Refinement of Utterance Fluency Feature Extraction and Automated Scoring of L2 Oral Fluency with Dialogic Features

Ryuki Matsuura* Shungo Suzuki* Mao Saeki* Tetsuji Ogawa* and Yoichi Matsuyama*
* Waseda university, Tokyo, Japan
E-mail: {matsuura, ssuzuki, saeki, ogawa, matsuyama}@pcl.cs.waseda.ac.jp

***Abstract*—We propose an automated scoring method of fluency that is compatible with second language dialogic responses. Human judgements of L2 oral fluency in dialogue tasks has different nature from scoring of monologue, and it is necessary to capture a dialogic aspect of fluency. Because utterances in dialogue tend to be less fluent than in monologue, procedures such as pruning disfluency words and classifying pauses by their syntactic locations are essential for automated scoring systems to extract utterance features strongly correlated to human ratings. However, existing automated utterance feature extractors have suffered from the difficulties to detect disfluency words and pauses locations due to the technical challenges. Moreover, conventional automated scoring methods of L2 spoken dialogue often predict oral proficiency for each turn, and dialogic features have not been considered properly. To address these gaps between the nature of fluency in dialogue and existing automated scoring methodologies, we refine an automated utterance feature extractor and design a fluency scoring model based on dialogic features. Experiments showed that the substantial agreement of disfluency word and pause location detection between our feature extractor and human (Cohen's $\kappa > 0.61$). We also found that the proposed scoring method outperformed in predicting subjective fluency scores (QW $\kappa = 0.833$) than a conventional turn-level scoring model (0.654) and even a manual rating (0.799). We additionally compared the current assessment approach considering disfluency features and pause location, and it improved the accuracy on predicting subjective fluency scores. These results may suggest what and how utterance and dialogic features should be utilized in automated scoring of spoken dialogue.**

I. Introduction

Fluency is one of the most important oral skills for successful communication in a second language (L2) because of its strong correlation with the ease of understanding L2 utterances for listeners [1]. It has also been said that fluent speech can hold the listeners' attention, and thus an optimal level of fluency is essential for L2 learners to make themselves understood in conversations [2]. From the perspective of effective L2 skill learning, it is crucial for learners to evaluate their current status of proficiency so that they can set realistic learning goals with regard to readiness and pedagogical demands. Motivated by such pedagogical needs, scholars have thus attempted to establish valid methods to assess L2 oral fluency and have developed automated fluency scoring technologies for L2 speech [3], [4], [5]. However, previous studies have focused on automated scoring of L2 oral proficiency in spoken monologues. Although acoustic characteristics of monologic speech are different from those of dialogic one [6], scholars have suffered from the lack of knowledge of valid operationalization of dialogic fluency features and the technologies capturing those features. To extend L2 research and pedagogies of L2 fluency, therefore, it is expected to develop an automated scoring system compatible with spoken dialogue with the better understanding of how different features contribute to human ratings of fluency in dialogues.

In the literature on second language acquisition (SLA), L2 oral fluency in utterance level has traditionally been assumed to have three distinctive aspects: speed of speech, pausing behavior, and disfluency phenomena (e.g., self-correction and repetition) [7]. As for a speed-related feature, one of the most common measures is articulation rate, which is calculated as the number of syllables per speech duration in seconds excluding pauses. Meanwhile, pausing behavior is typically analyzed in terms of the frequency and duration of silent pauses and also has been found to consistently predict learners' fluency levels. A feature of disfluency phenomena is also captured by the frequency of self-correction and repetitions in an utterance. In addition to those three aspects of utterance fluency, it has been found, albeit only recently, that the smoothness of interactions can make a unique contribution to human judgments of L2 oral fluency in dialogue [6]. Different studies have tried to extract features such a dialogic aspect of fluency, including the number of turns, the duration of silent intervals between a speaker and an interlocutor, and the number of repeating words delivered by an interlocutor [8], [9]. Due to the limited understanding of the role of dialogic features in fluency assessment, only few studies have examined the contribution of dialogic fluency features in the application to the automated scoring systems of fluency.

SLA research has identified speed, pause, and disfluency features that are associated with human ratings of fluency. Nevertheless, many existing automated scoring methods of fluency have failed to fully apply those findings. Firstly, the existing automated scoring systems often calculate utterance features such as articulation rate based on the number of syllables including repeated and corrected words (i.e., disfluency words), but it may lead to reduce the predictive power of fluency levels [10]. Moreover, the detection of disfluency words plays a vital role particularly in dialogic utterances, where learners need to spontaneously respond to the interlocutor's utterances, compared to extended monologues [11]. Secondly, pause features are rarely calculated without considering their locations (i.e., within a clause or between clauses) to score L2 fluency automatically, although pauses within the clause are more strongly correlated with the human judgments of L2 fluency than those at clausal boundaries both in monologues [7] and dialogues [8]. Finally, conventional automated fluency scoring systems of spoken dialogue have only considered the utterance features. In dialogue, several turns can consist of only a few words (e.g., "exactly", "I agree with that"), and such short turns may not provide sufficient information to judge learners' fluency level [12]. Therefore, it is required to capture the dialogue-level tendencies of speed, pause, and disfluency characteristics by computing the statistics of utterance features. Furthermore, the smoothness of dialogue was not explicitly used for the automated fluency scoring.

Motivated by the gaps between the findings in the field of SLA and existing automated scoring systems of fluency, current study aims to refine the extraction method of utterance-level speed, pause, and disfluency features and design the dialogue-level features to automatically predict the L2 oral fluency of spoken dialogue. In extracting utterance features, we adopted the procedures of pruning disfluency words and classifying pause locations (i.e., classifying pauses between mid-clause and end-clause). We used an oral proficiency interview as the speech elicitation task, which required learners to produce highly spontaneous speech. Both pruning disfluency words and classifying pause locations are thus expected to play an essential role in the prediction of fluency level. Moreover, the extraction of dialogue-level features are achieved by statistics pooling [13]. In the experiments, we preliminary explore how accurately our extractor detects disfluency words to be pruned and classify pause locations in interview responses by comparing with human annotations. We also examine the performance of dialogic features in L2 fluency judgement experiments using our proposed model, a conventional turn-level prediction model, and a manual rating by in-service language teachers.

The contributions of this study are as follows: (i) we develop first utterance-level speed, pause, and disfluency feature extractor incorporating both pruning of disfluency words and classification of pause locations, (ii) we design the dialogue-level fluency features to automatically predict L2 oral fluency of spoken dialogue.

A rest of the paper is organized as follows. Section II reviews previous work on the relationship between utterance fluency features and the human judgments of fluency, existing utterance feature extraction methods, and automated scoring systems for dialogic responses. Section III describes the proposed refinement of the speed, breakdown, and repair feature extractor and the current automated L2 oral fluency scoring model based on dialogic features. Section IV reports on the first experiment which evaluates the accuracy of our proposed feature extraction methods. Section V explains the results of the second experiment which compared our L2 fluency scoring model with a conventional turn-level prediction model and a manual rating. Finally, section VI concludes the paper by highlighting the contributions of the study and future directions for the automated assessment of L2 fluency.

II. Related work

A. Predicting subjective judgements of L2 oral fluency

To better understand the mechanisms of human-based evaluation of L2 oral fluency, previous studies have extensively examined how and what temporal speech characteristics contribute to subjective judgements of fluency [14]. There is a methodological consensus that oral fluency has three sub-dimensions—speed fluency (i.e., speed of delivery), breakdown fluency (i.e., frequency and duration of pauses), and repair fluency (i.e., disfluency phenomena such as repetition) [15], [16], and thus temporal predictors should cover these three dimensions [16]. A recent meta-analysis demonstrated that the predictive power of temporal features in subjective judgements of fluency can vary across those sub-dimensions of fluency [7]. Speed fluency and pause frequency (breakdown fluency) were strongly associated with fluency judgement scores ($r = 0.62$, 0.59, respectively), whereas pause duration (breakdown fluency) was moderately linked to fluency ratings ($r = 0.46$). Meanwhile, repair fluency measures were weakly but significantly related to fluency judgements ($r = 0.20$). Recently, researchers have also admitted that the nature of L2 fluency can differ between monologic and dialogic speaking tasks [6]. For example, utterance fluency measures can vary depending on the turn and topics in dialogue, even though they are calculated from same speakers' utterances [11], [17]. Given the distributions of turns affect temporal features of speech, it is thus necessary to capture the distributions of speed, breakdown, and repair fluency features across turns for predicting fluency ratings. Furthermore, several dialogue-specific fluency features, such as pauses between turns, have found to contribute to listener-based judgements of fluency in dialogic speaking tasks (e.g., paired discussion tasks) [9]. These findings, therefore, suggest that the automated scoring of L2 oral fluency should include a whole range of fluency features covering speed, breakdown, and repair fluency and dialogic fluency.

B. Automated extraction of fluency features

To extract aforementioned fluency features, existing fluency assessment tools count the number of syllables produced in speech either with or without transcription data of the speech. De Jong and Wempe proposed a method to detect the number of syllables and silent intervals only based on acoustic features of speech such as the intensity peaks and contours of the speech signals [18]. The correlation coefficient between the number of syllables counted manually and automatically achieved $r = 0.71$ in L2 English. Moreover, filled pauses are also detected with the assistance of the pitch information in the recent methods [19], [20]. The accuracy score of the filled pause detection exceeded 80% on several dataset of English speech produced by English speakers (L1) and Dutch speakers (L2) [19]. However, De Jong and colleagues admitted that in the context of L2 speech, these methods without written transcription of speech may not necessarily be applicable to other groups of L2 learners with different L1 backgrounds, due to the influence of L1 accents on English syllable structures [19]. In addition, despite the promising performance of detecting syllables and filled pauses, the lack of written transcription may threaten the validity of other fluency features. For instance, it is highly challenging to detect disfluency words, which is essential information for repair fluency measures. Similarly, the method may not classify silent pauses by syntactic locations (i.e., within or between clauses), though silent pauses can have different predictive power for fluency judgements according to their locations [6], [7], [8]. Chen and Yoon, then, proposed a detector of *structural events* [21] consisting of clause boundaries and the startpoints of reparandums, that is, word sequences repaired by a speaker, for their fluency feature extraction [22]. They examined the performance of structural event detector on the automatically transcribed speech using automated speech recognition (ASR) system. The results showed that F1 scores for detecting clause boundaries and reparandum startpoints were 0.690 and 0.304 respectively [23]. Although these studies extracted various features related to the triad of the utterance fluency, the exclusion of disfluency words (i.e., pruning) has not been sufficiently considered because structural event detection has not identified words themselves which compose reparandums. Without pruning disfluency words, the number of syllables can unfairly increase by repeating or correcting spoken utterances [10], subsequently lowering the validity of utterance fluency feature calculation. Therefore, to develop a valid automated scoring system for fluency assessment, our utterance fluency feature extractor incorporates not only the detection of pause locations with regard to clause boundaries but also that of disfluency words.

C. Automated Scoring for Dialogue Tasks

In the literature of automated L2 speech assessment, researchers have commonly regarded dialogue speech as a collection of individual turns and thus have aimed to predict the score of individual turns. Ref. [12] proposed an automated scoring method to predict the scores of fluency, pronunciation and intonation for individual turns. However, they found that it was difficult to predict the scores of those speaking abilities for each turn because dialogue usually contains short utterances which does not have sufficient information for the scoring. Qian and colleagues also predicted the overall score of L2 speech turn by turn. They adopted End-to-End Memory Network (MemE2E) [24] to consider the histories of dialogue, showing its effectiveness for the turn-level score prediction [25]. These turn-level scoring approaches have also applied to score the entire dialogue holistically. For instance, [26] assigned the score of the whole dialogue to its individual turns and predicted the score of individual turns from a set of linguistic features. Importantly, they then regarded the median of all the predicted by-turn scores as the score of the whole dialogue. This approach, albeit allowing for data augmentation [27], may fail to reflect the nature of L2 dialogic speech characteristics and thus can lower the predictive power of the assessment systems. From the fluency rating perspective, the entire dialogue score should be predicted using both statistics of the utterance-level features (i.e., speed, breakdown, and repair fluency) and dialogic fluency features. We, therefore, propose and evaluate the use of the statistics pooling to predict the fluency score of the entire dialogue by capturing the variability of temporal and dialogic features across turns.

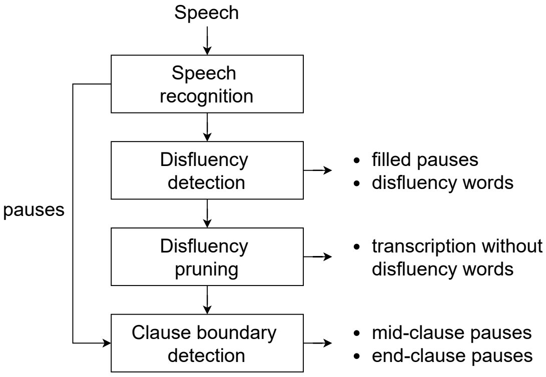


Fig. 1: Detection steps of pauses and disfluency words.

III. Automated fluency scoring method of L2 spoken dialogue

In this section, we describe our refinement of the extraction method of utterance features (i.e., speed, breakdown, and repair fluency features) which incorporates the disfluency word pruning and the pause location classification. Moreover, we propose a method to predict fluency scores tailored to dialogic speech, considering the statistics of utterance-level information and the features that reflect the smoothness of dialogue.

A. Refinement of utterance fluency feature extraction

The proposed utterance fluency feature extraction consists of an automated detector of silent pauses and disfluency words and an extractor of features related to speed, breakdown, and repair fluency for each turn.

1) Automated detection of pauses and disfluency words: We followed the detection method of [5]. Figure 1 depicts how pauses and disfluency words are detected in this method. This method consists of four phases: speech recognition, disfluency detection, disfluency pruning, and clause boundary detection. The phase of speech recognition predicts appropriate word sequences from learners' utterances. We use Asynchronous Speech-to-Text provided by Rev.ai[1], producing the time-aligned transcription for individual words. We thus adopt the time-aligned information to detect silent pauses. According to [7], [28], silences longer than 250ms tend to have strong predictive power for human-based fluency judgments. The current

[1] https://www.rev.ai/async

TABLE I: List of speed, breakdown, repair and dialogic fluency features.

Type	Feature	Description
Speed fluency	Articulation rate	Number of syllables per speech duration excluding pauses.
Breakdown fluency	Mid-clause pause ratio	Number of mid-clause pauses per syllables.
	End-clause pause ratio	Number of end-clause pauses per syllables.
	Filled pause ratio	Number of filled pauses per syllables.
	Mid-clause pause duration	Mean duration of mid-clause pauses.
	End-clause pause duration	Mean duration of end-clause pauses.
Repair fluency	Disfluency ratio	Number of disfluency words per syllables.
Dialogic fluency	Number of between-turn pauses	Number of between-turn pauses divided equally among participants.
	Between-turn pause duration	Mean duration of between-turn pauses.
	Number of turns	Number of turns.
	Mean length of turns	Number of syllables divided by number of turns.
	Number of other repetitions	Number of repeated words of interlocutors' speech.

method follows this threshold for silent pauses. Disfluency detection and disfluency pruning identify and remove disfluency phenomena, such as filled pauses (e.g., ”ah,” ”huh”), repetitions, and self-corrections. More specifically, filled pauses are first detected and removed from the transcriptions generated in the speech recognition phase. Disfluency words are then detected in the transcriptions without filled pauses by BERT [29] based model. This model, which is built by fine-tuning BERT using Switchboard reannotated dataset [30], classifies whether each token belongs to a reparandum or not. Tokens predicted as disfluency words are then removed to improve the accuracy of subsequent processing. Finally, clause boundary detection, which is needed for distinguishing pause locations (mid-clause pause (MCP) vs. end-clause pauses (ECP)), is achieved by analyzing the dependencies of the primary and subordinate clauses using the parser of Stanford CoreNLP [31].

2) Automated extraction of utterance fluency features: Using the written transcriptions, pauses, and disfluency words, we calculated fluency features related to speed, breakdown, and repair fluency in each turn. Extracted features are listed in Table I.

Speed fluency feature: As a speed fluency feature, we calculated articulation rate (AR). Let N_{syl} be the number of syllables, $p \in P$ be pauses in an utterance u. AR is then calculated as:

$$\text{AR} = \frac{N_{syl}}{d(u) - \sum_{p \in P} d(p)}, \tag{1}$$

where $d(\cdot)$ is the duration in seconds. Note that the number of syllables is counted from transcriptions after disfluency words are pruned [10]. If the disfluency words are not pruned, the number of syllables could be infinitely increased by repetitions and corrected words, resulting in unfairly high values for AR. For instance, the repetition of the same word multiple times quickly increases the number of syllables N_{syl} and subsequently increase AR, even though such an utterance should be considered less fluent.

Breakdown fluency feature: The features related to silent pauses should be distinguished by their syntactic locations because of the strong predictive power of MCPs for the human judgments of L2 oral fluency [7]. Accordingly, we computed pause ratio and mean pause duration separately for MCPs and ECPs. Mid-clause pause ratio (MCPR) is calculated by (2) and mid-clause pause duration (MCPD) is by (3)

$$\text{MCPR} = \frac{|P_{mc}|}{N_{syl}}, \tag{2}$$

$$\text{MCPD} = \frac{\sum_{p_{mc} \in P_{mc}} d(p_{mc})}{|P_{mc}|}, \tag{3}$$

$$\tag{4}$$

where $p_{mc} \in P_{mc}$ is MCP and $|P_{mc}|$ is the number of MCPs. In the same way, end-clause pause ratio (ECPR) and end-clause pause duration (ECPD) were computed. In addition, we also computed filled pause ratio (FPR). Let N_{fp} be the number of filled pauses. FPR is then obtained by

$$\text{FPR} = \frac{N_{fp}}{N_{syl}} \; . \tag{5}$$

For the calculation of pause frequency and duration measures, we use N_{syl} after pruning disfluency words for the same reason as the speed fluency feature.

Repair fluency feature: We employ the disfluency ratio (DR) as a feature reflective of the repair fluency. Using pruned N_{syl}, DR is calculated as:

$$\text{DR} = \frac{N_r}{N_{syl}}, \tag{6}$$

where N_r is the number of reparandums.

B. Automated scoring of L2 oral fluency using dialogic features

We propose the automated scoring model with the statistics pooling to obtain entire dialogue-level features across turns, as visualized in Figure 2. We incorporate an attention mechanism into the statistics pooling [32] to avoid equally weighting all turns because short turns can have confounding information to score fluency levels. The statistics pooling provides the entire dialogue-level features by calculating the mean and standard deviation of the aforementioned turn-level features (as indicated as 1 to 7 in Figure 2). Moreover, the statistics pooling model can include the following dialogic fluency features (see Table I) that characterize the smoothness of interactions: number of turns (NT), mean length of turn (MLT), number of between-turn pauses (NTP), between-turn pause duration

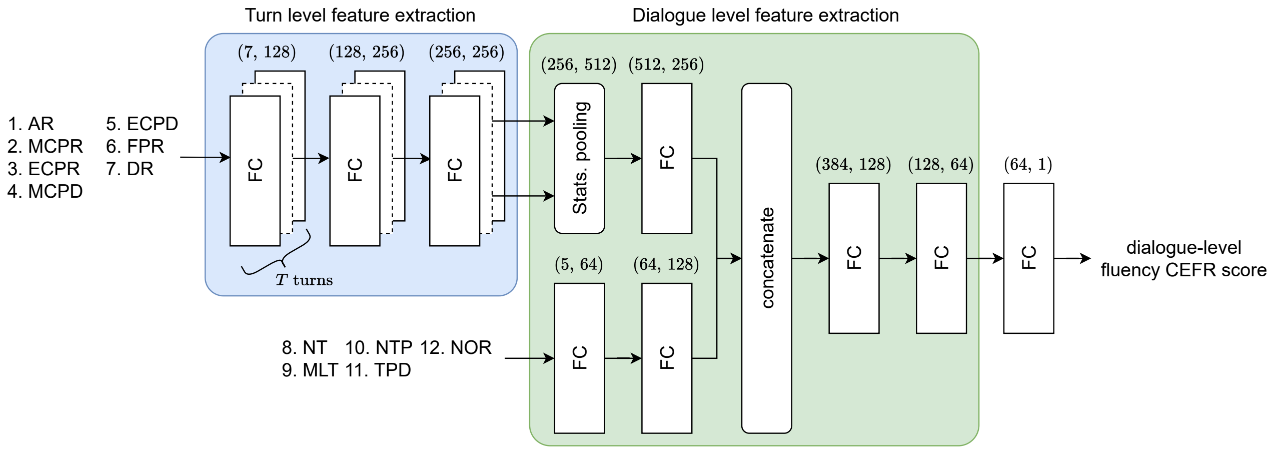


Fig. 2: Architecture of automated L2 oral fluency scoring model using statistics pooling.

(TPD), and number of other-repetition (NOR) (indicated as 8 to 12 in Figure 2). Here, between-turn pauses refer to the pauses between the speaker's utterance and the interlocutor's following one. These features are fed into fully-connected (FC) layers and concatenated with the pooling layer's output. This concatenated feature vector is then weighted by FC layers. The model predicts the fluency score of the entire dialogue. This model also avoids overfitting by dropout ($p = 0.05$).

IV. Evaluation of pause and disfluency word detection

Using the human-agent interaction data [33], the first experiment aims to test the accuracy of detecting pauses and disfluency words needed for scoring L2 oral fluency, compared with the corresponding manual annotations. Target fluency phenomena were automatically detected by the method described in section III.

A. Human-agent interaction data

In order to collect authentic human-agent interaction data, our precursor study developed the Wizard of Oz (WoZ) system [34]. In the WoZ system, trained human interviewers operate the utterances and motions of a conversational agent so that users can have a pseudo-experience of interacting with the agent that works fully automatically. Eighty-five Japanese learners of English completed an interview task with the conversational agent with the assistance of the WoZ system. The whole procedure of the interview task followed the guideline of the ACTFL Oral Proficiency Interview (OPI) [35]. Each interview consisted of seven different topics, while each topic was adaptively selected according to the user's immediate performance. The interaction between users and the agent lasted approximately nine minutes (*Range* = 4.49–16.28)

B. Manual annotation of pauses and disfluency words

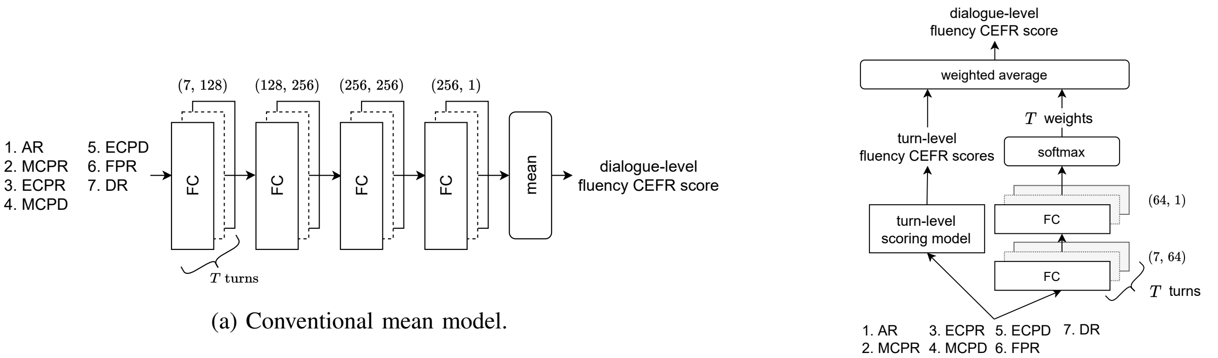
In order to evaluate the reliability of the automated utterance fluency feature extraction, we collected a manual annotation of disfluency phenomena as the ground truth. Following previous studies [6], [7], [8], it is necessary to annotate silent pauses and their syntactic locations, as well as disfluency words (filler word, self-correction, and repetition) for the calculation of fluency features (see Section III-A). All the interview speech samples were first manually transcribed. Two trained research assistants then annotated the speech data for the aforementioned fluency phenomena. In a one-hour training session, they received the explicit instruction on the goal of the project and fluency-relevant phenomena. After they confirmed the clear understanding of those phenomena, they analyzed responses to 35 topics from five interviews randomly selected from the current dataset to check the inter-coder reliability. The Cronbach's α indices suggested the high level of consistency between the two coders for silent pauses ($\alpha = 0.959$) and for disfluency words ($\alpha = 0.983$). Finally, the whole dataset was divided into two subsets, and the two coders were assigned to either of them. As a result, they annotated 7,621 MCPs, 2,446 ECPs and 8,562 disfluency words from the whole 85 interviews.

C. Results and Discussion

We evaluated our automated feature extraction method for pauses and disfluency words, using the indices of Cronbach's α, Cohen's κ, precision, recall, and F1 score, considering the current detection as a multiclass word classification. While the human coders annotated pauses and disfluency words on the manual transcription, automated detection was conducted on the transcription generated by ASR. The word error rate (WER) of ASR was 27.3%, which was more accurate than that reported in [36] (WER = 28.5%). To evaluate the reliability of the current feature extraction method while considering ASR errors, we used NIST's SCTK[2] to align the detection results by our method and the human coders [23].

As a result, 13,078 MCPs, 2,624 ECPs and 6,497 disfluency words were annotated automatically. The indices of Cronbach's α demonstrated that the high level of consistency between the human coders and our method both for silent pauses ($\alpha = 0.999$) and disfluency words ($\alpha = 0.999$), whereas those

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



(a) Conventional mean model.

(b) Conventional weighted average (WA) model.

Fig. 3: Architecture of turn-level L2 oral fluency scoring model for dialogue tasks.

TABLE II: Precision, recall and F1 score of pause and disfluency word detection between human and proposed method.

class	precision	recall	F1 score
MCP	0.519	0.891	0.656
ECP	0.508	0.545	0.526
no pause	0.985	0.868	0.923
disfluency word	0.832	0.631	0.718
otherwise	0.929	0.974	0.951

Note. MCP = mid-clause pause, ECP = end-clause pause

of Cohen's κ indicated the substantial agreement ($\kappa > 0.61$) between them for silent pauses ($\kappa = 0.613$) and for disfluency words ($\kappa = 0.674$). Precision, recall, and F1 score of pause and disfluency word detection are summarized in Table II. In terms of F1 score, the performance of disfluency word detection in the current study was higher than [23] (F1 $= 0.304$). These results confirmed that the current method for annotating pauses and disfluency words has substantial reliability and agreement with human annotators. However, care should be taken not to overestimate the current detection method. Firstly, MCP and ECP detection may have some room for improvement in accuracy (see Table II). Precision of MCP may have been affected by the large difference in the number between manual and automated detection. Due to the discrepancies in the time alignment of words, unnecessary pauses may have been falsely detected. Regarding the ECP detection, both precision and recall were relatively low. ASR errors may have led to an incorrect insertion of clause boundaries. As with MCP, the time alignment estimated by ASR may have also caused the inaccurate detection of ECPs. To solve these problems, future works should adopt forced alignment techniques to the transcription generated by ASR for the sake of accurate time alignment of individual words. Similarly, the dependency parser that is robust to speech recognition errors might also be expected to be developed in future studies.

V. Predicting oral fluency judgements

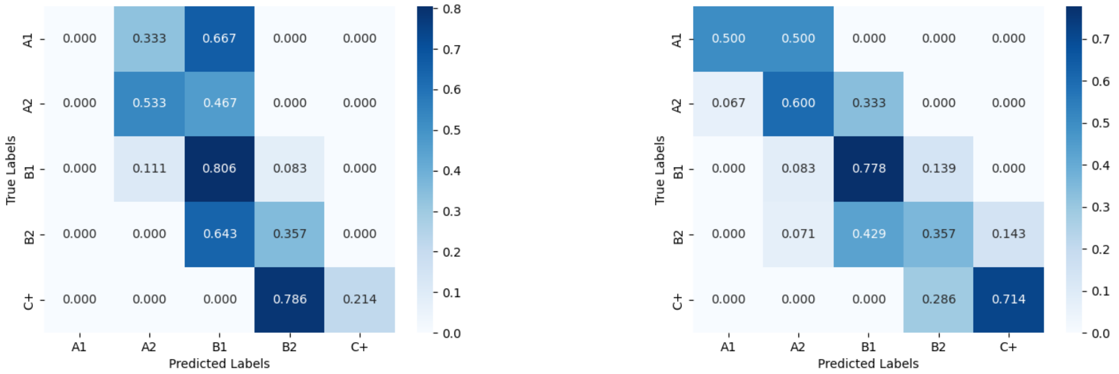
We evaluated the predictability of our proposed scoring model, compared with a conventional model. The conventional model produced the entire dialogue fluency score as the mean value of predicted fluency scores of individual turns [26], [27]. In contrast, our proposed model directly predicted the dialogue-level fluency score using dialogic features extracted with the statistics pooling. Both of conventional and proposed models were trained with the same dataset in the first experiment.

A. Human-based judgements of fluency

In order to train our fluency scoring model, the interview data were scored for fluency by an English-speaking in-service language instructor with more than 10 years of English teaching and assessment experience (henceforth, rater). We employed the Common European Framework of Reference for Language (CEFR) [37] as a scale for fluency. The CEFR scale has 6 levels: A1, A2, B1, B2, C1, and C2. A1 is the lowest level, and C2 represents the highest one. The number of learners in the each level of the CEFR scale was 6, 15, 36, 14, 12, 2, respectively. Due to the limited number of C2 level learners ($n = 2$), we merged C1 and C2 level learners into a single group of C+ level [34]. To measure the human-human agreement of fluency scoring, all 85 interviews were independently assessed by another English-speaking native teacher with more than 10 years of English teaching experience. As a result, inter-rater agreement was 0.832 in Pearson's r correlation coefficient and 0.799 in quadratic weighted (QW) κ, indicating the optimal level of agreement.

B. Turn-level scoring model

We developed a conventional turn-level scoring model using multi layer perceptron (MLP). In line with previous studies [6], [7], [8], we extracted the following seven utterance fluency features to train the model: AR, MCPR, ECPR, MCPD, ECPD, FPR, and DR (indicated as 1 to 7 in Figure 3). We assigned the fluency score of a whole interview to all of its individual turns and trained the model at the turn-level, following previous studies [26], [27]. The predicted scores of individual turns were aggregated using statistical calculation to obtain the entire dialogue score. We compared mode, median, and mean as the calculation method of the dialogue-level score and found that mean provided best performance (conventional mean model) in the current dataset. Additionally, we used the weighted average of turn-level fluency scores as the entire dialogue score because the characteristics of speed, breakdown, and



(a) Conventional model (WA).

(b) Proposed model.

Fig. 4: Confusion matrix of fluency CEFR prediction by conventional and proposed model.

TABLE III: Pearson's r correlation coefficient and quadratic weighted κ of CEFR fluency predicted by conventional model, statistics pooling model, and human rater.

model	Pearson's r	QW κ
conventional (mean)	0.709	0.503
conventional (WA)	0.721	0.654
stats. pooling	**0.836**	**0.833**
human-human	0.834	0.799

repair fluency vary for each utterance in dialogue. The weights $\boldsymbol{w} = \{w_1, \ldots, w_T\}$ for T turns were computed by MLP with parameter θ and utterance fluency features, analogous to the attention mechanism (conventional WA model). Here, we optimized the parameter θ so that the mean squared error (MSE) between the weighted average of turn-level fluency scores and the entire dialogue score is minimized. The model consisted of four FC layers, and weights are output from two FC layers and a softmax layer as shown in Figure 3. We employed dropout ($p = 0.2$) for the regularization.

C. Results and Discussion

We evaluated the proposed model in terms of the agreement between human and automated scoring. The agreement was measured by 5-fold cross-validation with Pearson's r correlation coefficient and QW κ. We also examined an human-human agreement as the benchmark based on the extent to which the fluency scores judged by human raters can accord. We used MSE for loss function and Adam [38] for optimization of parameters. All utterance fluency features for training those scoring models were extracted by the proposed method in Section III.

The agreements of predicted fluency scores of the entire interview by the conventional models and the statistics pooling model are summarized in Table III and confusion matrices are shown in Figure 4. The statistics pooling model achieved the higher alignment with the manual annotation than the conventional model. In addition, both Pearson's r and QW κ of the proposed model also outperformed that of the human-human agreement. Comparing two conventional models, it was found that the agreement of WA model is higher than the one aggregated with mean. We assumed that the conventional WA model might be able to successfully account for differences in utterance features across turns and topics. Ref. [39] reported that the easier topic difficulties, the more fluently L2 learners speak. Taken together, it may be inappropriate to treat all turns equally in the fluency evaluation of spoken dialogue. We thus considered that the conventional WA model is more valid baseline than the mean model. We compared the conventional WA model and our proposed model. We found that the inclusion of dialogic-specific features, such as statistics of each turn and dialogic fluency, improved the predictive power for L2 oral fluency levels of spoken dialogue. Especially the predictability for beginner levels (A1 and A2 levels on the CEFR scale) was lower in the conventional model. Since it is known that utterances by beginner level learners tend to consist of a few words [40], it might be difficult to rate their fluency as A1 or A2 level for each turn [12]. In contrast, our proposed model might enable to extract proper features that could distinguish the novice levels incorporating dialogue-level statistics of utterance fluency features. As we compared the statistics pooling model with human ratings, the performance of our model was higher than that of human. Therefore, our model might be expected to be used as a substitute for human judgment for L2 speaking ability. However, the rates of fluency scores which our proposed model correctly predicted were 0.500, 0.600, 0.778, 0.357, 0.714 for each CEFR level, respectively. This indicated that the prediction accuracy was not so high, especially for A1, A2, and B2. One possible explanation for it might lie in the inability of the current scoring model to utilize the topic difficulties explicitly. The interview used in the experiment was designed to ask questions at the +1 difficulty level, depending on the learners' immediate performance to probe the upper limit of their ability [34]. For example, while B2-level learners might speech less fluent with lots of pauses and repeating and correcting words in advanced

TABLE IV: Pearson's r correlation coefficient and quadratic weighted κ between ground truth and predicted fluency score with and without refinement of utterance fluency feature extraction.

method	**pruning**	**pause loc.**	**Pearon's r**	**QW κ**
(a)			0.817	0.813
(b)		✓	0.806	0.801
(c)	✓		0.819	0.818
(d)	✓	✓	**0.836**	**0.833**

Note. pruning = disfluency word pruning,
pause loc. = pause location classification

C+ level topics, the proposed model scored fluency without explicitly considering what difficulty topics those utterances were observed from. In the future work, we could investigate the automated fluency scoring system of spoken dialogue utilizing topic difficulty.

D. Ablation study

To evaluate the refinement of utterance fluency feature extraction with the disfluency word pruning and the pause location classification, we conducted a follow-up ablation study. Given that dialogic speaking tasks such as an interview require speakers to spontaneously respond to the interlocutor's questions [11], the disfluency word pruning is essential to extract utterance-level speed, breakdown, and repair features, especially in dialogue tasks. Meanwhile, regarding pausing behavior, the necessity of pause location classification is promising due to the strong correlation of MCP-based features with human judgements of L2 oral fluency [7]. Therefore, we compared the performance of automated fluency scoring in terms of the agreement with the ground truth across the following four conditions;

(a) without both of the disfluency word pruning and pause location classification (corresponding to [19], [18], [20]),
(b) with only the disfluency word pruning,
(c) with only the pause location classification (corresponding to [23]),
(d) with both of the disfluency word pruning and pause location classification.

Note that, in the methods (a) and (b), pause ratio and mean pause duration were computed instead of MCPR, ECPR, MCPD, and ECPD, due to the lack of the function of pause location classifications. On the other hand, the methods (a) and (c) calculated all fluency features without pruning of disfluency words, and thus those measures were computed based on the number of syllables produced including disfluency words. Other experimental conditions of the methods (a)–(c) are identical to the proposed statistical pooling model.

Table IV shows Perason's correlation coefficients and QW κ between four different conditions and the ground truth scores. The results suggested that the performance of fluency score predictions was highest with the method (d), confirming the effectiveness of the disfluency word pruning and pause location classification. In addition, the difference in both QW κ and Perason's r across the four methods indicates that the disfluency word pruning plays a meaningful role in predicting fluency scores. In detail, the performance of fluency score prediction tended to be high when the disfluency word pruning was added regardless of whether the pause location classification was integrated. Therefore, these findings suggest that disfluency words should be pruned for the utterance fluency feature extraction. Meanwhile, the effectiveness of pause location classification might be limited. Comparing the methods (c) and (d), the improvement of the agreement between predicted score and the ground truth was found (QW κ; 0.818 vs. 0.833). However, the performance of automated fluency scoring was degraded by adding the breakdown fluency features based on MCPs and ECPs when there was no disfluency word pruning (QW κ; 0.813 (a) vs. 0.801 (b)). Utterance length (i.e. the number of syllables) including reparandums reduced the discriminability of MCPR and ECPR, and that might have negative impact on automated scoring. In other words, the contributions of the disfluency word pruning and pause locatoion classification to the predictability of fluency scores might be interrelated. It is thus plausible to argue that the pause location classification should be applied to automated fluency scoring systems together with the disfluency word pruning.

VI. Conclusions

In this study, we developed and validated the automated scoring system of L2 fluency in dialogue speech, with the focus on the refinement of the utterance fluency feature extractor and the scoring model. Firstly, we evaluated the automated detection method of pause locations and disfluency words. The results showed that the agreement between human coders and our method were substantially high. Secondly, we compared the statistics pooling model for L2 fluency scoring of dialogue responses with a conventional model and a manual annotation. The results demonstrated that the statistics pooling model outperformed not only the conventional model but also the manual annotation in terms of the agreement with the ground truth. Finally, the ablation study confirmed that the effectiveness of refinement of speed, breakdown, and repair fluency feature extraction method, highlighting the interrelationship between the disfluency word pruning and the pause location classification.

Despite these contributions to the automated assessment of oral fluency, the current study suggests that the accuracy of pause location classification as well as the disfluency word detection could be further improved. One possible direction for future studies is thus to improve the automated detection of disfluency phenomena, especially pause locations. We will integrate some audio processing techniques such as forced alignment for more accurate silent pause detection. Moreover, the current automated L2 fluency scoring may have ignored the effects of topic difficulty on the predictability of fluency features for human ratings. This line of research could be extended by integrating speaking task design features into the automated scoring models explicitly.

VII. Acknowledgements

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).

References

[1] S. Suzuki and J. Kormos, “Linguistics Dimensions of Comprehensibility and Perceived FluencyL: An Investigation of Complexity, Accuracy, and Fluency in Second Language Argumentative Speech,” *Studies in Second Language Acquisition*, vol. 42, no. 1, p. 143–167, 2020.

[2] P. Lennon, “The Lexical Element in Spoken Second Language Fluency,” in *Perspectives on fluency*, H. Riggenbach, Ed. Ann Arbor: University of Michigan Press, 2000, pp. 25–42.

[3] S. Mao, Z. Wu, J. Jiang, P. Liu, and F. K. Soong, “NN-based Ordinal Regression for Assessing Fluency of ESL Speech,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7420–7424.

[4] Y. Shen, A. Yasukagawa, D. Saito, N. Minematsu, and K. Saito, “Optimized Prediction of Fluency of L2 English Based on Interpretable Network Using Quantity of Phonation and Quality of Pronunciation,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 698–704.

[5] R. Matsuura, S. Suzuki, M. Saeki, T. Ogawa, and Y. Matsuyama, “Automated Scoring of L2 Fluency Based on Detection of Disfluency Words and Pause Locations,” in *Proceeding of The 2022 Spring meeting of the Acoustical Society of Japan*, 2022, pp. 1351–1354.

[6] P. Tavakoli, “Fluency in Monologic and Dialogic Task Performance: Challenges in Defining and Measuring L2 Fluency,” *International Review of Applied Linguistics in Language Teaching*, vol. 54, no. 2, pp. 133–150, 2016.

[7] S. Suzuki, J. Kormos, and T. Uchihara, “The Relationship between Utterance and Perceived Fluency: A Meta-Analysis of Correlational Studies,” *The Modern Language Journal*, vol. 105, no. 2, pp. 435–463, 2021.

[8] P. Peltonen, “Temporal Fluency and Problem-Solving in Interaction: An Exploratory Study of Fluency Resources in L2 Dialogue,” *System*, vol. 70, pp. 1–13, 2017.

[9] ——, “Connections between Measured and Assessed Fluency in L2 Peer Interaction: A Problem-Solving Perspective,” *International Review of Applied Linguistics in Language Teaching*, pp. 1–29, 2021.

[10] R. Ellis and B. Gary, *Analysing Learner Language*, ser. Oxford applied linguistics. Oxford: Oxford University Press, 2005.

[11] P. Tavakoli and C. Wright, *Second Language Speech Fluency: From Research to Practice*. Cambridge University Press, 2020.

[12] V. Ramanarayanan, P. L. Lange, K. Evanini, H. R. Molloy, and D. Suendermann-Oeft, “Human and Automated Scoring of Fluency, Pronunciation and Intonation During Human–Machine Spoken Dialog Interactions,” in *Proc. Interspeech 2017*, 2017, pp. 1711–1715.

[13] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep Neural Network Embeddings for Text-Independent Speaker Verification,” in *Proc. Interspeech 2017*, 2017, pp. 999–1003.

[14] N. Segalowitz, *Cognitive Bases of Second Language Fluency*. London & New York: Routledge, 2010.

[15] P. Tavakoli and P. Skehan, “Strategic Planning, Task Structure, and Performance Testing,” in *Planning and task performance in a second language*, R. Ellis, Ed. Amsterdam: John Benjamins, 2005, pp. 239–273.

[16] S. Suzuki and J. Kormos, “The Multidimensionality of Second Language Oral Fluency: Interfacing Cognitive Fluency and Utterance Fluency,” *Studies in Second Language Acquisition*, 2022.

[17] B. Heather, L. Silvia, D., B. Jonathan, E., S. Michael, F., and B. Susan, E., “Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender,” *Language and Speech*, vol. 44, no. 2, pp. 123–147, 2001.

[18] N. H. De Jong and T. Wempe, “PRAAT Script to Detect Syllable Nuclei and Measure Speech Rate Automatically,” *Behavior Research Methods*, vol. 41, pp. 385–390, 2009.

[19] N. H. de Jong, J. Pacilly, and W. Heeren, “PRAAT Scripts to Measure Speed Fluency and Breakdown Fluency in Speech Automatically,” *Assessment in Education: Principles, Policy & Practice*, vol. 28, no. 4, pp. 456–476, 2021.

[20] R. Rose, “Fluidity: Real-time Feedback on Acoustic Measures of Second Language Speech Fluency,” in *Proc. Speech Prosody 2020*, 2020, pp. 774–778.

[21] L. Chen, J. Tetreault, and X. Xi, “Towards Using Structural Events to Assess Non-Native Speech,” in *IUNLPBEA ’10*. USA: Association for Computational Linguistics, 2010, p. 74–79.

[22] L. Chen and S.-Y. Yoon, “Detecting Structural Events for Assessing Non-Native Speech,” in *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*. Portland, Oregon: Association for Computational Linguistics, 2011, pp. 38–45.

[23] ——, “Application of Structural Events Detected on ASR Outputs for Automated Speaking Assessment,” in *Proc. Interspeech 2012*, 2012, pp. 767–770.

[24] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, “End-to-End Memory Networks,” in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, 2015, p. 2440–2448.

[25] Y. Qian *et al.*, “Neural Approaches to Automated Speech Scoring of Monologue and Dialogue Responses,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8112–8116.

[26] V. Ramanarayanan, M. Mulholland, and Y. Qian, “Scoring Interactional Aspects of Human-Machine Dialog for Language Learning and Assessment using Text Features,” in *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, 2019, pp. 103–109.

[27] M. Saeki, Y. Matsuyama, S. Kobashikawa, T. Ogawa, and T. Kobayashi, “Analysis of Multimodal Features for Speaking Proficiency Scoring in an Interview Dialogue,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 629–635.

[28] N. H. De Jong and H. R. Bosker, “Choosing a Threshold for Silent Pauses to Measure Second Language Fluency,” in *Proceeding of The 6th Workshop on Disfluency in Spontaneous Speech*, 2013, pp. 17–20.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[30] V. Zayats, T. Tran, R. Wright, C. Mansfield, and M. Ostendorf, “Disfluencies and Human Speech Transcription Errors,” in *Proc. Interspeech 2019*, 2019, pp. 3088–3092.

[31] C. D. Mannin *et al.*, “The Stanford CoreNLP Natural Language Processing Toolkit,” in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60.

[32] K. Okabe, T. Koshinaka, and K. Shinoda, “Attentive Statistics Pooling for Deep Speaker Embedding,” in *Proc. Interspeech 2018*, 2018, pp. 2252–2256.

[33] M. Saeki *et al.*, “InteLLA: A Speaking Proficiency Assessment Conversational Agent,” in *The 12th Dialog System Symposium*, 2021, pp. 15–20.

[34] M. Saeki, W. Demkow, T. Kobayashi, and Y. Matsuyama, “A WoZ Study for an Incremental Proficiency Scoring Interview Agent Eliciting Ratable Samples,” in *12th International Workshop on Spoken Dialog System Technology (IWSDS 2021)*, 2021.

[35] J. E. Liskin–Gasparro, “The ACTFL Proficiency Guidelines and the Oral Proficiency Interview: A Brief History and Analysis of Their Survival,” *Foreign Language Annals*, vol. 36, pp. 483–490, 2003.

[36] L. Chen *et al.*, “Automated Scoring of Nonnative Speech Using the SpeechRater^{SM} v.5.0 Engine,” *ETS Research Report Series*, vol. 2018, no. 1, pp. 1–31, 2018.

[37] Council of Europe, *Common European Framework of Reference For Languages: Learning, Teaching, Assessment*. Cambridge University Press, 2018.

[38] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[39] M. Rahimpour and H. Fatemeh, “Topic Familiarity Effect on Accuracy, Complexity, and Fluency of L2 Oral Output,” *The Journal of Asia TEFL*, vol. 4, pp. 191–211, 2007.

[40] P. Tavakoli, F. Nakatsuhara, and A.-M. Hunter, *Scoring Validity of the Aptis Speaking Test: Investigating Fluency Across Tasks and Levels of Proficiency*. British Council, 2017, accessed 14 May 2022 at https://www.britishcouncil.org/sites/default/files/tava koli_et_al_layout.pdf.