Vol.2022-MUS-134 No.47
Vol.2022-SLP-142 No.47
2022/6/18

対話特徴を用いた第二言語発話の流暢性自動採点

松浦 瑠希[1,a)]　鈴木駿吾[1]　佐伯真於[1]　小川哲司[1]　藤江真也[2]　松山洋一[1]

概要：本稿では，対話における第二言語発話の流暢性自動採点について調査を行う．現実の対人コミュニケーションは対話形式で行われることが多く，第二言語発話の流暢性の評価についても，実際の対話を模擬した課題で行われることが望ましい．対話音声の採点では，流暢性スコアは各ターンの質的な現象だけでなく，ターンを跨いだ対話特有の現象に基づいて与えられる．しかし，既存の自動採点の多くはターンごとに発話特徴の抽出とスコアの予測をしており，対話特徴が考慮されていない．そこで，本研究では発話単位の特徴の統計量とターンを跨いだ特徴の双方を用いた第二言語発話の流暢性自動採点方式を提案する．85 人の日本自衛後学習者によるインタビューデータを用いて人手による流暢性スコアの予測実験を行ったところ，ターンごとに特徴抽出・スコア予測を行う従来法に対して，正解スコアと予測スコアの正解率，一致率，相関係数において提案法の有効性が示された．

Automated Scoring of L2 Oral Fluency Using Dialogic Features

1. はじめに

第二言語の発話能力の採点には，学習者や講師に対して学習の進捗状況や，次の学習目標を伝えるという重要な役割がある．そのため，より多くの学習者に対して第二言語発話の採点をすべきであるが，人による採点には多額の費用や時間がかかる，評価者による属人性がある等の課題が存在する [1]．また，流暢性は第二言語の発話能力を評価する指標の１つであり，総合的な評価との相関が高いことが報告されている [2]．したがって，第二言語学習者を対象とした発話流暢性の自動採点器の開発は，教育的な高い意義がある．

これまでの第二言語発話の流暢性自動採点では，独話音声に対するものが主な研究対象となっている [3], [4]．独話に対する自動採点では，一般的に，１つの発話から流暢性や発音，語彙や文法等に関係する特徴を抽出し，流暢性スコアを予測する [5]．一方で，現実の対人コミュニケーションは対話形式で行われることが多いため，対話音声に対する第二言語発話の自動採点が望まれる．対話における流暢性評価は，個々のターンに閉じた現象 (発話の速度や言い淀みの発生率など) のみならず，ターンを跨ぐ対話ならではの現象（対話相手とのインタラクションやその履歴，話題の遷移など）に基づいて行われており [6]，独話音声のための自動採点技術をそのまま転用することはできない．しかし，対話における自動採点ではターンごとに抽出された特徴のみを用いており [7], [8]，対話特有の特徴を考慮したものは少ない．それに対し本研究では，対話特有の特徴を用いた第二言語発話の流暢性自動採点方式を提案する．また，ターンごとに特徴抽出・流暢性スコア予測を行う従来方式に対する有効性を，インタビューデータを用いた実験による明らかにする．

本稿の構成は以下の通りである．第 2 章では，人による第二言語発話の流暢性採点と，対話音声を対象とした自動採点に関する先行研究について概観する．第 3 章では，提案手法について説明する．第 4 章では，評価実験とその結果について報告する．第 5 章では，実験の考察を行う．最後に第 6 章で，結論を述べる．

2. 関連研究

2.1 第二言語発話の流暢性採点

第二言語発話の流暢性には発話速度 (speed fluency)，ポーズ (breakdown fluency)，発話修正 (repair fluency) の 3 つの側面があるとされている [9]．流暢性研究では，この 3 つの側面が，人による流暢性評価に対してどのように寄与しているのかについて調査されてきた [10]．鈴木らによるメタ分析では，独話音声について，人による流暢

[1] 早稲田大学 Waseda University
[2] 千葉工業大学 Chiba Institute of Technology
[a)] matsuura@pcl.cs.waseda.ac.jp

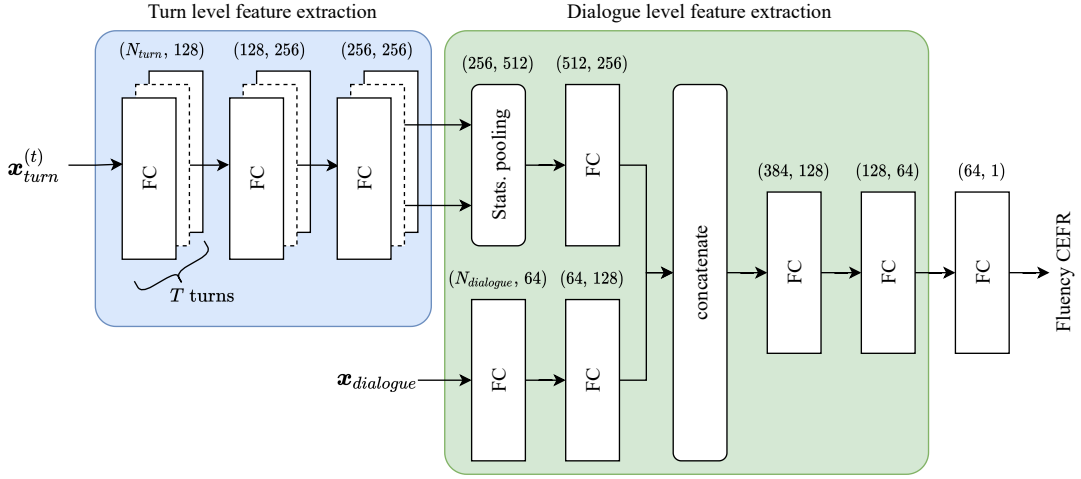


図 1: 統計プーリングを用いた流暢性自動採点器のアーキテクチャ．図中の $x_{turn}^{(t)}, x_{dialogue}, N_{turn}, N_{dialogue}$ はそれぞれ，t 番目のターンから抽出された流暢性特徴，対話流暢性特徴，流暢性特徴の次元数，対話流暢性特徴の次元数を表す．

Fig. 1 Architecture of automated oral fluency scoring using statistics pooling. Where $x_{turn}^{(t)}, x_{dialogue}, N_{turn}, N_{dialogue}$ are fluency feature extracted from t-th turn, dialogic fluency feature, dimension of fluency feature and dimension of dialogic fluency feature, respectively.

性評価と発話速度の相関が $r = 0.62$，ポーズ発生率との相関が $r = -0.59$ であり強い関係性があること，ポーズ長との相関が $r = -0.46$ と十分な関係性があることがわかった [11]．また，人による流暢性評価と発話修正の間には $r = -0.20$ と，有意な相関関係があった．一方で，対話音声における流暢性評価の性質は，独話音声とは異なることが指摘されている [6]．対話音声では同一話者の発話であっても，ターンや話題によって流暢性特徴が変わり得ることが知られている [12], [13]．よって，流暢性採点では，対話全体における発話速度，ポーズ，発話修正の傾向を捉える必要がある．また，対話における発話では，第二言語の習熟度が低い学習者と高い学習者の間で，ターン間に発生したポーズの平均長や対話相手の発話単語の繰り返し数などの対話流暢性 (dialogic fluency) の尺度が，統計的に有意に異なるという結果が得られている [14]．さらに，対話音声に対する人による流暢性評価と対話流暢性との間には相関関係 (e.g. ターンポーズ平均長: $r = -0.943$) があることが報告された [15]．したがって，対話音声における流暢性採点では，発話速度，ポーズ，発話修正だけでなく，対話流暢性も考慮する必要がある．

2.2 対話音声における第二言語発話の自動採点

対話音声における第二言語発話の自動採点では，ターンごとにスコアを予測することが一般的である．Ramanarayanan らは，各ターンに人手で採点された流暢性，発音，イントネーション・ストレスのスコアを，ターンごとに独立に予測する方法を提案した [7] ．しかし，対話には，採点に使える情報が含まれないような短い発話も多く，ターンごとの独立した自動採点は難しいという指摘がなされている．Qian らは，発話産出，言語使用，内容の 3 つの側面から第二言語発話の総合的な習熟度をターンごとに予測する方法を提案しており，内容に関する特徴抽出に End-to-End Memory Network (MemE2E) [16] を用いることで，対話の履歴を考慮する工夫をしている [17]．対話全体の自動採点においても，ターンごとにスコアを予測し，それらの統計量を計算することが多い．対話音声における第二言語発話のインタラクション性自動採点では，対話全体に対して採点された 1 つのインタラクション性のスコアを個々のターン全てに割り振ってから，ターン単位で自動採点器を学習し，予測スコアの中央値を対話全体のスコアとする方法が提案されている [18]．このような手法では，学習データの水増しが可能である [8] が，対話特徴を用いることができないという課題がある．そこで本研究では，ターンごとに抽出される特徴の統計量と対話特有のターンを跨いだ特徴の双方を用いた第二言語発話の流暢性自動採点方式を提案する．

3. 対話特徴を用いた流暢性自動採点

本研究では，統計プーリング [19] を用いて発話特徴の統計量とターンを跨いだ特徴の抽出をし，第二言語発話の流暢性自動採点を行う．提案する流暢性自動採点器のアーキテクチャを**図 1** に示す．2 章で述べたとおり，対話音声における流暢性採点においては，個々のターンの流暢性特徴ではなく，対話全体における傾向を捉える必要がある．そのため，ターンごとに抽出される発話速度，ポーズ，発話修正特徴に対して統計プーリングで平均と標準偏差を計算

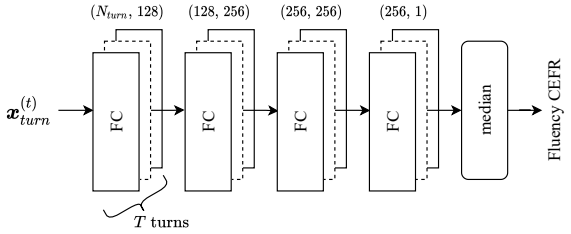


図 2: 従来法による流暢性自動採点器のアーキテクチャ.
Fig. 2 Architecture of conventional automated oral fluency scoring.

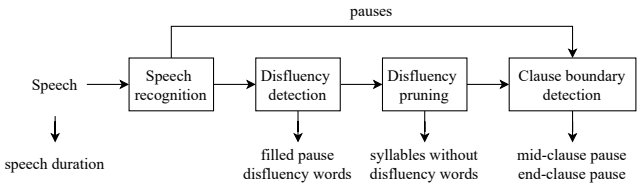


図 3: 非流暢現象の自動アノテーション過程.
Fig. 3 Annotation step of disfluency phenomena.

することで，対話単位での流暢性特徴を抽出する．また，対話においては，数単語のみからなる流暢性採点への寄与が低いターンもあるため，注意機構を用いて各ターンに適当な重み付けをしてから，平均と標準偏差の計算をしている [20]．さらに，対話特有の流暢性の性質を捉えるため，発話速度，ポーズ，発話修正特徴の平均・標準偏差に対して，対話流暢性特徴を結合する．得られた対話単位の流暢性特徴を全結合層入力し，対話全体に与えられた流暢性スコアを予測する．

4. 流暢性自動採点実験

提案法による流暢性スコアの予測精度について，従来法と比較することで評価をした．また，複数の評価者による採点の一致率を基準として，人と提案法の比較も行った．実験において，3 人の専門家による英語インタビューの流暢性採点結果を予測すべき流暢性スコアとした．従来法は対話の各ターンから予測された流暢性スコアの中央値をインタビュー全体の流暢性スコアとした．一方で，提案法は対話特徴抽出部で抽出した対話単位の流暢性特徴から直接的にインタビュー全体の流暢性スコアを予測した．

4.1 実験条件

流暢性スコアの予測は，正解率 (accuracy)，重み付きカッパ係数 (quadratic weighted κ)，相関係数 (Pearson's r) の 3 つの指標を用いて，5 分割交差検証により行った．また，評価者間の採点は，Krippendorff's α と級内相関係数 (intraclass correlation coefficient) を用いて一致率を計算し，提案法の重み付きカッパ係数，相関係数との比較を行った [3], [18]．実験では，ターンごとに流暢性スコアを予測し，それらの中央値を対話全体のスコアとするニューラルネットを従来法とした [8], [18]．提案法と従来法のネットワークの構成とパラメータは図 1，**図 2** に示す通りで，Optimizer はそれぞれ学習率 0.0020, 0.0015 の Adam[21] とした．また，両手法ともドロップアウト (提案法: $p = 0.05$，従来法: $p = 0.2$) による正則化を行った．

4.1.1 WoZ インタビューデータ

流暢性自動採点器の学習と評価には，Wizard of Oz (WoZ) 法によって操作された対話エージェントと人間によるインタビューデータ [22] を用いた．インタビューは 7 つの異なる話題から構成されており，アメリカ外国語教育審議会 (ACTFL) が提案する Oral Proficiency Interview (OPI) [23] を参考に，学習者の第二言語の習熟度を逐次的に評価し，それに合わせて次の話題の難易度を動的に変えた．本研究では，85 人の日本人英語学習者に対して行われたインタビューデータ用い，各インタビューは約 9 分程度であった．

4.1.2 人による流暢性採点

インタビューの流暢性スコアは，3 人の評価者によって行われた [24]．流暢性は Common European Framework of Reference for Languages (CEFR) [25] に基づいて評価された．CEFR は A1, A2, B1, B2, C1, C2 の 6 段階のレベルから構成されており，A1 が最も低く，C2 が最も高い評価を表している．評価者はいずれも，10 年以上の英語教育および評価の経験がある現職の英会話講師であった．**図 2** に示すとおり，3 人の評価者間の流暢性スコアの一致率は Krippendorf's α で 0.804，級内相関係数で 0.894 と高い一致率が確認できた．最終的な流暢性スコアは，多相ラッシュ分析 [26] によって評価者の厳しさを統制した値を用いた．その結果，各流暢性スコアの人数は A1 から C2 まで順に，5, 21, 34, 15, 9, 1 となり，C2 レベルの学習者が 1 人のみであったため，C1 レベルと合わせて C+ レベルとした．

4.1.3 流暢性特徴

本研究で用いた流暢性特徴は**表 1** にまとめた．発話速度，ポーズ，発話修正に関する流暢性特徴は [4] に従い自動抽出をした．流暢性特徴の抽出に必要な自動アノテーションの過程は**図 3** に示す．実験では，0.25 ミリ秒以上の無音区間をポーズとして扱い [11], [27]，音声認識とポーズ検出に Rev.ai[*1] の提供する Asynchronous Speech-to-Text，言い淀み検出に Switchboard reannotated dataset [28] でファインチューニングした BERT [29]，節境界検出器に Stanford CoreNLP [30] の構文解析器を用いた．音声認識の単語誤り率は 27.3% であり，[31] で報告されているもの (28.5%) より高い精度を示していた．人と流暢性特徴抽出器のアノテーション精度を比較したところ，節内・節間ポーズ分

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

表 1: 流暢性特徴のリスト.

Table 1 List of fluency features.

Type	Parameter	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 between 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.

表 2: 3 人の評価者による流暢性スコアの Krippendorff's α と級内相関係数.

Table 2 Krippendorff's α and intraclass correlation coefficients by three raters.

metrics	human
Krippendorff's α	0.804
Intraclass correlation	0.894

類の Cronbach's α と Cohen's κ はそれぞれ $\alpha = 0.999$, $\kappa = 0.613$, 言い淀み検出はそれぞれ $\alpha = 0.999$, $\kappa = 0.674$ と高い一貫性と十分な一致率が確認された [32].

実験では，従来法の入力を調音速度 (articulation rate)，節内ポーズ発生率 (mid-clause pause ratio)，節間ポーズ発生率 (end-clause pause ratio)，フィラー発生率 (filled pause ratio)，節内ポーズ平均長 (mid-clause pause duration)，節間ポーズ平均長 (end-clause pause duration)，言い淀み発生率 (disfluency ratio) とした．提案法では，従来法の入力に加えて，ターンポーズ数 (number of between-turn pauses)，ターンポーズ平均長 (between-turn pause duration)，ターン数 (number of turns)，ターン平均長 (mean length of turn)，対話相手の発話単語の繰り返し数 (number of other repetitions) を，対話流暢性特徴として入力した．

4.2 実験結果

提案法，従来法による流暢性スコアの予測を 5 回行った結果の平均と標準偏差を**表 3** に示す．また，正解率が最も高くなった流暢性スコアを予測した結果の混同行列を**図 4** に示す．結果より，提案法による流暢性スコア予測の方が，正解率，一致率，相関係数の全てにおいて従来法の精度を超えることがわかった．特に重み付きカッパ係数より，提案法による予測流暢性スコアと正解ラベルとの一致率は，従来法を大きく超えることがわかった．しかし，提案法による予測の一致率と相関は，評価者間の一致率，級内相関より低いことがわかった．混同行列を比べると，従来法は殆どの流暢性スコアを B1 レベルと予測してしまっており，特に流暢性スコアが A1，A2 レベルの予測において誤りが多かった．一方で，提案法では，A1 レベルを除き，6 割を超えるサンプルにおいて正解と同じ流暢性スコアの予測をしており，誤った予測においても殆どが 1 レベル以内の誤差に収まっていた．

表 3: 提案法と従来法による予測流暢性スコアの正解率，重み付きカッパ係数，相関係数の平均と標準偏差

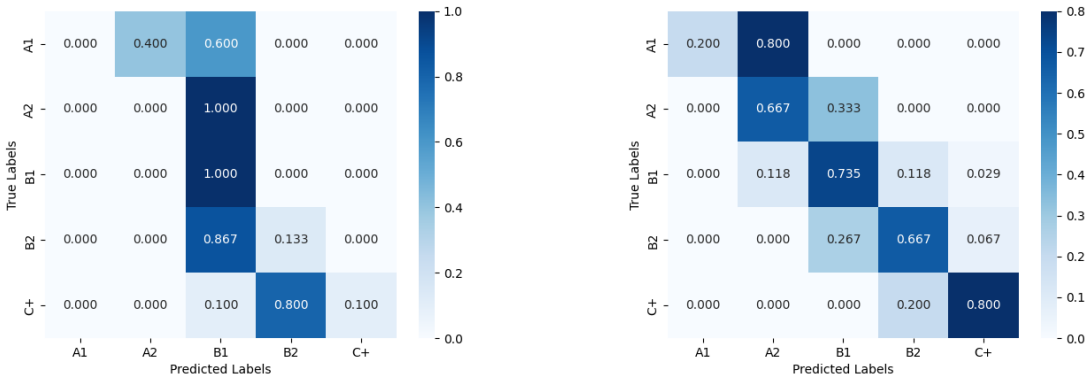
Table 3 Mean and standard deviation of accuracy, quadratic weighted κ and pearson's r correlation coefficient of fluency CEFR prediction by proposed and conventional models.

metrics	conventional	proposed
Accuracy	0.424 ± 0.008	**0.633 ± 0.035**
QW κ	0.398 ± 0.051	**0.792 ± 0.040**
Pearson's r	0.701 ± 0.009	**0.870 ± 0.012**

5. 議論

本稿では，対話特徴を用いた流暢性自動採点について調査を行った．ターンごとに流暢性スコアを予測し，それらの中央値を対話全体のスコアとする従来法と比べ，統計プーリングによる対話特徴抽出を行う提案法の方が，正解値と予測値の正解率，一致率，相関がより高くなることがわかった．このことから，対話音声を対象とした流暢性自動採点では対話特徴を考慮することの有効性が確認できた．

一方で，提案法による正解スコアと予測スコアの混合行列を見ると，A1 レベルの予測の殆どが誤りであることがわかる．原因として，A1 レベルと判定されたインタビュー数が 5 と少ないこと，発話を数語で終了する場合があり，流暢性特徴を上手く抽出できないことが多いことが考えら



(a) Conventional model.　　(b) Proposed model.

図 4: 従来法と提案法による流暢性スコア予測結果の混同行列.

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

れる [33]．今後の課題として，習熟度の低い学習者のデータを更に収集すること，発語数に頑健な流暢性特徴を考えることが必要である．

また，流暢性スコアの一致率について，提案法は人の精度には及ばないことがわかった．これは，対話特徴抽出部によって得られた特徴量が，対話の性質を十分に捉えられていないことが原因だと考えている．具体的には，まず，インタビュー中の話題の難易度の影響を考慮することができていない．本実験で用いたインタビューは，学習者のパフォーマンスによって話題の内容や難易度が動的に変化する形式となっている．また，話題の難易度や親密度は発話速度，ポーズ，発話修正特徴の値に影響を与えるため，異なるいくつかの話題で抽出された流暢性特徴を厳密に同一に扱うことはできない．しかし，統計プーリングでは，流暢性特徴が話題に依存しないという仮定を置いて，平均・標準偏差を計算している．今後の研究では，話題の難易度を考慮した対話特徴の抽出方法について調査する．次に，インタビュー中に生じる対話の破綻を捉えることができていない．話題が動的に変化するインタビューにおいて，話題の難易度が学習者の第二言語能力を超えるとき，ブレイクダウンと呼ばれる対話破綻が起こることがある [23]．また，ブレイクダウンが発生すると発話途中で無言になったり，繰り返しや言い直しが頻発したりすることが知られており [24]，流暢性特徴との関係生が大きい．したがって，ブレイクダウンの考慮が流暢性自動採点においても有効だと考えられる．ただし，ブレイクダウンを厳密に定義，アノテーションすることは難しく，かつ，個人によってブレイクダウンによる発話速度，ポーズ，発話修正への影響も変化する (e.g. 発話がゆっくりになる，言い淀みが増える)．そこで，学習者個々人の発話速度，ポーズ，発話修正特徴の分布内での差から，ブレイクダウンに相当する特徴を抽出することを検討する予定だ．

6. まとめ

本研究では，対話特有の特徴を用いた第二言語発話の流暢性採点について調査を行った．提案法の評価実験では，正解スコアと予測スコアの正解率，一致率，相関について，流暢性特徴の抽出と流暢性スコアの予測をターンごとに独立に行う従来法と比較をした．その結果，従来法に比べ，3つの評価指標の全てが向上することを確認し，提案法の有効性を明らかにした．ただし，A1 レベルの予測について誤りが多いこともわかった．今後は習熟度が低い学習者のデータ収集や，より適した流暢性特徴について検討をする予定である．また，提案した統計プーリングによる対話特徴抽出部にもいくつかの課題があることを考察した．今後，話題の難易度やブレイクダウンを考慮した対話特徴の抽出方法について調査をしていく必要があると考えている．

謝辞　この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の委託業務 (JPNP20006)「人と共に成長するオンライン語学学習支援 AI システムの開発」の結果得られたものです．

参考文献

[1] Brown, A.: Interviewer Variation and the Co-construction of Speaking Proficiency, *Language Testing*, Vol. 20, No. 1, pp. 1–25 (2003).
[2] Suzuki, S. and 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*, Vol. 42, No. 1, p. 143–167 (2020).
[3] Shen, Y., Yasukagawa, A., Saito, D., Minematsu, N. and Saito, K.: Optimized Prediction of Fluency of L2 English Based on Interpretable Network Using Quantity of Phonation and Quality of Pronunciation, *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 698–704 (2021).
[4] 松浦瑠希，鈴木駿吾，佐伯真於，小川哲司，松山洋一：

言い淀みとポーズ位置検出に基づく第二言語発話の流暢性自動採点，日本音響学会研究発表会講演論文集，pp. 1351–1354 (2022).

[5] Zechner, K. and Evanini, K.: *Automated Speaking Assessment Using Language Technologies to Score Spontaneous Speech 1st Edition*, Routledge, New York (2019).

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

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

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

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

[11] Suzuki, S., Kormos, J. and Uchihara, T.: 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).

[12] Heather, B., Silvia, D., L., Jonathan, E., B., Michael, F., S. and Susan, E., B.: Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender, *Language and Speech*, Vol. 44, No. 2, pp. 123–147 (2001).

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

[14] Peltonen, P.: Temporal Fluency and Problem-Solving in Interaction: An Exploratory Study of Fluency Resources in L2 Dialogue, *System*, Vol. 70, pp. 1–13 (2017).

[15] Peltonen, P.: 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).

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

[17] Qian, Y., Lange, P., Evanini, K., Pugh, R., Ubale, R., Mulholland, M. and Wang, X.: Neural Approaches to Automated Speech Scoring of Monologue and Dialogue Responses, *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8112–8116 (2019).

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

[19] Snyder, D., Garcia-Romero, D., Povey, D. and Khudanpur, S.: Deep Neural Network Embeddings for Text-Independent Speaker Verification, *Proc. Interspeech 2017*, pp. 999–1003 (2017).

[20] Okabe, K., Koshinaka, T. and Shinoda, K.: Attentive Statistics Pooling for Deep Speaker Embedding, *Proc. Interspeech 2018*, pp. 2252–2256 (2018).

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

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

[23] Liskin–Gasparro, J. E.: 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).

[24] 佐伯真於，松浦瑠希，鈴木駿吾，宮城琴佳，小林哲則，松山洋一：InteLLA: 適応的な質問戦略を有するスピーキング能力判定会話エージェント，人工知能学会研究会資料言語・音声理解と対話処理研究会，Vol. 93, pp. 15–20 (2021).

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

[26] Linacre, J. M.: "Many-Facet Rasch Measurement" (1989).

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

[28] Zayats, V., Tran, T., Wright, R., Mansfield, C. and Ostendorf, M.: Disfluencies and Human Speech Transcription Errors, *Proc. Interspeech 2019*, pp. 3088–3092 (2019).

[29] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *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).

[30] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J. and McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit, *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60 (2014).

[31] 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. and Gyawali, B.: 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).

[32] Matsuura, R., Suzuki, S., Saeki, M., Ogawa, T., Kobashikawa, S. and Matsuyama, Y.: Automated Scoring of L2 Oral Fluency in a Dialogue TaskBased on Disfluency Phenomena Annotation and the Statistics Pooling (forthcoming).

[33] Tavakoli, P., Nakatsuhara, F. and Hunter, A.-M.: *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/tavakoli_et_al_layout.pdf.