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*

☆松浦瑠希, 鈴木駿吾 (早大・GCS 機構)，佐伯真於，
小川哲司（早大・理工），松山洋一（早大・GCS 機構）

1　はじめに

第二言語学習者による発話の流暢性自動採点における，言い淀み除去とポーズ位置の節内・節間判定を組み込んだ自動特徴抽出方法を提案する．

第二言語発話の流暢性は，発話の「速度」「ポーズ」「修正」の 3 つの側面から評価される．発話の速度ではポーズを含まない単位発話時間あたりの音節数で計算される調音速度，ポーズについてはその平均長や発生率，修正では発話における言い淀みの割合等の特徴量が用いられる．一方で，流暢性特徴の算出過程において，自動採点の妥当性に影響を与える問題がいくつかある．まずは，発話の速度特徴における言い淀みである．言い淀んだ単語が含まれたままの発話文章で調音速度を計算すると，流暢性評価の弁別性がなくなる恐れがある [1]．次に，ポーズの発生箇所によるその特徴量への影響である．節中で発生するポーズと流暢性評価の相関が強いことがわかっており [2]，発話のポーズ特徴は節内・節間の発生位置で分けて考えるべきだと指摘されている [3]．しかし，流暢性の自動採点のための特徴抽出において，これら 2 つの課題を自動で行うような研究は未だ例を見ない．

本研究では，不要な言い淀みの除去と節内・節間ポーズ分類を組み込んだ流暢性特徴抽出器を構築し，上述の 2 つの課題を解決する形で妥当性の高い流暢性の自動採点を実現することを試みる．また，提案する特徴抽出器を用いて，専門家による 4 種類の独話課題に対する流暢性の採点結果を予測する．流暢性特徴の抽出を人手で行った場合と性能を比較することで，構築した特徴抽出器の有効性を明らかにする．

2　流暢性特徴抽出

流暢性の測定に寄与する発話速度特徴，ポーズ特徴，発話修正特徴の抽出方法について述べる．

発話速度，ポーズ，発話修正に関する特徴として用いるパラメータを表 1 に示す．また，これら流暢性特徴の抽出過程を図 1 に示す．提案する流暢性特徴抽出器は，音声認識器，無音区間検出器，言い淀み検出器，言い淀み除去器，節境界検出器から成る．

Table 1　Fluency features

Type	Parameter
Speed	Articulation rate
Pause	Mid-clause pause ratio
	End-clause pause ratio
	Mean pause duration
Repair	Disfluency ratio

2.1　発話速度特徴

発話速度特徴には調音速度 (Articulation rate) を用いる．調音速度は，単位発話時間あたりの単語もしくは音節数（単語や音節の数をポーズを除いた発話長で割った値）として得られる．このとき，流暢性を正しく評価するためには，発話された単語のうち言い淀みに相当するものを除いて計算すべきである．例えば，習熟度の異なる 2 人の英語学習者による以下の発話に対して，調音速度を計測することを考える．発話時間はともに 10 秒とする．

A: {I, I am}, I live in Tokyo with my family.

B: I live in Tokyo, more specifically in Shinjuku, with friends.

このとき，話者 A の発話では，文頭の “I, I am” は，後続する “I” によって言い直されていることから，言い淀みと判断できる．一方，話者 B の発話には言い淀みがなく，話者 A よりも流暢と言える．いま，言い淀みを除かなければ話者 A と B の発話はともに 10 単語であり，調音速度はともに 1.0 単語/秒である．一方で，言い淀み単語を除くと話者 A の発話は 7 単語となるので，調音速度は話者 A が 0.7 単語/秒，話者 B が 1.0 単語/秒となる．このように，言い淀み単語を除いて計算することで，調音速度は流暢性について弁別的になり，直感とも合う．

言い淀みの検出および除去は，音声認識によって得られる単語ごとにそれが言い淀みか否かを識別することで行った．この言い淀み検出器は，BERT [4] と 1 層の識別層で構築し，各単語ごとに言い淀みか，言い淀みの修正か，それ以外かの 3 つを識別するようファインチューニングすることで得た (図 2)．この検

*Automated scoring of L2 fluency based on detection of disfluency words and pause locations. by MATSUURA, Ryuki, SUZUKI, Shungo (GCS research organization), SAEKI, Mao, OGAWA, Tetsuji (Waseda University), MATSUYAMA, Yoichi (GCS research organization)

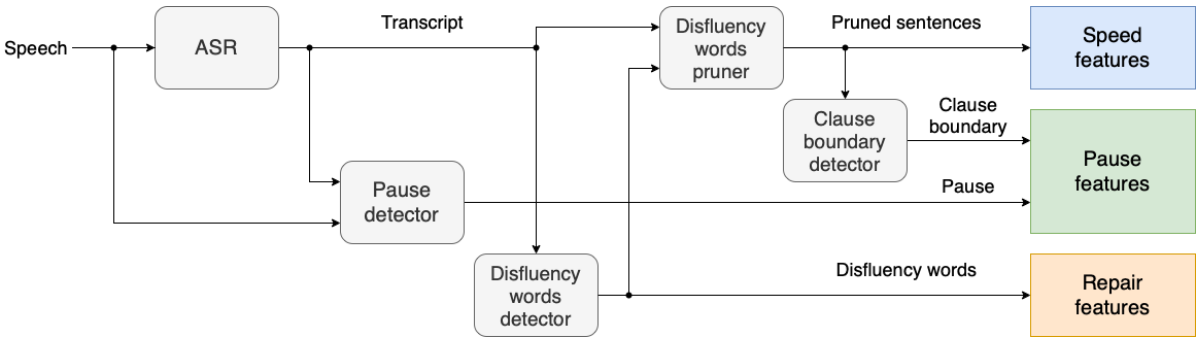

Fig. 1　Schematic diagram of fluency feature extraction.

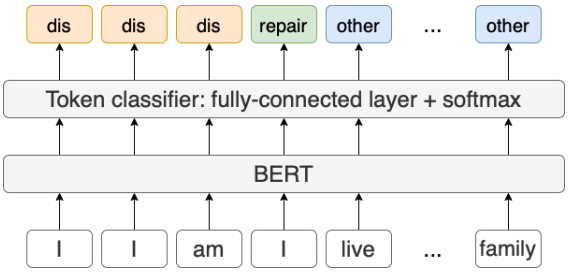

Fig. 2　Architecture of disfluency words detector. The “dis” tag means disfluency word.

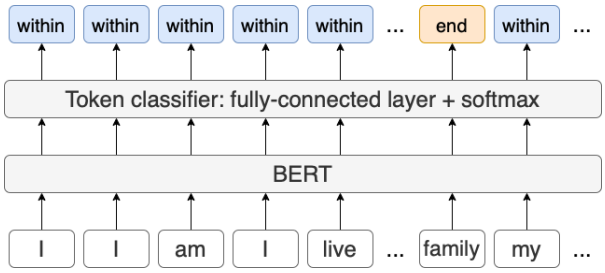

Fig. 3　Architecture of sentence end words detector.

出器により言い淀みとして検出された単語を除いたうえで，発話内の音節数を計数した．また，音声認識結果と音声信号の強制アライメントにより無音以外の単語に割り当てられた時間長の総和として発話時間を得た．以上のようにして得た音節数と発話時間から調音速度が計算できる．このとき，構文解析および言い淀み検出は文単位で処理を行うため，音声認識結果から文末単語の検出を行った．図 3 に示すように，文末単語検出器も BERT と 1 層の識別層からなる構造とし，各単語が文末か否かを識別するようファインチューニングすることで構築した．

2.2　ポーズ特徴

発話中のポーズに関連した特徴量として，節内ポーズ発生率，節間ポーズ発生率，平均ポーズ長を算出する．ポーズ発生率はポーズ数を音節数で割った値であり，平均ポーズ長はポーズの長さの平均値である．なお，250 ミリ秒以上の長さの無音区間をポーズとした [5]．

ポーズ発生率を節内，節間に分けて計算しているのは，第二言語の習熟度が低いほど語彙や文法知識の不足による言語産出の困難を反映して節内ポーズの数が増える傾向があり [6, 7, 8]，流暢性評価においても，節内ポーズ発生率の弁別性が高い [2, 3] という知見に基づいている．このとき，流暢性を正しく評価するためには，ポーズの位置が節内か節間かを正確に推定する必要がある．

本研究では，これを節境界検出により実現する．言い淀みを含まない文に対して，構文解析器によって推定された従属節の係り受け構造から節の境界を見つけ出すとともに，発話中のポーズを検出する．ここで，節中にあるポーズを節内ポーズ，節境界にあるポーズを節間ポーズとして計数し，発話文中の音節数で割ることで，節内ポーズ発生率，節間ポーズ発生率を算出した．また，全てのポーズからポーズの平均長を計算した．

2.3　発話修正特徴

発話の修正に関する特徴量として，言い淀み発生率を算出する．言い淀み発生率は，言い淀み単語数を発話された単語の音節数で割った値である．ここでは，言い淀み検出器によって検出された言い淀み単語を数え上げ，音声認識結果より得られる音節数で割ることで算出された．

3　関連研究

代表的な流暢性特徴の自動抽出方法は，音響特徴を用いるものが挙げられる [9, 10]．この方法では，発話音声の音圧レベルのピーク数と相対的な差から音節数や無音区間を推定し，発話速度とポーズに関連する特徴量を算出する．しかし，音響特徴のみによる流暢性特徴抽出では，言語情報を用いないため言い淀みや節の境界の自動検出が行えない．したがって，発話の修正に関連する特徴量は算出できず，言い淀み

Table 2　Correlation between true fluency scores and automated fluency scoring results using manually and automatically extracted features ($df = 125$)

task	extract method	r	T	p-value
argumentation	manual	0.815	1.533	0.128
	auto	0.780		
picture narration	manual	0.787	3.746	< 0.01
	auto	0.692		
reading-to-speech	manual	0.718	0.455	0.650
	auto	0.713		
reading-while-listening-to-speech	manual	0.769	0.105	0.917
	auto	0.755		

Table 3　Correlation coefficients and mean squared error between manually and automatically extracted articulation rate with and without pruning of disfluency words

	r_{AR}	MSE_{AR}
without pruning	0.615	0.303
with pruning (proposal)	**0.646**	**0.286**

除去をした調音速度や節内・節間ポーズの発生率等の特徴量も得られない．

Educational Testing Service(ETS) が開発した SpeechRater [11] は，音声認識による自動書き起こし文章から，節の境界と文中の言い淀み単語系列の終了箇所を推測することで，発話修正特徴の抽出や節内・節間ポーズ分類の課題を解決している [12]．ただし，SpeechRater は各単語に対して，言い淀みか否かの推定を行わないため，言い淀みの語の特定及び除去は不可能である．本研究の手法では，SpeechRater の特徴抽出方法を拡張することで，言い淀み除去と節内・節間ポーズ分類の自動化を行う．

4　流暢性推定実験

提案法により自動で抽出した流暢性特徴を用いたときの流暢性予測性能と，提案法における性能の上限値である，人手で算出した流暢性特徴を用いたときの流暢性予測性能を比較することで，提案法の精度を評価した．ここでは，2 人の専門家が流暢性を評価した結果を予測すべき流暢性スコアとした．

4.1　実験条件

流暢性特徴を用いた流暢性スコアの予測 (流暢性自動採点) は，重回帰モデルを用いて行った．流暢性予測性能の評価は，実際の流暢性スコアと予測されたスコアの相関係数を指標とし，3 分割交差検定により行った．また，流暢性スコア (ground truth) は，評価者 2 名による 9 段階評価に対して，多相ラッシュ分析 [13] で評価者の厳しさを統制した値を用いた．なお，評価者の負担を考慮して，流暢性評価は発話音声の冒頭 1 分程度について行った．

実験には，128 人の日本語母語話者による計 512 の英語の独話音声データを用いた [14]．発話課題[1]は，(a) 予め与えられた質問に対して意見を述べる「意見述べ」，(b) 絵から読み取れるストーリを説明する「絵描写」，(c) 文章を読み，内容を説明する「読み上げ音声なし再話」，(d) 読み上げ音声を聞きながら文章を読み，内容を説明する「読み上げ音声あり再話」の 4 種類を用いた．

提案する特徴抽出の性能の上限値を得るために，発話の書き起こし，言い淀みの特定と除去，ポーズの特定と節内・節間ポーズの分類を全て人手で行い，流暢性特徴を抽出した．流暢性特徴の自動抽出においては，音声認識に Google Cloud Speech to Text[2]，無音区間検出器に Montreal Forced Alignment [15]，節境界検出器に Stanford CoreNLP [16] の構文解析器を使用した．また，言い淀み検出器および文末単語検出器は，言い淀み単語が含まれる対話コーパスである Switchboard reannotated corpus [17] を用いて，Hugging Face が提供する事前学習済み BERT [18] をファインチューニングをした．

4.2　実験結果

手動と自動で算出した流暢性特徴による流暢性スコアの予測性能と，それらの比較結果を表 2 に示す．実験結果から，いくつかの独話課題において，提案法による流暢性特徴抽出の妥当性を確認できた．意見述べ，読み上げ音声なし再話，読み上げ音声あり再話の

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

[2] https://cloud.google.com/speech-to-text

3 課題について，自動抽出された流暢性特徴による採点性能はどれも手動算出されたものより低いものの，統計的な有意差 ($p > 0.1$) はなかった．一方で，提案法の精度は，発話課題による言い淀みやポーズの発生しやすさに影響を受けることが示唆された．絵描写課題において，手動抽出された流暢性特徴より提案法は統計的に有意 ($p < 0.01$) に低い予測性能であることがわかった．また，絵描写のような発話課題では，十分な言語能力を習得していなくても目的の達成のために何らかの情報を伝える必要があることから，語彙検索等によるポーズや発言の修正による言い淀みが多くなると，議論されている [14]．発話中に言い淀みやポーズが増えてしまうことで，流暢性特徴抽出の過程で無視できない程の誤差が蓄積され，流暢性の採点性能に影響を与えてしまったと考察する．

4.3 言い淀み自動除去の妥当性に関する分析

言い淀みの自動除去の妥当性検証のため，言い淀み除去をする場合，しない場合で 2 つの調音速度を自動算出した．提案法の上限値として人手で発話の書き起こしと言い淀みの除去を行い，調音速度を計算し，自動算出した値との比較を行った．自動算出した調音速度は，手動算出した調音速度との相関係数と平均二乗誤差で評価した．分析の結果を表 3 に示す．言い淀みの自動除去を行った調音速度と手動算出された値との相関が有意に高くなる ($p < 0.01$) こと，平均二乗誤差が小さくなることから，言い淀み自動除去の有効性を確認できた．

5 まとめ

妥当性の高い第二言語発話の流暢性自動採点を目的に，言い淀み除去とポーズ位置の節内・節間分類を組み込んだ特徴抽出方法を提案し，その有効性について調査した．実験の結果，複数の独話課題において，人手で算出したものと同等の採点性能を達成できることから，提案する流暢性特徴抽出の有効性を確認できた．ただし，言い淀みやポーズが多く発生しやすい発話課題においては，提案法の精度が下がることがわかった．詳細分析からは，言い淀み自動除去の妥当性，つまり，手動算出された調音速度との誤差を言い淀みの自動除去によって小さくできることを確認できた．今後は，言い淀みやポーズが多い発話においても，妥当性の高い流暢性自動採点が実現できるような流暢性特徴抽出方法について検討する予定である．

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) の結果得られたものです．また，NTT 人間情報研究所に日本人英語音声認識エンジンをご提供いただきました．

参考文献

[1] R. Ellis & G. Barkhuizen. “Analysing learner language”. 2005.

[2] J. Kahng. “The effect of pause location on perceived fluency”. In *Applied Psycholinguistics*, pages 569–591, 2018.

[3] S. Suzuki *et al.* “The relationship between utteranceand perceived fluency: A meta-analysis of correlational studies”. In *The Modern Language Journal*, pages 435–463, 2021.

[4] D. Jacob *et al.* “BERT: Pre-training of deep bidirectional transformers for language understanding”. In *NAACL HLT*, pages 4171–4186, 2019.

[5] N. H. De Jong & H. R. Bosker. “Choosing a threshold for silent pauses to measure second language fluency”. In *Proc. of DiSS 2013*, pages 17–20, 2013.

[6] J. Cenoz. “Pause and communication strategies in second language speech”. In *ERIC*, 1998.

[7] J. Kahng. “Exploring utterance and cognitive fluency of l1 and l2 english speakers: Temporal measures and stimulated recall”. In *Language Learning*, pages 809–854, 2014.

[8] N. H. De Jong. “Predicting pauses in l1 and l2 speech: the effects of utterance boundaries and word frequency”. In *International Review of Applied Linguistics in Language Teaching*, pages 113–132, 2016.

[9] R. L. Rose. “Fluidity: Real-time feedback on acoustic measures of second language speech fluency”. In *Proc. of the International Conference on Speech Prosody*, pages 774–778, 2020.

[10] N. H. De Jong *et al.* “PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically”. In *Assessment in Education: Principles, Policy & Practice*, pages 456–476, 2021.

[11] L. Chen *et al.* “Automated scoring of nonnative speechusing the speechrater^{SM} v.5.0 engine”. 2018.

[12] L. Chen & S. Yoon. “Application of structural events detected on asr outputs for automated speaking assessment”. In *Proc. of INTERSPEECH2012*, pages 767–770, 2012.

[13] J. M. Linacre. “Many-facet rasch measurement”. 1989.

[14] S. Suzuki & J. Kormos. “The multidimensionality of second language oral fluency: Interfacing cognitive fluency and utterance fluency”. In *Studies in Second Language Acquisition*, in press.

[15] M. McAuliffe *et al.* “Montreal forced aligner: Trainable text-speech alignment using kaldi”. In *Proc. of INTERSPEECH2017*, pages 498–502, 2017.

[16] C. Manning *et al.* “The stanford corenlp natural language processing toolkit”. In *Proc. of ACL (System Demonstrations)*, pages 55–60, 2014.

[17] V. Zayats *et al.* “Disfluencies and human speech transcription errors”. In *Proc. of INTERSPEECH2019*, pages 3088–3092, 2019.

[18] V. Sanh *et al.* “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”. In *arXiv*, 2020.