Prompt-independent Automated Scoring of L2 Oral Fluency by Capturing Prompt Effects

Ryuki Matsuura[1][0000−0002−9429−6257] and Shungo Suzuki[1][0000−0002−6327−3298]

Waseda University, Tokyo, Japan
{matsuura,ssuzuki}@pcl.cs.waseda.ac.jp

Abstract. We propose a prompt-independent automated scoring method of second language (L2) oral fluency, which is robust to different cognitive demands of speaking prompts. When human examiners assess L2 learners' oral fluency, they can consider the effects of different task prompts on speaking performance, systematically adjusting their evaluation criteria across prompts. However, conventional automated scoring methods tend to ignore such variability in speaking performance caused by prompt design and use prompt-specific features of speech. Their robustness is thus arguably limited to a specific prompt used in model training. To address this challenge, we operationalize prompt effects in terms of conceptual, linguistic and phonological features of speech and embed them, as well as a set of temporal features of speech, into a scoring model. We examined the agreement between true and predicted fluency scores in four different L2 English monologue prompts. The proposed method outperformed a conventional method which used only temporal features ($\kappa = 0.863$ vs. 0.797). The detailed analysis showed that the conceptual and phonological features improved the performance of automated scoring. Meanwhile, the effectiveness of the linguistic features was not confirmed possibly because it may largely reflect redundant information to capture the prompt demands. These results suggest that the robustness of the automated fluency scoring should be achieved by careful consideration of what characteristics of L2 speech reflect the prompt effects.

Keywords: Automated speech scoring · Fluency · Prompt-independent scoring · foreign language learning · L2 speech

1 Introduction

It is crucial for second language (L2) learners to acquire an optimal level of fluency. In real-world communication, listeners tend to be distracted when speech is less fluent [4]. To encourage the acquisition of fluency, it is helpful to evaluate learners' current level and set realistic learning goals. However, the evaluation of L2 oral fluency by trained examiners requires time, cost and expert knowledge, which reduces learners' opportunities to attain fluency in an effective manner. Previous work has thus addressed automated speech scoring (ASS) for L2 fluency.

The existing ASS models have been trained to predict subjective fluency ratings based on temporal features of speech. The fluency rating is found to be

significantly associated with speed of delivery, pausing behavior and disfluency phenomena (e.g., self-repair and repetition) [13]. Building on this finding, ASS models of fluency have commonly used the corresponding temporal features, such as speech rate, pause duration and repaired word frequency, as input (e.g., [6, 7, 10]). While the models using the temporal features can approach human performance in a specific prompt that is used for the training, they have failed to maintain the same level of prediction accuracy in other prompts. This low applicability of the trained model to new speaking contexts may provide incorrect scores to learners and subsequently may hinder effective L2 speech learning. It is thus expected to develop a prompt-independent ASS system for fluency, which is robust to various speaking prompts.

The aforementioned problem is possibly due to an ignorance of variability in temporal features caused by different cognitive demands across prompts. L2 research consistently showed that variability in speaking performance across different prompts is explained by the fact that different speaking prompts impose different quality of demands on components of speech production processes, such as content planning and linguistic encoding [9]. Despite such variability of speech performance, human examiners can evaluate L2 fluency consistently across multiple prompts. Previous studies on L2 fluency found that the fluency evaluation is influenced not only by temporal features of speech but also by non-temporal ones (e.g., content, grammar and pronunciation), suggesting that human examiners adjust evaluation criteria by intuitively inferring the prompt effects from a whole range of speech characteristics [9].

To address this challenge of the prompt-specific ASS, the current study aims to operationalize the cognitive demands of prompts and incorporate conceptual, linguistic and phonological features of speech as well as temporal ones into the fluency scoring system. We examined the effectiveness of these features by comparing the performance of subjective fluency score predictions between the proposed scoring method and a conventional method based only on temporal features as the input. In an experiment, we used four monologue prompts to extract these features of speech and construct ASS models. We also conducted a follow-up analysis to examine the relative importance of them in relation to the prediction accuracy of the models.

2 Related Work

Previous research on ASS for fluency has focused on prompt-specific systems, which are developed and evaluated using the same prompt. For example, Shen et al. [10] trained and tested a regression model to predict fluency scores in a picture description prompt, and it outperformed human performance (r = 0.956 vs. 0.910). While the prompt-specific models have achieved high prediction performance, the rate of prediction accuracy often declines when the models are tested with new prompts. Matsuura et al. [6] predicted human ratings of fluency in an oral proficiency interview task consisting of multiple question-format prompts. Their prompt-specific ASS model showed that the classification accu-

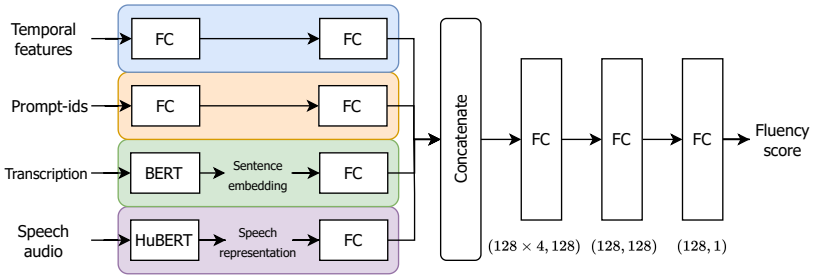

Fig. 1. Architecture of prompt-independent automated fluency scoring model

racy was low for novice and pre-advanced learners. The strong reliance of their prompt-specific models on a target prompt may have lowered the prediction accuracy of fluency scores, highlighting the necessity of prompt-independent ASS systems. To the best of our knowledge, such systems have not been developed in the domain of speaking assessment. Meanwhile, there are several successful scoring models for assessing L2 essays. Despite the difference in the modality between speaking and writing, one major issue is the methods for extracting and controlling for the effects of features that are highly subject to prompt design. In the automated essay scoring, semantic information of essays is considered to be prompt-specific, and thus alternatively some general features, such as grammatical structure and text length, are used as a proxy for the prompt-specific characteristics of essays [8]. When it comes to the oral fluency scoring, given temporal features of speech are, by nature, highly subject to the prompt design, an adaptation of the models to various prompts should be achieved by manipulating non-temporal features of speech. We therefore propose an incorporation of conceptual, linguistic and phonological features into the temporal ones to develop the prompt-independent fluency scoring system.

3 Prompt-Independent Automated Fluency Scoring

We propose the ASS method for fluency with conceptual, linguistic and phonological features as well as temporal features (see Figure 1). Three dimensions of temporal characteristics of speech are captured, following [6]. Speed feature is represented by articulation rate, speech rate and mean length of run; breakdown feature includes mid-clause pause ratio, end-clause pause ratio, filled pause ratio, mid-clause pause duration, end-clause pause duration and mean pause duration; and repair feature is measured by disfluency ratio, repetition ratio and self-repair ratio. To extract the conceptual feature, we assume that speech content is largely controlled by prompt design and thus use prompt-ids as an input of the ASS model. As for the linguistic feature, a sentence embedding is extracted by means of BERT [1], which can capture lexical and grammatical characteristics of sentences [3]. In our study, a hidden state of [CLS] token is used as the sentence embedding. For the phonological feature, we utilize the pre-trained HuBERT [2] and extract speech representation, which is reflective of phonetic characteristics of speech [14]. The extracted features are mapped to 128 dimension vectors by

fully-connected (FC) layers and concatenated. This vector is further fed into three FC layers, and finally fluency score is predicted. Moreover, the dropout is adopted to avoid an overfitting to the training data. We employ Adam as an optimizer, and hyper-parameters are determined by the tuning.

4 Experiment

To examine the proposed prompt-independent fluency scoring model, we used English monologue speech elicited from four speaking tasks differing in the nature of cognitive demands [12]: an argumentative speech, a related picture narrative, a reading-to-speaking (RtoS) and a reading-while-listening-to-speaking (RwLtoS)[1]. The argumentative prompt demanded an ability of content planning, in which learners argued for or against a given statement with valid reasons. Meanwhile, with the picture narrative prompt, learners were demanded to use their ability to describe a cartoon, where the content of speech is largely predefined by the visual prompt. The RtoS was a prompt to give a oral summary of a given English written text and differs from the picture description in that the essential vocabulary was presented in the text. The requirement for the RwLtoS prompt was same for the RtoS, whereas the RwLtoS also provides phonological information of the vocabulary by the audio-recording of the text. All four tasks were completed by 128 Japanese learners of English, and speech duration was around two minutes on average. The first one-minute excerpts of 512 speech were scored for fluency by two Ph.D. students in Applied Linguistics independently. They evaluated L2 oral fluency on a nine-point scale, focusing on temporal aspects of speech, and the inter-rater reliability was confirmed (Cronbach's $\alpha = 0.819$). As the ground truth of fluency score, we removed raters' severity by conducting many-facet Rasch analysis [5], and the estimated fluency scores for each speech sample were re-scaled on a six-point scale. To train and evaluate the model, speech samples were split into 384 training, 64 validation and 64 test set, with a equal balance of four prompts.

Considering the ordinality of the fluency scores, we evaluated the fluency prediction performance using quadratic weighted κ (QWK) and Peasron's correlation coefficient (PCC) between the predicted and true fluency scores. To confirm the effectiveness of the non-temporal features for the prompt-independent ASS, we compared the proposed method with the conventional one which sorely depends on temporal features of speech. Moreover, we compared different combinations of the conceptual, linguistic and phonological features to investigate which features can capture the cognitive demands of the prompts. The basic architecture and training setting of each model was identical to the proposed method except for the type of features to be concatenated.

5 Results and Discussion

We summarized the results of the experiment in Table 1. First, the proposed method outperformed the conventional one in terms of QWK and PCC. This

[1] https://osf.io/zrwmn/?view_only=0eeb1c966cb64afc9834acf80a42ad7e

Table 1. QWK and PCC of fluency score predictions.

model					QWK	PCC
temporal				(conventional)	0.797	0.904
temporal + conceptual + linguistic + phonological				(proposed)	**0.863**	**0.932**
temporal + conceptual					0.799	0.867
temporal	+ linguistic				0.830	0.917
temporal		+ phonological			0.859	0.931
temporal + conceptual + linguistic					0.853	0.929
temporal + conceptual		+ phonological			**0.896**	0.929
temporal	+ linguistic + phonological				0.870	0.919

result indicates that the non-temporal features can increase the prediction accuracy of fluency scores when multiple prompts were included. We also compared the various combinations of the conceptual, linguistic and phonological features. When adding one of the non-temporal features, all methods achieved higher performance than the conventional method but lower than the proposed one. Meanwhile, the models which included two types of the features had the same or higher agreement than the one with one feature. This result suggests that combining at least two non-temporal features is effective in improving the robustness of the fluency scoring model to multiple prompts. Moreover, the combination of the phonological feature and either conceptual or linguistic features approached the same level of performance of the proposed method. These findings suggest that the phonological feature strongly contribute to the prompt-independent prediction, whereas the effectiveness of the linguistic feature may need further investigation. However, these results should be interpreted carefully. According to Suzuki and Kormos [11], the listeners' perception of L2 oral fluency is related to the linguistic dimension of the speech. This discrepancy between the previous and current findings might be attributed to the information involved in the BERT's sentence embedding. BERT is known to capture the sentiment and semantic characteristics of the text as well as lexical and grammatical ones [3]. In contrast, raters in the current study may not have relied on such sentimental aspects of speech to adjust the effect of the cognitive demands by the prompts because they were instructed to focus solely on temporal aspects of speech. Therefore, it might be better to disentangle essential information from the information-rich features to capture more refined prompt effects which improve the performance of the ASS. In future work, it should be investigated how to extract the lexical and grammatical features, which can enhance the performance of the prompt-independent ASS system. Although the effectiveness of the prompt-independent ASS by capturing the prompt effects were confirmed, it should be noted that the prompt-ids were used to capture the conceptual dimension of speech. Since the prompt-ids are determined by the type of prompts used in the training set, and conceptual features of completely new prompts cannot be captured, the current method reduces the robustness of the ASS model to such prompts. To solve this problem, future works should design the extraction method of conceptual features from speech transcriptions using natural language processing techniques, such as speech act classification.

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186 (2019). https://doi.org/http://dx.doi.org/10.18653/v1/N19-1423
2. Hsu, W.N., et al.: HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Trans. Audio, Speech and Lang. Proc. **29**, 3451–3460 (oct 2021). https://doi.org/10.1109/TASLP.2021.3122291
3. Jawahar, G., Sagot, B., Seddah, D.: What Does BERT Learn about the Structure of Language? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3651–3657. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1356, https://aclanthology.org/P19-1356
4. Lennon, P.: The Lexical Element in Spoken Second Language Fluency. In: Riggenbach, H. (ed.) Perspectives on fluency, pp. 25–42. University of Michigan Press, Ann Arbor (2000)
5. Linacre, J.M.: Many-Facet Rasch Measurement. MESA Press, Chicago (1989)
6. Matsuura, R., Suzuki, S., Saeki, M., Ogawa, T., Matsuyama, Y.: Refinement of Utterance Fluency Feature Extraction and Automated Scoring of L2 Oral Fluency with Dialogic Features. In: 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 1312–1320 (2022). https://doi.org/10.23919/APSIPAASC55919.2022.9980148
7. Ramanarayanan, V., Lange, P.L., Evanini, K., Molloy, H.R., Suendermann-Oeft, D.: Human and Automated Scoring of Fluency, Pronunciation and Intonation During Human–Machine Spoken Dialog Interactions. In: Proc. Interspeech 2017. pp. 1711–1715 (2017)
8. Ridley, R., He, L., Dai, X., Huang, S., Chen, J.: Prompt Agnostic Essay Scorer: A Domain Generalization Approach to Cross-prompt Automated Essay Scoring. arXiv (2020). https://doi.org/10.48550/ARXIV.2008.01441
9. Segalowitz, N.: Cognitive Bases of Second Language Fluency. Routledge, London & New York (2010). https://doi.org/https://doi.org/10.4324/9780203851357
10. Shen, Y., Yasukagawa, A., Saito, D., Minematsu, N., Saito, K.: 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). pp. 698–704 (2021). https://doi.org/10.1109/SLT48900.2021.9383458
11. Suzuki, S., Kormos, J.: Linguistic Dimensions of Comprehensibility and Perceived Fluency: An Investigation of Complexity, Accuracy, and Fluency in Second Language Argumentative Speech. Studies in Second Language Acquisition **42**(1), 143–167 (2020). https://doi.org/10.1017/S0272263119000421
12. Suzuki, S., Kormos, J.: The Multidimensionality of Second Language Oral Fluency: Interfacing Cognitive Fluency and Utterance Fluency. Studies in Second Language Acquisition (2022). https://doi.org/10.1017/S0272263121000899
13. Suzuki, S., Kormos, J., Uchihara, T.: The Relationship Between Utterance and Perceived Fluency: A Meta-analysis of Correlational Studies. The Modern Language Journal **105**(2), 435–463 (2021). https://doi.org/10.1111/modl.12706
14. Wells, D., Tang, H., Richmond, K.: Phonetic Analysis of Self-supervised Representations of English Speech. In: Proc. Interspeech 2022. pp. 3583–3587 (2022). https://doi.org/10.21437/Interspeech.2022-10884