, the corpus contains 29,936 MCPs and 9,218 ECPs. In addition, the initial one-minute segments of 512 speech samples were evaluated in terms of fluency by two raters independently. Following a meta-analysis (Suzuki et al., 2021), they used a nine-point scale (1 = "*Not fluent at all*", 9 = "*Very fluent*") with a focus on temporal aspects of speech. The inter-rater reliability reached .819 with a Cronbach's α, indicating a substantial agreement. To obtain rating scores controlling for task difficulty and rater severity, we conducted a Rasch analysis (Linacre, 1994), which is a psychometric technique to estimate latent capacity given measurement influence.

English dialogue corpus

To evaluate the annotation system with speech data in dialogic tasks, we also employed English dialogue corpus of interview speech

data collected through the Wizard of Oz (WoZ) system (Saeki et al., 2022). In the WoZ system, trained human interviewers operate the pre-determined utterances and motions of a spoken dialogue system to minimize interlocutors' variability (cf. Galaczi & Taylor, 2018). The dialogues were collected from 85 Japanese English learners using an interview task. Each interview consisted of seven different topics and lasted nine minutes on average.

The dialogue corpus also includes manual transcriptions, disfluency words and pause location annotations, and oral fluency ratings. Pauses between turns were excluded because dialogic-specific features were beyond the scope of validation. Regarding the temporal features, the corpus includes 3,935 disfluency words, 7,574 MCPs, and 2,414 ECPs. The learners' oral fluency was evaluated by three raters on the CEFR scale of fluency (Council of Europe, 2020). The inter-rater reliability was .804 in terms of Krippendorff's α, indicating a substantial agreement. Another Rasch analysis (Linacre, 1994) was performed to calculate fluency scores while controlling for rater's severity.

Architecture of automatic fluency feature annotation system

The current study targeted a set of UF measures shown in the Table 1. Since a meta-analysis (Suzuki et al., 2021) showed that the each frequency of filled pauses, repetitions, self-corrections, and false starts have negligible predictive power of fluency judgments, we excluded filled pauses from the scope of the annotation system and categorized others as disfluency words without distinction. To obtain UF measures, we followed an annotation method in Matsuura et al. (2022). Fig. 1 shows the pipeline to automatically annotate disfluency words and pauses. This system consists of six modules: speech recognition, pause detection, sentence segmentation, disfluency detection, disfluency pruning, and clause boundary detection.[1]

Speech recognition

The first module, *speech recognition*, takes a speech signal as the input and generates a transcription as the output. We employed Asynchronous Speech-to-Text provided by Rev.ai.[2] The WER of Rev.ai's Asynchronous Speech-to-Text achieved 16.7% for all the corpora comparing their manual and automatic transcriptions by a python package, jiwer.[3] This WER was lower than the other ASR systems which have been used for L2 speech processing (e.g., Coulange et al., 2024); WER of Google Speech-to-Text[4] was 37.2%; and WER of Whisper (Radford et al., 2022) was 20.4%. In addition, the WER of the ASR system adopted by *SpeechRate*^{SM} v.5.0 (Chen et al., 2018) is reported to be 28.5%, indicating that substantial accuracy of Rev.ai's Asynchronous Speech-to-Text.

Pause detection

The second module is *pause detection* which identifies silence intervals longer than a predetermined threshold as the output from the speech input. To detect silence intervals, we employed deep neural network based Forced Alignment (FA),[5] consisting of pre-trained Wav2Vec2.0 (Baevski et al., 2020) and the connectionist temporal classification segmentation algorithm (Graves et al., 2006). This FA does not require a dictionary to map words and phonemes and can handle less frequent words. After detecting silence intervals, this module classifies them as silent pauses or not based on a threshold. We determined 250 milliseconds as the cut-off point because shorter intervals are less representative of L2 proficiency, while longer ones confound pause ratio and duration (de Jong & Bosker, 2013). The pause detection accuracy in terms of Cohen's kappa was .741, indicating the substantial reliability of the current method.

Sentence segmentation

The *sentence segmentation* module receives an automatic transcription as the input and returns the individual sentences as the output. The segmentation was realized by predicting words that would be located at the end of the sentence in the transcription. To predict sentence-ending words, we adopted a pre-trained DistilBERT[6] (Sanh et al., 2020). The DistilBERT was fine-tuned using a Switchboard reannotated dataset (Zayats et al., 2019), which contains L1 spontaneous English dialogue transcriptions with disfluency word annotations. The accuracy score of the sentence segmentation achieved 97.5% in the test data split from the same fine-tuning dataset, indicating high reliability.

Disfluency detection

The disfluency detection module identifies repeated and corrected words from the segmented sentences. Using a BERT-based language model, the disfluency words were detected by predicting whether each word in input sentences is reformulated or not (cf. Lu et al., 2022; Skidmore & Moore, 2023). The current system employed pre-trained RoBERTa[7] (Liu et al., 2019) because of its high-performance in the L2 text processing (e.g., Kyle et al., 2022). We decided to fine-tune the model using a large L1 and a small L2 English corpus ($\boldsymbol{N_{word}} = 1,487,896,\ 14,193$ for L1 and L2 corpora, respectively). As the L1 corpus, we employed the Switchboard reannotated dataset (Zayats et al., 2019). By fine-tuning the model with this L1 data in advance, we expected the model to learn

[1] Scripts for the annotation system are available at https://github.com/RusCucumber/FluencyFeatureAnnotator
[2] https://www.rev.ai/async
[3] https://pypi.org/project/jiwer/
[4] https://cloud.google.com/speech-to-text?hl=en
[5] https://pytorch.org/audio/main/tutorials/ctc_forced_alignment_api_tutorial.html
[6] https://huggingface.co/distilbert/distilbert-base-cased
[7] https://huggingface.co/FacebookAI/roberta-base

Table 1
List of speed, breakdown, repair, and composite measures.

Type	Feature	Description
Speed	Articulation rate (AR)	Number of syllables per speech duration excluding pauses.
Breakdown	Mid-clause pause ratio (MCPR)	Number of mid-clause silent pauses per syllables.
	End-clause pause ratio (ECPR)	Number of end-clause silent pauses per syllables.
	Pause ratio (PR)	Number of silent pauses irrespective of pause location.
	Mid-clause pause duration (MCPD)	Mean duration of mid-clause silent pauses.
	End-clause pause duration (ECPD)	Mean duration of end-clause silent pauses.
	Mean pause duration (MPD)	Mean duration of silent pauses irrespective of pause location.
Repair	Disfluency ratio (DR)	Number of disfluency words (repetitions, self-correction, and false-starts) per syllables.
Composite	Speech rate (SR)	Number of syllables per speech duration.
	Mean length of run (MLR)	Mean number of syllables of speech separated by pauses.

foundational features for the subsequent fine-tuning with the relatively small L2 corpus. Concerning the L2 corpus, we used monologue samples with disfluency word annotations collected from 110 Japanese English learners using a different argumentative task from the test monologue corpus (see Takizawa, 2024). We transcribed the speech samples using Rev.ai's Asynchronous Speech-to-Text to obtain automatic transcriptions. They were then automatically aligned with disfluency word annotations using the jiwer. Using both manual and automatic transcriptions, we further fine-tuned the disfluency detection module to minimize the gap between training and test datasets. The Cohen's kappa between manual and automatic disfluency detection in the test data of the automatically transcribed L2 corpus was .737, suggesting the substantial degree of agreement.

Disfluency pruning

The function of the fourth module, *disfluency pruning*, is to remove disfluency words detected in the disfluency detection module. This module is essential to calculate UF measures primarily because the disfluency pruning standardizes the number of syllables (Suzuki & Révész, 2023). It can also decrease prediction errors in the subsequent clause boundary detection module because its L1 training datasets include relatively few disfluency words (see Kyle et al., 2022).

Clause boundary detection

The last module is *clause boundary detection* which classifies detected silent pauses into either MCPs or ECPs. The prediction of clause boundary locations is achieved by a dependency parser. In this study, we utilized the RoBERTa-based parser which was fine-tuned by L1 written English text and L2 English dialogue corpora (Kyle et al., 2022). The high reliability of the parser in L2 speech was reported as .938 in terms of F1.

Analysis

Evaluation of automatic fluency annotation

To test the accuracy of annotating disfluency words and pause locations, we compared manual and automatic annotations using both monologue and dialogue corpora. More specifically, we regarded the current disfluency word and pause location annotation as an NLP task to predict whether target events occurred after each word. We evaluated disfluency word and pause location detection in terms of the following four metrics: *Cohen's kappa* indicates the agreement between manual and automatic annotations; *Precision* is defined as the proportion of true positives in the automatically detected items; *Recall* refers to how well target items are automatically detected; and *F1* is integrated metrics of Precision and Recall. The strength of agreement in terms of Cohen's kappa was interpreted as moderate ($\kappa > .41$), substantial ($\kappa > .61$), and almost perfect ($\kappa > .81$) (Landis & Koch, 1977). To calculate the four metrics considering ASR errors, we aligned the manual and automatic annotation results using NIST's SCTK[8] following Chen and Yoon (2012).

Moreover, disfluency words were manually added to the pruned transcriptions of the monologue corpus by four trained coders. After a one-hour orientation, they coded 20 speech samples randomly chosen from the monologue corpus to check the inter-coder reliability. The mean WER of four coders' transcriptions was 7.277 %, confirming high inter-coder reliability. Afterwards, the whole dataset was divided into four subsets, each of which was assigned to a different coder. As a result, they added 10,525 disfluency words in the whole 512 monologue transcriptions.

To examine annotation system's robustness to speaking tasks and listener-based fluency judgments, we calculated Cohen's kappa, precision, recall, and F1 for the test dataset divided by tasks (i.e., argumentative, picture narrative, RtoS, RwLtoS, and interview) and PF score groups (i.e., low, mid, and high). The low, medium, and high fluency groups include samples with ratings of 1–3, 4–6, and 7–9 in monologue and A1-A2, B1-B2, and C1-C2 in dialogue, respectively. Table 2 shows the sample sizes of each divided dataset. According to power analyses, the minimum sample sizes to show substantial agreement in terms of Cohen's kappa (i.e., $\kappa > .61$) with power $\beta = .8$ were 151 and 67 for disfluency detection and pause location classification, respectively (Donner & Eliasziw, 1987), which the current divided datasets satisfied.

[8] https://github.com/usnistgov/SCTK

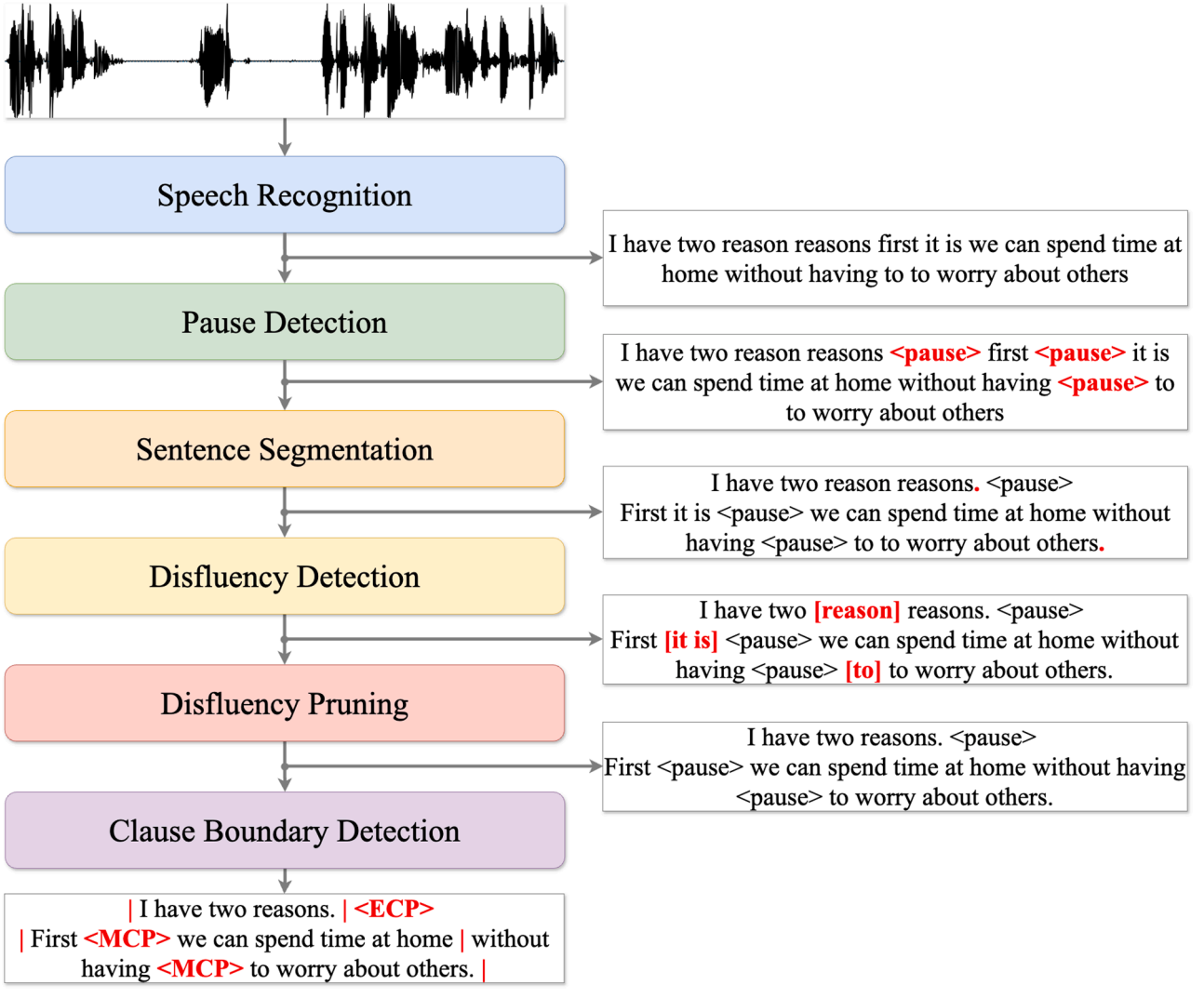


Fig. 1. The architecture of automated temporal feature annotation system.

Evaluation of the UF measure reliability and robustness

To evaluate the automatically calculated UF measures in terms of reliability and the robustness to speaking tasks and listener-based fluency judgments, we conducted correlation analyses and prediction of subjective judgments of L2 oral fluency (cf., de Jong et al., 2021). We calculated Person's correlation coefficients between the UF measures derived from manually and automatically annotated data. Comparing correlation coefficients across tasks, we evaluated the robustness of the current system to the speaking tasks differing in cognitive demands and modality. As for the stability to listener-based fluency judgments, we evaluated the system across the low, mid, and high fluency score groups. The strength of correlation coefficients was interpreted as small ($r \geq .25$), medium ($r \geq .40$), and large ($r \geq .60$) (Plonsky & Oswald, 2014).

In addition, we constructed a series of linear regression models to examine the extent to which the automatically calculated UF measures can predict listener-based fluency ratings. As the ground truth for fluency ratings, we used continuous logit scores with rater severity controlled by Rasch Analyses (Linacre, 1994). The regression models were evaluated in terms of R^2 (Chen et al., 2018; de Jong et al., 2021). Due to the relatively small size of each corpus (i.e., 128 and 85 for monologic and dialogic corpora, respectively), we utilized the stratified k-fold cross-validation technique to separate data into training and test sets considering the distribution of fluency judgments. To keep the ratio of training data to test data at 8:2, k was set to five. Moreover, to evaluate the construct validity of automatic measures, we compared the relative importance of manual and automatic predictor variables in regression models. More specifically, among five regression models trained on manual measures, we picked up a best model in terms of R^2. A best automatic model was chosen to be the one trained on the same training data as the best manual model. Afterwards, employing dominance analyses (Budescu, 1993), which calculates the contribution of each predictor variable to R^2, we compared relative importance of manual and automatic measures.[9]

Results

The agreements of disfluency detection and pause location classification across the five tasks are summarized in Table 3. Note that we excluded one speech sample in the RtoS task because WER was too high (123 %) to accurately annotate temporal features possibly due to many Japanese fillers. Cohen's kappa coefficients indicated a substantial agreement of both disfluency detection and pause location classification between human and the system.

Figs. 2-5 illustrate Cohen's kappa, precision, recall, and F1 between the manual and automatic annotations of disfluency words and pause locations across the tasks and PF score groups. The automatic disfluency detection achieved substantial agreement with manual annotation across the monologic tasks and PF score groups ($\kappa = .635 - .734$), while the agreement was moderate in the dialogic task in all PF score groups ($\kappa = .558 - .607$). As for the pause location classification, Cohen's kappa was slightly lower than 0.61 in the subset of high PF score group in the interview task ($\kappa = .596$), but the other subsets yielded substantial agreement ($\kappa = .626 - .749$).

Agreement of utterance fluency measures between human and the system

Table 4 summarizes the correlation coefficients between the manual and automatic UF measures across the speaking tasks. Fig. 6 depicts a scatter plot of manual and automatic measures of the argumentative tasks (see Appendices A-D, for scatter plots of the other tasks). The correlation aggregated from all speaking tasks was found strong in all the UF measures except for AR and ECPR, both of which exhibited moderate effect sizes. The analyses also demonstrated that correlation patterns of UF measures in each task were similar to those when aggregated by tasks whereas speech rate in the interview task had a medium correlation, compared with strong correlations in other subsets.

Despite the generally strong correlations between the manual and automatic measures in the task-level analyses, more nuanced relationships between correlation coefficients and PF score groups were visualized in Fig. 7. First, the correlations between the automatic and manual MPD, DR, and SR were relatively stable across PF scores. However, the correlations of SR in dialogic tasks were relatively lower, especially for low and mid PF score groups, and their 95 % CIs were wide. Second, the correlations of MCPD and MLR linearly changed as a function of the PF group levels. Third, the correlation patterns of MCPR, ECPD, and PR across PF score groups differed by task modality. For instance, in the monologic tasks, the correlations of MCPR increased towards the higher PF score group but decreased in the interview task.

Predictive power of automatic utterance fluency measures

The mean R^2 of five-fold cross validation are summarized in Table 5. Surprisingly, the R^2 for the automatic measures were higher than manual ones in the argumentative, picture narrative, and RtoS tasks. On the other hand, the RwLtoS and interview tasks yielded lower R^2 for the automatic measures.

Fig. 8 depicts the results of dominance analyses for the best regression models in the five-fold cross validation. In the regression model based on manual measures, composite measures tended to be relatively most important. Although the relative importance of each measure differed across the tasks, pause duration (e.g., MCPD, ECPD, MPD) and pause ratio (e.g., MCPR, PR) measures followed composite ones in the monologic tasks. Meanwhile AR and pause ratio measures (e.g., MCPR, PR) tended to be relatively important in

[9] Scripts for analyses are available at https://github.com/RusCucumber/fluency_feature_annotation_experiment

Table 2
The number of words in speech samples in test datasets.

	Low	Mid	High	Total
Argumentative	3,109	7,066	4,448	14,623
Picture Narrative	3,891	9,492	5,435	18,818
RtoS	4,210	9,223	5,795	19,228
RwLtoS	3,930	9,622	5,699	19,251
Interview	7,301	24,792	6,664	38,757

Table 3
Results of overall evaluation of manual and automatic annotation.

Category	κ	95 % CI	N_{manual}	N_{auto}	Precision	Recall	F1
Disfluency Detection							
	0.658	[0.652, 0.665]	14,460	12,922	0.736	0.658	0.694
Pause Location Classification							
MCP	0.707	[0.704, 0.711]	29,936	36,853	0.678	0.834	0.748
ECP			9,218	8,657	0.698	0.655	0.679

the interview task. In the regression models based on automatic measures, composite and pause duration and ratio measures showed the similar patterns to the manual ones. However, the automatic AR was found more relatively important, compared with the corresponding manual one in the monologic tasks, and vice versa in the interview task.

Discussion

Accuracy of automatic temporal feature annotation

To evaluate the annotation accuracy of disfluency detection and pause location classification, we compared automatic and manual annotation using Cohen's kappa. The substantial agreement was confirmed for both disfluency detection and pause location classification in all monologic speaking tasks regardless of PF score groups. Meanwhile, in the dialogic interview task, the substantial agreements were observed for disfluency detection for all PF score groups. However, pause location classification yielded a moderate agreement with manual ones for high PF score group while demonstrating substantial agreements for low and mid PF score groups. These results suggest that the current system can robustly annotate disfluency words and pause locations for Japanese English learners' speech samples particularly elicited from monologic tasks.

The substantial agreements of disfluency detection for monologues elicited from different tasks were achieved possibly because training and test datasets were consistent in the monologic modality. The F1 of disfluency detection was .784 in the test dataset split from the training monologic dataset (argumentative), while those in the three unseen monologic test datasets (picture narrative, RtoS, and RwLtoS) ranged from .676 to .766, resulting in .02–.11 accuracy degradation. Meanwhile, the F1 in the interview task was slightly lower than in the monologic tasks approximately by .15–.20 (.579–.648). Previous research reported that the mismatch of task modality between training and test datasets led to an approximately .20 decrease in F1 under different task design conditions and an approximately .17 decrease even under the same task design conditions (Skidmore & Moore, 2023). These findings thus indicate that the robustness of disfluency detection could be moderated by the difference in task modality rather than task designs between training and test datasets.

The moderator effects of task modality on disfluency detection accuracy could be explained by the co-constructive and spontaneous nature of dialogic discourse. For example, in the current interview dataset, some students started their turn by repeating the interviewer's expressions to display their listenership and then started a new utterance with that same expression as the subject, and consequently a repetition occurred (see Excerpt 1).

The current annotation system falsely detected the first "My house" in Excerpt 1 as disfluency words potentially because this type of repetition is annotated as self-repetition in the training monologic dataset. In contrast, human coders may have judged this repetition as an *other-repetition*, which makes the ongoing dialogue more cohesive (Peltonen, 2017a). To improve in the accuracy of disfluency detection for dialogic samples, it would be necessary not only to mix monologic and dialogic corpora for training but also to utilize interlocutor's utterances to avoid detecting other repetitions as disfluency words.

The pause location classification yielded high F1s for MCP (.618–.809) and for ECP (.529–.777), which were comparable to or higher than conventional systems (e.g., $\boldsymbol{F1_{MCP}} = .656$, $\boldsymbol{F1_{ECP}} = .526$; Matsuura et al., 2022). However, care should be taken not to overestimate the current annotation system due to the relatively low overall accuracy of disfluency detection ($\boldsymbol{F1} = .694$). In the interview task for the high PF score group, where the lowest F1 for disfluency detection was found (.579), the agreement of pause location classification was moderate. Considering the substantial agreement of pause location classification in the other settings, the error propagation from the disfluency detection module might be the major reason for the vulnerability of this module. To mitigate the error propagation in a cascaded annotation system for L2 speech, an optimization technique may have been effective to use the actual outputs of upstream modules as inputs for downstream modules (Lu et al., 2022).

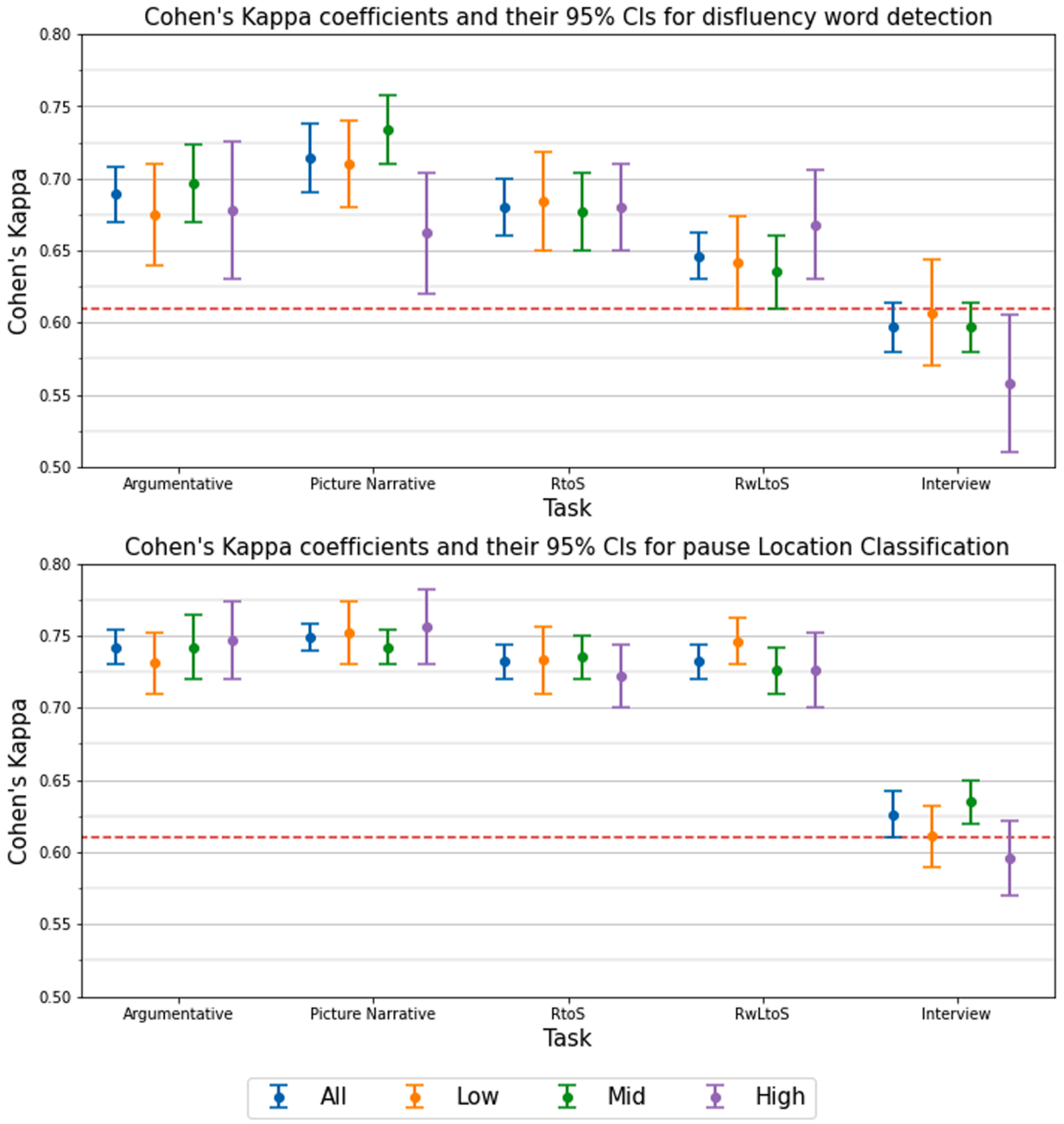


Fig. 2. Cohen's Kappa and Their 95 % CIs for Manual and Automatic Disfluency Detection and Pause Location Classification Across Tasks and PF Score Groups.

Validity of automatic utterance fluency measures

RQ2 addressed the correlations between automatic and manual UF measures as the concurrent validity evidence of our system. We found a general tendency of strong correlations between those two calculation methods. First, MPD, DR, and SR had strong correlations with the corresponding manual measures throughout all subsets in general, showing the robustness of these UF measures to various speech samples. From a language assessment perspective, the measure of SR has been found to be a reliable indicator of L2 proficiency levels (Tavakoli et al., 2020). The current system may thus have the potential of being adopted in automated scoring systems of speaking skills. Second, the strength of correlations varied in MCPD and MLR across PF levels, as well as in ECPD, MCPR, and PR across tasks and PF levels. For instance, automatic MCPR had a strong correlation with the manual measure in most subsets, with the exception of monologues for the low PF group and dialogues for the high PF group. MCPR is associated with cognitive fluency (Suzuki & Kormos, 2023; see also Suzuki & Révész, 2023) and is found to be sensitive to L2 fluency development (e.g., Suzuki & Hanzawa,

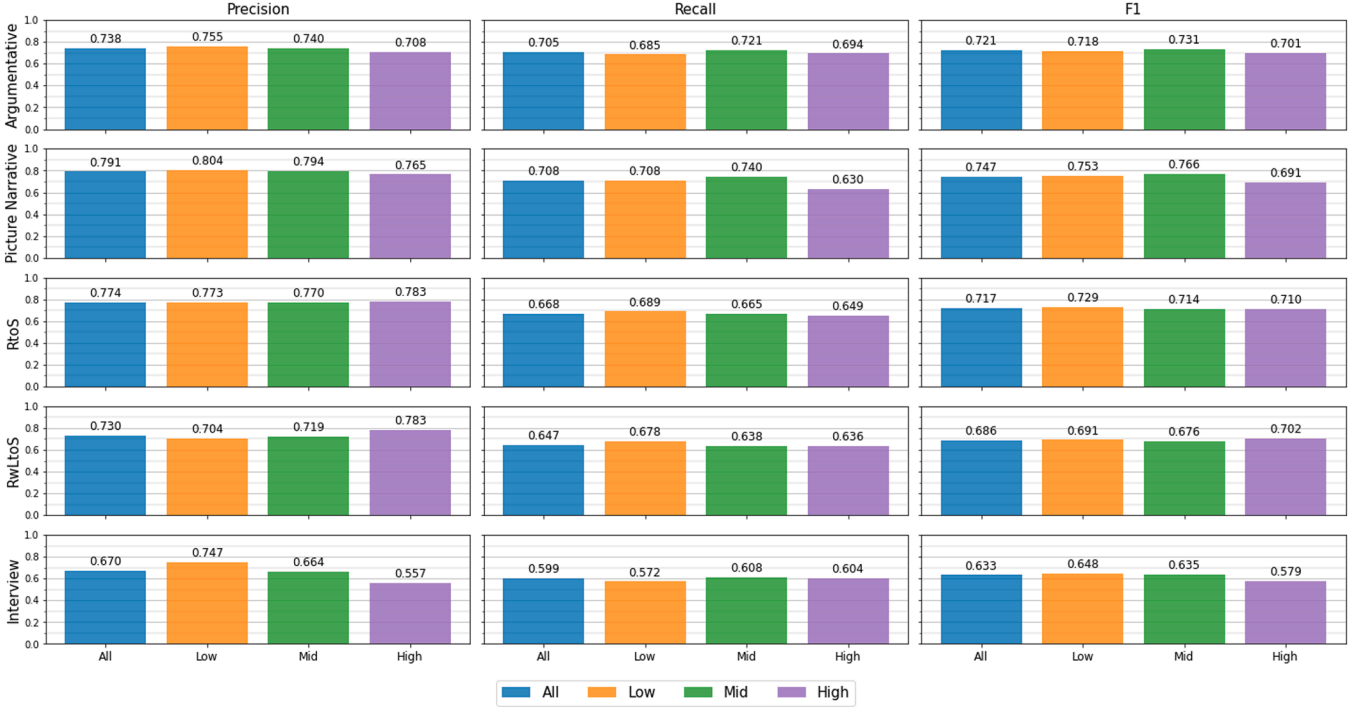


Fig. 3. Precision, recall, and F1 for disfluency detection across tasks and PF score groups.

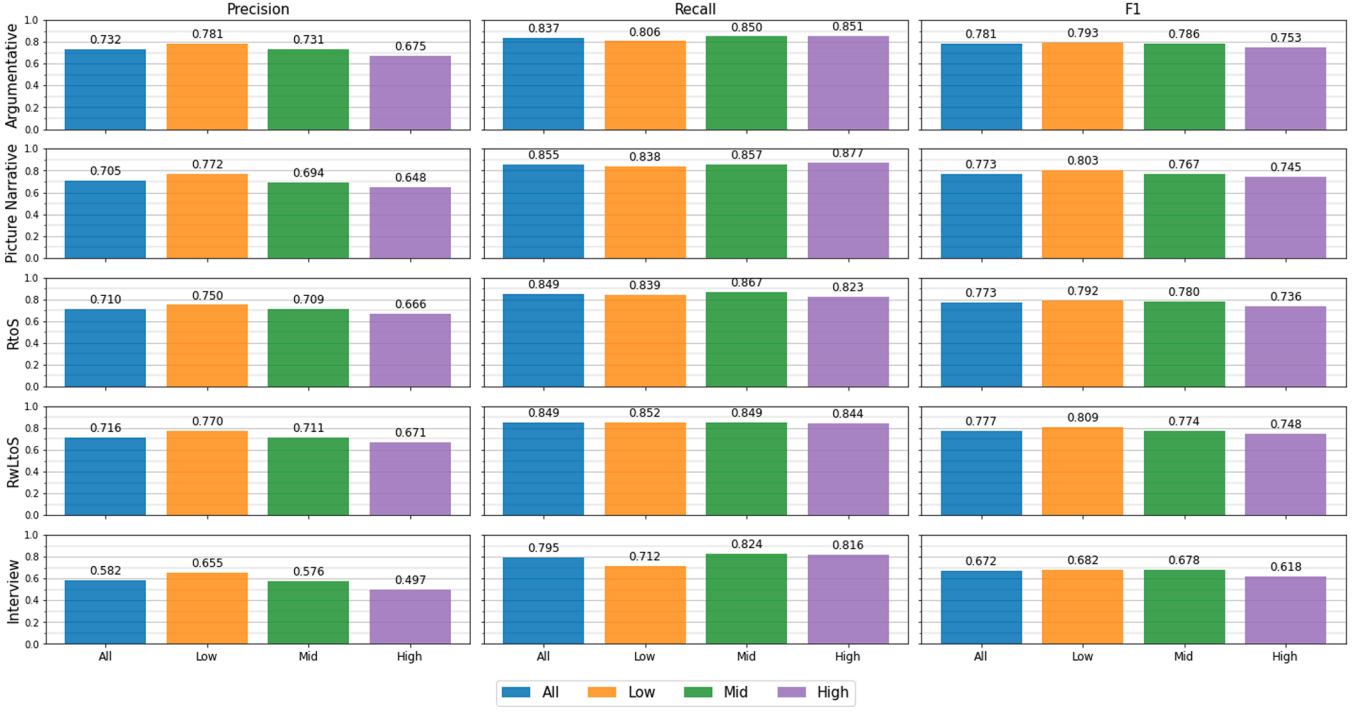


Fig. 4. Precision, recall, and F1 for MCP classification across tasks and PF score groups.

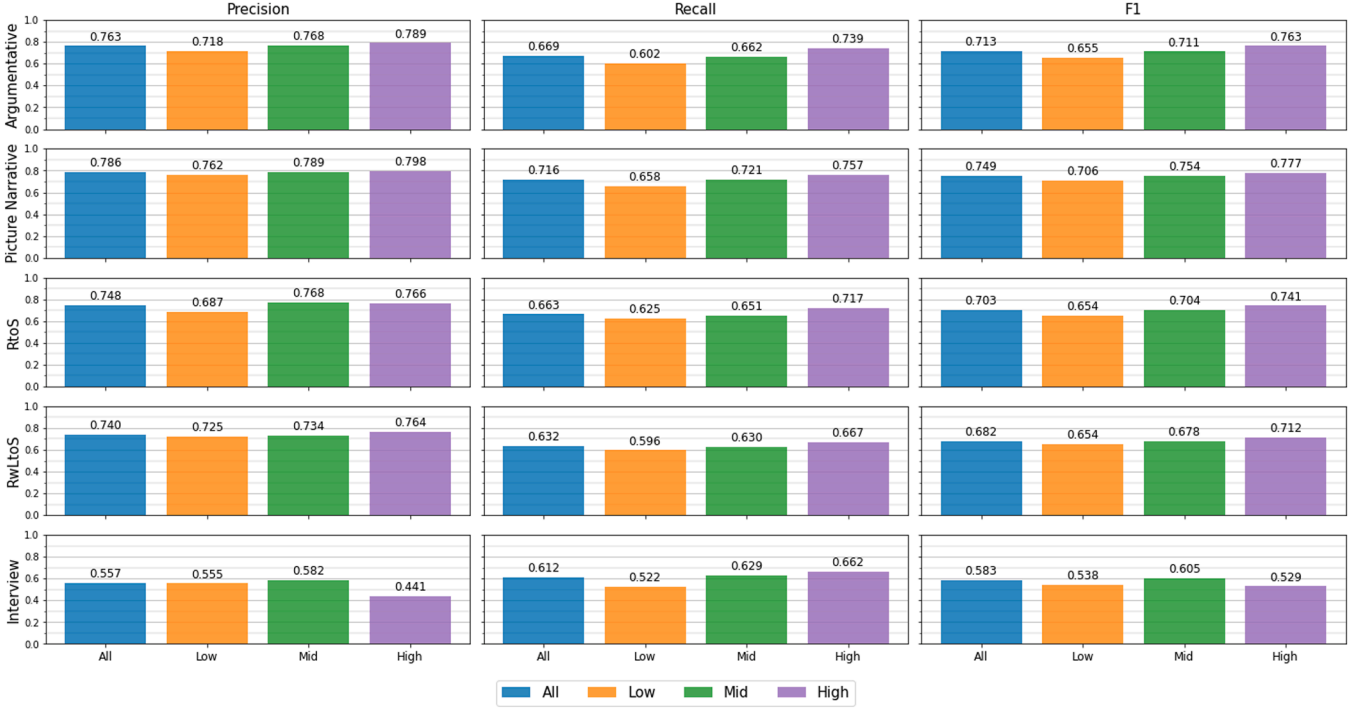


Fig. 5. Precision, recall, and F1 for ECP classification across tasks and PF score groups.

Table 4
Results of correlation analyses between manually and automatically calculated utterance fluency measures.

Measure	All	Argumentative	Picture	RtoS	RwLtoS	Interview
Articulation Rate	0.582***	0.604***	0.663***	0.635***	0.657***	0.506***
	[0.526, 0.633]	[0.480, 0.703]	[0.553, 0.750]	[0.519, 0.729]	[0.546, 0.745]	[0.328, 0.649]
Mid-Clause Pause Ratio	0.804***	0.798***	0.817***	0.789***	0.773***	0.844***
	[0.773, 0.830]	[0.725, 0.853]	[0.750, 0.868]	[0.713, 0.847]	[0.693, 0.835]	[0.769, 0.896]
End-Clause Pause Ratio	0.553***	0.459***	0.616***	0.375***	0.435***	0.470***
	[0.495, 0.607]	[0.310, 0.586]	[0.496, 0.713]	[0.215, 0.516]	[0.283, 0.566]	[0.286, 0.621]
Mid-Clause Pause Duration	0.848***	0.894***	0.865***	0.801***	0.890***	0.816***
	[0.824, 0.869]	[0.853, 0.924]	[0.814, 0.903]	[0.729, 0.856]	[0.848, 0.922]	[0.730, 0.877]
End-Clause Pause Duration	0.719***	0.645***	0.690***	0.802***	0.604***	0.716***
	[0.678, 0.756]	[0.531, 0.736]	[0.587, 0.771]	[0.730, 0.857]	[0.481, 0.704]	[0.593, 0.806]
Pause Ratio	0.798***	0.780***	0.823***	0.777***	0.750***	0.833***
	[0.767, 0.825]	[0.701, 0.840]	[0.758, 0.872]	[0.697, 0.838]	[0.662, 0.817]	[0.754, 0.888]
Mean Pause Duration	0.858***	0.885***	0.844***	0.883***	0.890***	0.791***
	[0.835, 0.878]	[0.841, 0.918]	[0.786, 0.888]	[0.838, 0.916]	[0.847, 0.921]	[0.695, 0.859]
Disfluency Ratio	0.818***	0.868***	0.825***	0.828***	0.745***	0.885***
	[0.789, 0.843]	[0.818, 0.905]	[0.761, 0.874]	[0.764, 0.876]	[0.656, 0.813]	[0.828, 0.924]
Speech Rate	0.897***	0.959***	0.972***	0.960***	0.954***	0.574***
	[0.880, 0.911]	[0.942, 0.971]	[0.960, 0.980]	[0.944, 0.972]	[0.935, 0.967]	[0.412, 0.702]
Mean Length of Run	0.865***	0.889***	0.882***	0.875***	0.889***	0.896***
	[0.843, 0.884]	[0.846, 0.920]	[0.836, 0.915]	[0.827, 0.910]	[0.846, 0.921]	[0.844, 0.931]

Note. ** $p < .01$; *** $p < .001$.

2022). While the minimal check of the automatic MCPR may be required according to speaking tasks and learners' proficiency levels, the current result may confirm the potential of the system for the use in fluency studies by minimizing labor-intensiveness of manual annotations. Nevertheless, as MCPR tends to have a relatively lower contribution to overall proficiency scores than other UF measures (Chen & Yoon, 2012), the computation of a full set of the current measures could mitigate the bias in automated scoring of other constructs. Finally, the correlations of AR and ECPR were weak-to-moderate consistently across tasks and PF groups. In particular, AR is one of the most crucial measures to improve the validity of automated scoring systems because it is arguably the only pure measure of speed fluency (Suzuki et al., 2021; Tavakoli et al., 2020), and it is strongly related to the communicative aspects of oral proficiency (Handley & Wang, 2023).

The strong correlations between the automatic and manual SR and MPD were confirmed possibly because automatic annotation errors may have been ignorable in the measure calculation process. For example, regardless of manual and automatic annotation, speech duration is constant, and thus the strength of the correlation of the two SRs should depend solely on the accuracy of the number of syllables counted. Although transcription errors are unavoidable in ASR for L2 speech (Knill et al., 2018), the syllable counts can approach a level comparable to those based on manual transcriptions if the ASR module can recognize syllable structures correctly. Meanwhile, pause detection errors could be non-negligible to automatically calculate AR. Falsely detected pauses in the automated condition lead to the longer total duration of pauses, possibly resulting in a relatively shorter phonation time than in the manual conditions. Given the number of syllables counted was close to the manual condition in the current study, the discrepancy of pause detection between the conditions could have lowered the correlation coefficients of AR, which uses the phonation time as the denominator for AR. Moreover, learners' PF levels may have moderated the degree to which automatic annotation errors can be reduced. For instance, the correlations of automatic and manual MCPD were weak or moderate in the high PF subset. As the frequency of MCP decreases as a function of fluency levels (Suzuki et al., 2021), when calculating the average duration of MCPs, the weights of falsely detected pauses might have been relatively large.

Predictive power of automatic utterance fluency measures

Finally, we examined automatic measures' predictive power for listener-based fluency judgments across tasks by training linear regression-based automated scoring systems. The mean R^2 of gold standard automated scoring systems, which were trained using manual measures, was high for monologic tasks ($\boldsymbol{R}^2 = .641 - .740$). On the other hand, lower explained variance was found for the interview task ($\boldsymbol{R}^2 = .510$). The relative high accuracy of automated scoring for monologues aligned with the meta-analyses of a UF-PF link (Suzuki et al., 2021), which reported the higher correlation coefficients between PF and UF measures in monologic tasks than in dialogic tasks. In other words, the difference in prediction accuracy between monologues and dialogues in the current study should be attributed to the difference in the construct of fluency between the modalities (Peltonen, 2021).

The follow-up dominance analyses for the gold standard system demonstrated that composite measures contributed most to the prediction of fluency judgments regardless of tasks. In the monologic tasks, apart from the composite measures, BDF, SF, and RF measurements were relatively important in that order, whereas in the dialogic task SF measure were relatively more important than BDF ones (Peltonen, 2017b, 2021; Suzuki et al., 2021). In addition, RF and PF have a weak correlation because RF may reflect individual speaking style (de Jong et al., 2015), and thus relatively lower contribution of the RF measure to PF scores should be valid. Taken together, these results suggest that the current gold standard automated scoring systems sufficiently simulated human judgments of oral fluency.

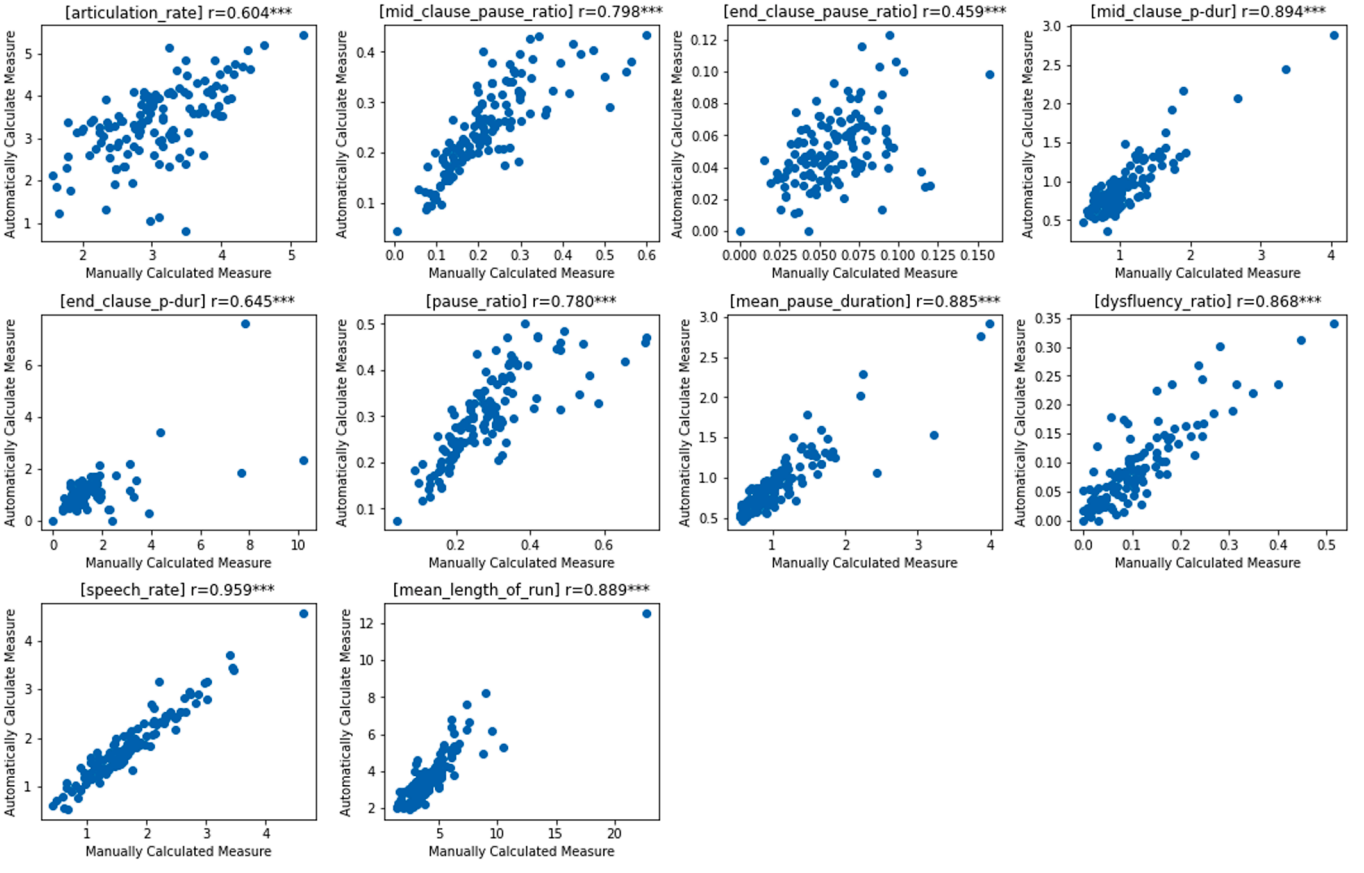


Fig. 6. Scatter plots of manual and automatic measures of argumentative task. *Note.* Blue, red, and green shades indicate strong, medium, and weak correlations, respectively.

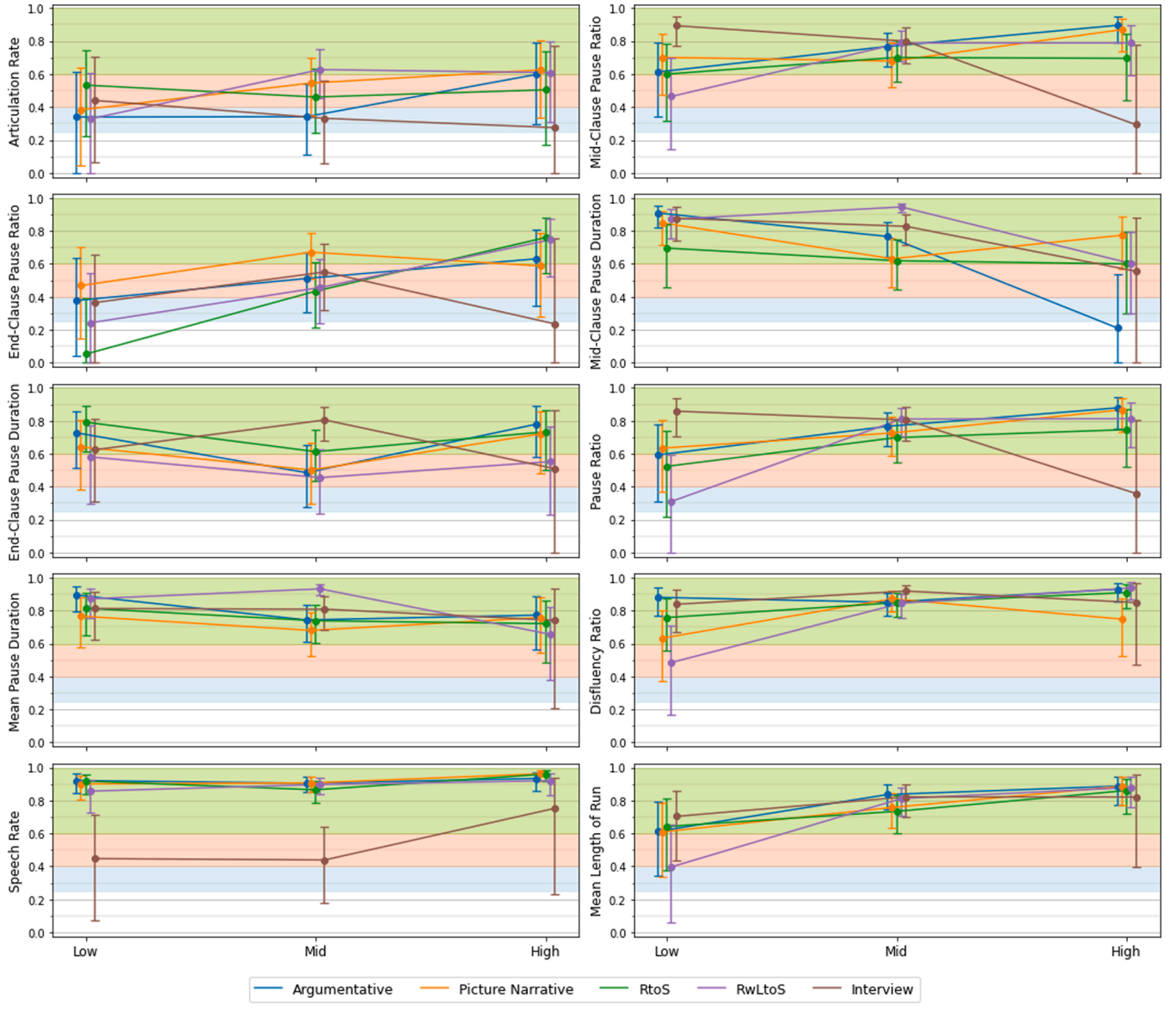


Fig. 7. Correlation coefficients and their 95 % CI between manual and automatic UF measures across tasks and PF score groups.
Note. Blue, red, and green shades indicate strong, medium, and weak correlations, respectively.

Table 5
Total R^2 for linear models predicting (mean) scores for perceived fluency.

Task	Manual Measures	Automatic Measures
Argumentative	0.696	0.740
Picture	0.628	0.745
RtoS	0.737	0.752
RwLtoS	0.695	0.548
Interview	0.510	0.328

The automated scoring analyses based on automatic measures showed higher accuracy of fluency judgment prediction in the monologic tasks ($R^2 = .641 - .783$). The dominance analyses also showed relatively strong contributions of automatic composite, SF, and BDF measures to the fluency judgment prediction. These results indicate the valid predictive power of automatic measures in monologic tasks. Although the accuracy of automated scoring is not directly comparable across studies due to various methodological differences, the current $\boldsymbol{R}^2$ scores were higher than the previous systems (e.g., $\boldsymbol{R}^2 = .\mathbf{320}$; de Jong et al., 2021). Especially, the accuracy of the current automatic scoring predictions outperformed that of SpeechRater^{SM} v5.0 for TOEFL iBT ($\boldsymbol{R}^2 = .584$; Chen et al., 2018), suggesting that our annotation system might have potential to be employed even in high-stakes assessment contexts. Despite the high accuracy of automated scoring for monologues, the analyses demonstrated that the prediction accuracy in the interview task were lower ($\boldsymbol{R}^2 = .336$) than the gold standard and conventional systems. One possible reason for the relatively low accuracy in the interview task might be the discrepancy between the manual and automatic AR measures. The dominance analyses showed the relatively lower importance of automatic AR than the manual one in the PF score prediction in the interview task. Given the potentially stronger link between SF and PF in dialogic speech (Peltonen, 2021), the distortion in automatic AR could have critically lowered the prediction accuracy of fluency judgment in the interview task.

Surprisingly, we found that the automatic UF measures explained more variance of PF scores than the manual measures in the monologic tasks. In addition, the dominance analyses demonstrated the stronger contribution of automatic AR than the manual one. These results indicate that the construct validity of the current automatic UF measures should be further tested. Especially, automatic AR might encompass pronunciation aspects of speech beyond the target construct of SF. The automatic AR could have been calculated lower than the manual one, especially in the low PF score group, because the number of falsely detected pauses might increase as their speech samples might have had relatively heavier accent. Moreover, listener-based fluency judgments might be biased by non-canonical accent speech because, from the perspective of listener perceptions, pronunciation aspects and temporal characteristics of speech might be hard to distinguish (Suzuki et al., 2021). For instance, syllable structure errors (e.g., vowel insertion) can affect the time duration for the same words. These potential biases in automatic AR and PF scores might cause their higher correlation and larger contributions of the AR to the PF rating score prediction in the model based on automatic UF measures, compared to the manual AR.

Conclusion

Temporal feature annotation systems, implemented by the cascade connection of ML-based modules, have suffered from declining accuracy in a new context with unfamiliar speech elicitation tasks and different speakers' proficiency levels, due to the common distribution gaps between training and real-world data. The current study thus validated an ML-based annotation system on L2 English datasets split by speaking tasks (four monologic and one dialogic) and listener-based fluency levels (low, mid, and high). The results showed the substantial agreement of disfluency detection and pause location classification between the system and experts for various monologic L2 speech. We also found a general tendency of strong correlations between automatic and manual UF measures regardless of speaking tasks, supporting the concurrent validity of our system. Moreover, automatic measures predicted listener-based fluency judgments more accurately in monologic tasks than conventional methods (Chen et al., 2018; de Jong et al., 2021).

One highlight of the current study is that the ML-based temporal feature annotation system stably yielded substantial annotation agreements, strong correlations between automatic and manual UF measures, and high prediction accuracy of fluency ratings in the four different monologic tasks. The current annotation system can thus be considered valid enough to be applied to monologues. However, disfluency detection errors can propagate and potentially lower the pause location classification accuracy. Moreover, the disfluency detection agreements and PF score prediction accuracy did not reach the predetermined thresholds in the dialogic interview task. The automatic AR also may encompass the variability of pronunciation quality due to pause detection errors. Applying the current system to studies that are interested in dialogic tasks and SF should require researchers to check and, if necessary, correct automatic annotations by experts. However, it would still mitigate labor-intensiveness and enhance the scalability of L2 speech research to a large extent. To develop a more robust system, it is necessary to create large datasets including comprehensive annotations (i.e., speech samples, transcriptions, disfluency words, and clause boundaries), various task modalities, and samples from various learner levels and apply the optimization technique to mitigate error propagation (Lu et al., 2022).

There are several methodological limitations in our validation study, one of which is the criteria for robustness evaluation. The current study focused on task types and proficiency levels as target variables due to their potential impacts on ML-based annotation systems. However, it is reported that ASR is affected by genders, L1 backgrounds, and ages (Feng et al., 2024). Additionally, we tested the system only in L2 English. Future studies are thus expected to evaluate the annotation systems taking these factors into account. Second, we only treated the oral proficiency interview as a dialogic task, while other task types, such as role-play (Skidmore & Moore, 2023) and problem-solving (Peltonen, 2021), are also widely used in pedagogical and assessment contexts. Finally, proficiency levels

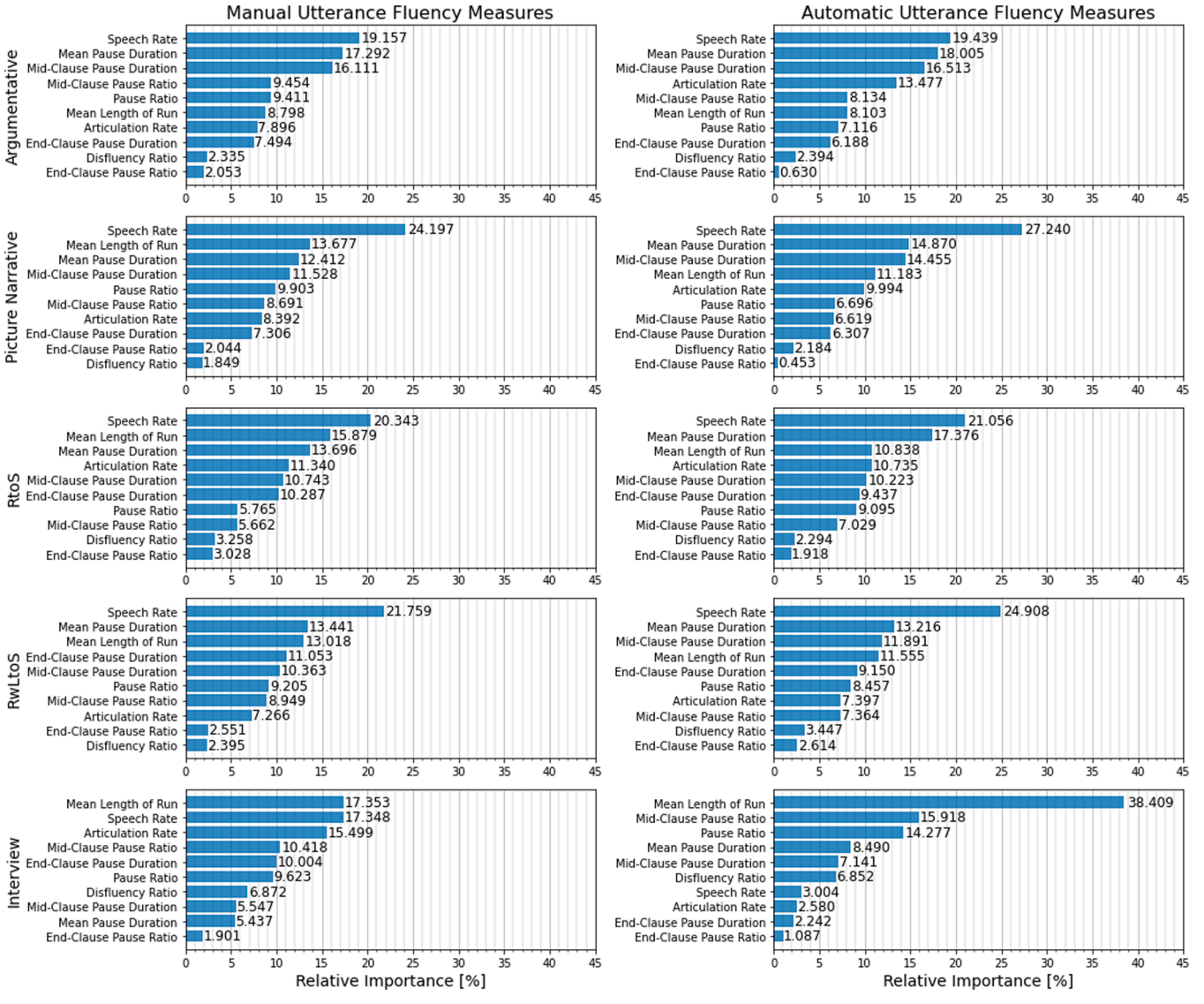


Fig. 8. Relative importance of manual and automatic measures in terms of PF score prediction.

Excerpt 1
Interview question and response with other repetition.

Interviewer:	*Can you tell me about your house?*
Student:	*My house, my house is not so big, but it's comfortable for me.*

are operationalized as PF ratings, and there are various ways of operationalizing oral proficiency with a particular emphasis on communicative aspects of L2 speech such as functional adequacy (de Jong et al., 2012; Handley & Wang, 2023). Another direction of future research is thus to replicate the analyses with more variety of task types and different assessment criteria.

CRediT authorship contribution statement

Ryuki Matsuura: Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Shungo Suzuki:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Kotaro Takizawa:** Writing – review & editing, Resources, Investigation, Data curation. **Mao Saeki:** Writing – review & editing, Resources, Investigation, Data curation. **Yoichi Matsuyama:** Writing – review & editing, Funding acquisition.

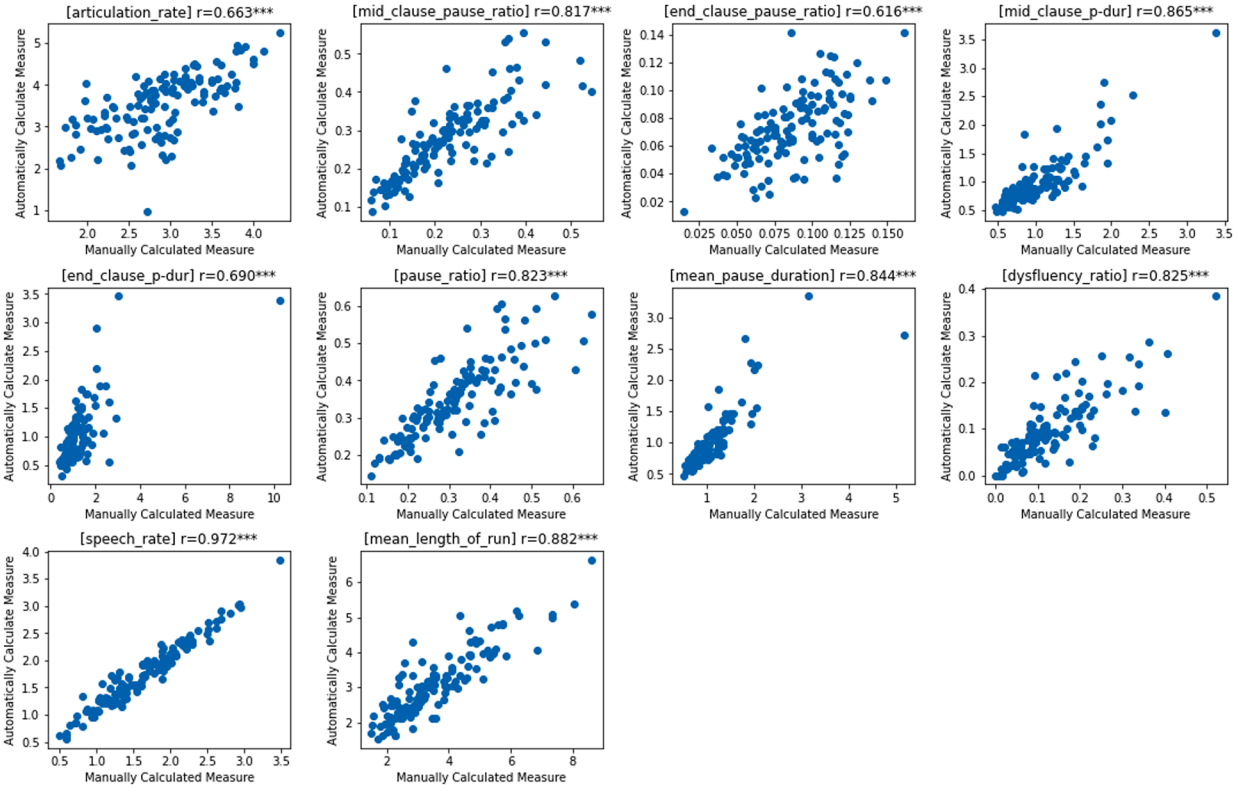
Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

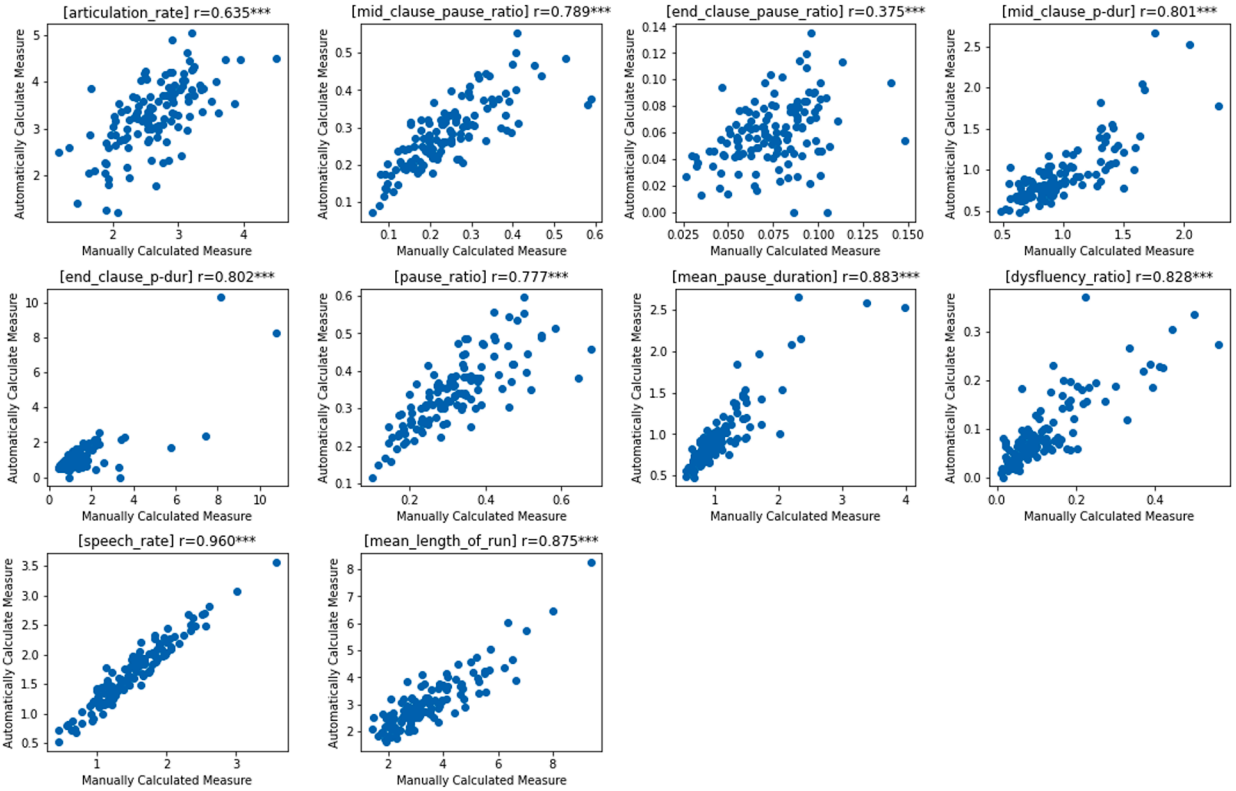
Acknowledgements

We acknowledge Judit Kormos for her help with the collection of fluency rating data for monologue speech. This paper is based on results obtained from a project, JPNP20006 ("Online Language Learning AI Assistant that Grows with People"), subsidized by the New Energy and Industrial Technology Development Organization (NEDO)

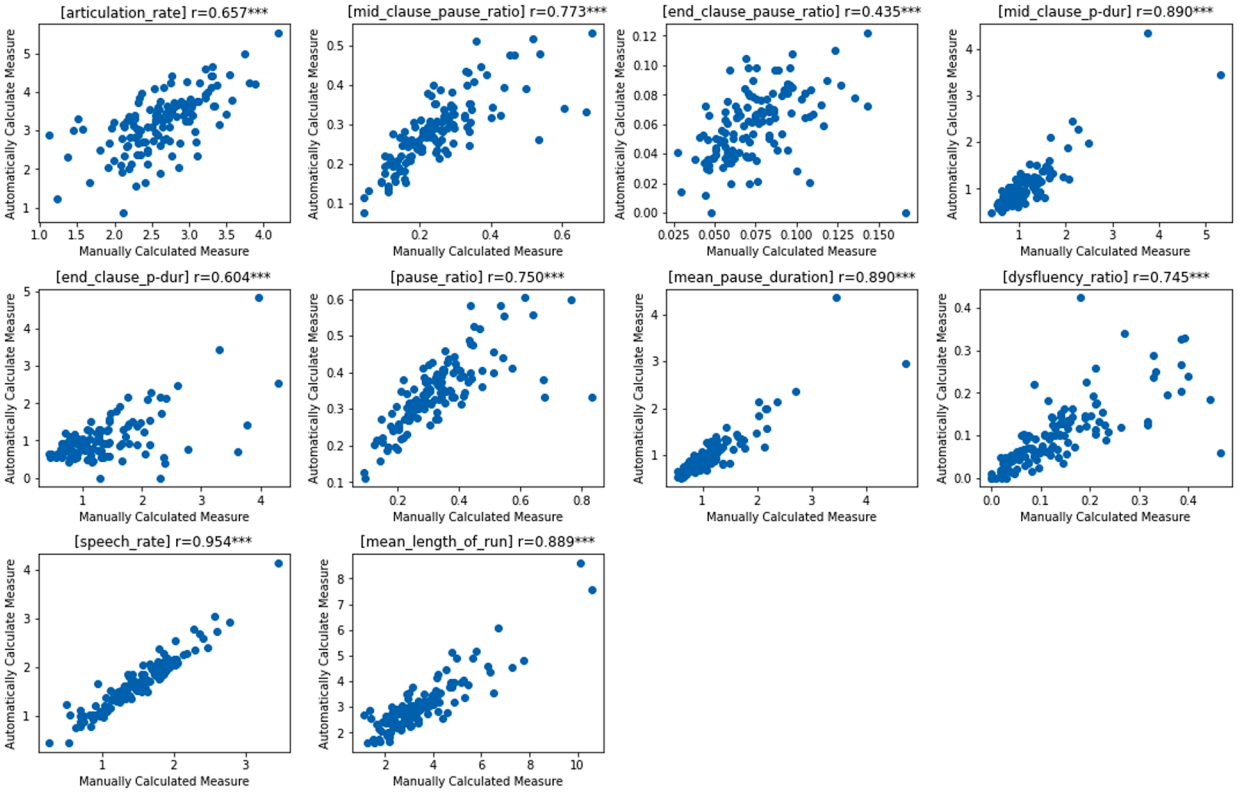
Appendix A. Scatter plots of automatic and manual measures of picture narrative task



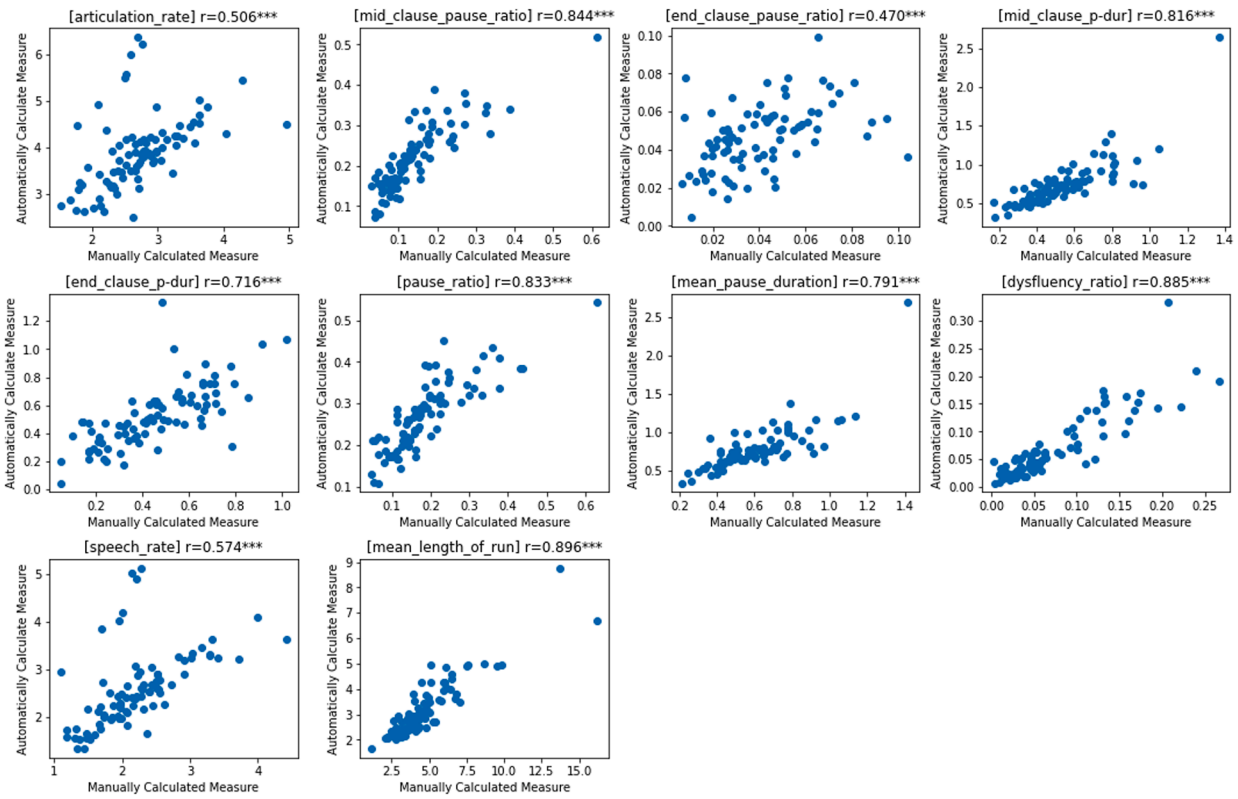
Appendix B. Scatter plots of automatic and manual measures of RtoS task



Appendix C. Scatter plots of automatic and manual measures of RwLtoS task



Appendix D. Scatter plots of automatic and manual measures of interview task



References

Al-Ghezi, R., Voskoboinik, K., Getman, Y., Von Zansen, A., Kallio, H., Kurimo, M., … Hildén, R. (2023). Automatic Speaking Assessment of Spontaneous L2 Finnish and Swedish. *Language Assessment Quarterly, 20*(4–5), 421–444. https://doi.org/10.1080/15434303.2023.2292265

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. arXiv. http://arxiv.org/abs/2006.11477.

Boersma, P., & Weenink, D. (2024). *Praat: Doing phonetics by computer* (Version 6.4.16) [Computer software]. http://www.praat.org/.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin, 114*(3), 542–551. https://doi.org/10.1037/0033-2909.114.3.542

Chen, L., & Yoon, S.-Y. (2011). Detecting structural events for assessing non-native speech. *Proceedings of the Sixth Workshop on INULPBEA, 11*, 38–45.

Chen, L., & Yoon, S.-Y. (2012). Application of structural events detected on ASR outputs for automated speaking assessment. *Proceedings of Interspeech, 2012*, 767–770. http://www.isca-speech.org/archive.

Chen, L., Zechner, K., Yoon, S. Y., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, C. M., Ma, M., Mundkowsky, R., Lu, C., Leong, C. W., & Gyawali, B. (2018). Automated scoring of nonnative speech using the SpeechRaterSM v. 5.0 engine. *ETS Research Report Series, 2018*(1). https://doi.org/10.1002/ets2.12198

Coulange, S., Kato, T., Rossato, S., & Masperi, M. (2024). Enhancing language learners' comprehensibility through automated analysis of pause positions and syllable prominence. *Languages, 9*(3), 78. https://doi.org/10.3390/languages9030078

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment—Companion volume*. https://doi.org/10.1002/9781118784235.eelt0114.pub2.

de Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *IRAL - International Review of Applied Linguistics in Language Teaching, 54*(2), 113–132. https://doi.org/10.1515/iral-2016-9993

de Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. *Proceedings of DiSS, 2013*, 17–20.

de Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics, 36*(2), 223–243. https://doi.org/10.1017/S0142716413000210

de Jong, N. H., Pacilly, J., & Heeren, W. (2021). PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically. *Assessment in Education: Principles, Policy and Practice, 28*(4), 456–476. https://doi.org/10.1080/0969594X.2021.1951162

de Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*(1), 5–34. https://doi.org/10.1017/S0272263111000489

de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*(2), 385–390. https://doi.org/10.3758/BRM.41.2.385

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv. 10.48550/arXiv.1810.04805.

Donner, A., & Eliasziw, M. (1987). Sample size requirements for reliability studies. *Statistics in Medicine, 6*(4), 441–448. https://doi.org/10.1002/sim.4780060404

Feng, S., Halpern, B. M., Kudina, O., & Scharenborg, O. (2024). Towards inclusive automatic speech recognition. *Computer Speech & Language, 84*, Article 101567. https://doi.org/10.1016/j.csl.2023.101567

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly, 15*(3), 219–236. https://doi.org/10.1080/15434303.2018.1453816

Graves, A., Fernandez, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd International Conference on Machine Learning*, 369–376. https://doi.org/10.1145/1143844.1143891

Handley, Z. L., & Wang, H. (2023). What do the measures of utterance fluency employed in automatic speech evaluation (ASE) tell us about oral proficiency? *Language Assessment Quarterly, 0*(0), 1–30. https://doi.org/10.1080/15434303.2023.2283839

Knill, K. M., Gales, M. J. F., Manakul, P. P., & Caines, A. P. (2019). Automatic grammatical error detection of non-native spoken learner english. *Proceedings of ICASSP, 2019*, 8127–8131. https://doi.org/10.1109/ICASSP.2019.8683080

Knill, K. M., Gales, M., Kyriakopoulos, K., Malinin, A., Ragni, A., Wang, Y., & Caines, A. (2018). Impact of ASR performance on free speaking language assessment. *Proceedings of Interspeech, 2018*, 1641–1645. https://doi.org/10.21437/Interspeech.2018-1312

Kormos, J. (2006). *Speech production and second language acquisition (pp. xxvii, 221)*. Lawrence Erlbaum Associates Publishers.

Kyle, K., Eguchi, M., Miller, A., & Sither, T. (2022). A dependency treebank of spoken second language english. *Proceedings of the 17th Workshop on IUNLPBEA '22,* 39–45. https://doi.org/10.18653/v1/2022.bea-1.7.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Linacre, J. M. (1994). *Many facet rasch measurement* (2 ed.). MESA Press.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv. https://doi.org/10.48550/arXiv.1907.11692.

Lu, Y., Gales, M., Fondazione, S. B., & Kessler, B. (2022). On Assessing and developing spoken "grammatical error correction. *Systems*, 51–60.

Matsuura, R., Suzuki, S., Saeki, M., Ogawa, T., & Matsuyama, Y. (2022). Refinement of utterance fluency feature extraction and automated scoring of L2 oral fluency with dialogic features. In *Proceedings of 2022 APSIPA ASC* (pp. 1312–1320). https://doi.org/10.23919/APSIPAASC55919.2022.9980148

Peltonen, P. (2017a). L2 fluency in spoken interaction: A case study on the use of other-repetitions and collaborative completions. *Näkökulmia Toisen Kielen Puheeseen. Insights into Second Language Speech, 10*, 118–138. https://doi.org/10.30660/afinla.73130

Peltonen, P. (2017b). Temporal fluency and problem-solving in interaction: An exploratory study of fluency resources in L2 dialogue. *System, 70*, 1–13. https://doi.org/10.1016/j.system.2017.08.009

Peltonen, P. (2021). Connections between measured and assessed fluency in L2 peer interaction: A problem-solving perspective. *IRAL - International Review of Applied Linguistics in Language Teaching, 60*(4), 983–1011. https://doi.org/10.1515/iral-2020-0030

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in l2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision*. arXiv. http://arxiv.org/abs/2212.04356.

Rose, R. L. (2020). Fluidity: Real-time feedback on acoustic measures of second language speech fluency. *Proceedings of the International Conference on Speech Prosody, 2020*, 774–778. https://doi.org/10.21437/SpeechProsody.2020-158

Saeki, M., Demkow, W., Kobayashi, T., & Matsuyama, Y. (2022). A woz study for an incremental proficiency scoring interview agent eliciting ratable samples. *Conversational AI for Natural Human-Centric Interaction. Lecture Notes in Electrical Engineering, 943*. https://doi.org/10.1007/978-981-19-5538-9_13

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv. https://doi.org/10.48550/arXiv.1910.01108.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge. https://doi.org/10.4324/9780203851357

Skidmore, L., & Moore, R. K. (2023). BERT models for spoken learner english disfluency detection. *Proceedings of SLaTE, 2023*, 91–92.

Suzuki, S., & Kormos, J. (2023). The multidimensionality of second language oral fluency: Interfacing cognitive fluency and utterance fluency. *Studies in Second Language Acquisition, 45*(1), 38–64. https://doi.org/10.1017/S0272263121000899

Suzuki, S., Kormos, J., & Uchihara, T. (2021). The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *Modern Language Journal, 105*(2), 435–463. https://doi.org/10.1111/modl.12706

Suzuki, S., & Révész, A. (2023). Measuring speaking and writing fluency: A methodological synthesis focusing on automaticity. In Y. Suzuki (Ed.), *Practice and automatization in second language research* (pp. 247–266). Routledge.

Suzuki, Y., & Hanzawa, K. (2022). Massed task repetition is a double-edged sword for fluency development: An EFL classroom study. *Studies in Second Language Acquisition, 44*(2), 536–561. https://doi.org/10.1017/S0272263121000358

Takizawa, K. (2024). What contributes to fluent L2 speech? Examining cognitive and utterance fluency link with underlying L2 collocational processing speed and accuracy. *Applied Psycholinguistics*, 1–26. https://doi.org/10.1017/S014271642400016X

Tao, J., Evanini, K., & Wang, X. (2014). The influence of automatic speech recognition accuracy on the performance of an automated speech assessment system. In *Proceedings of 2014 SLT* (pp. 294–299). https://doi.org/10.1109/SLT.2014.7078590

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *IRAL - International Review of Applied Linguistics in Language Teaching, 54*(2), 133–150. https://doi.org/10.1515/iral-2016-9994

Tavakoli, P., Kendon, G., Mazhurnaya, S., & Ziomek, A. (2023). Assessment of fluency in the test of english for educational purposes. *Language Testing, 40*(3), 607–629. https://doi.org/10.1177/02655322231151384

Tavakoli, P., Nakatsuhara, F., & Hunter, A. (2020). Aspects of fluency across assessed levels of speaking proficiency. *Modern Language Journal, 104*(1), 169–191. https://doi.org/10.1111/modl.12620

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–273). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.11.15tav.

Zayats, V., Tran, T., Wright, R., Mansfield, C., & Ostendorf, M. (2019). Disfluencies and human speech transcription errors. *Proceedings of Interspeech, 2019*, 3088–3092. https://doi.org/10.21437/Interspeech.2019-3134