日本人小学生による英語暗唱音声における語彙強勢位置の自動推定と母語話者評価 *

☆木村 竜也，Sylvain Coulange，加藤 恒夫 (同志社大)

1 はじめに

英語における語彙強勢は，発声中の単語分割と語義曖昧性解消の機能を果たし，聞き取りやすさに影響を与える．第二言語 (L2) 英語学習者が聞き取りやすい発声をするうえで正しい語彙強勢は重要な役割を果たす．しかし，母語 (L1) の韻律規則の影響を受けることもあり，簡単ではない．

英語学習者が聞き取りやすい発声に向けて効果的な訓練とフィードバックを受けられるように，Coulange らは最新の音声認識技術を活用し，英語自発発話におけるプロミネンスとポーズ位置を自動的に検出，分析するパイプライン処理を提案している [1][2]．これまでフランスの英語検定試験におけるフランス人学習者の自発発話に同手法を適用し，英語音声を分析してきたが，母語の異なる学習者の英語音声への適用と母語話者による評価との比較は行われていなかった．そこで，本研究では日本人学習者による少量の英語暗唱音声に提案手法を適用し，語彙強勢位置の観点で英語母語話者による評価と比較する．

ただし，英語母語話者による語彙強勢の評価は絶対の正解と思われがちであるが，必ずしもそうではない．菅原らは，同一語幹で接尾辞の異なる 21 種類の英単語対（例 domination vs. dominating）の発声から共通の語幹部分を切り出して強勢位置を推定する実験を，英語母語話者，日本語母語話者，韓国語母語話者を対象に行ったところ，英語母語話者は語頭音節に強勢を知覚する傾向を示した [3]．英語の語彙の約 60%，各種の自発発話における頻度であれば約 90% が語頭音節に第 1 強勢を持っており [4]，英語は語頭強勢に偏った言語といえる．母語話者でも，正確に強勢を知覚できる人と語彙の知識に影響を受けて強勢を知覚する人の間でばらつきを生じる可能性がある．

本研究では，日本人小学生による英語暗唱音声における語彙強勢位置について，Coulange らのパイプライン処理による自動推定と英語母語話者 10 名による評価の相関を分析する．さらに，母語話者評価においては，発音の正確さ，流暢さ，表現の豊かさ，聞き取りやすさについても評価してもらった．総合的な評価である聞き取りやすさを目的変数とする重回帰分析により聞き取りやすさに寄与する要因を分析する．

2 英語語彙強勢の自動推定と母語話者評価

2.1 英語語彙強勢位置推定パイプライン

Coulange らが提案したパイプライン処理 (Ver. MFA)[1]は，英語自発発話に含まれる多音節語の語彙強勢推定とポーズ検出を行う．自発発話を対象とするためテキストは用意しない．

分析対象の英語音声に対して，まず WhisperX [5] を用いて音声認識を行う．次に，音声認識結果をテキストとして Montreal Forced Aligner (MFA) [6] を実行し，音素と単語のアライメントを得る．並行して praat スクリプト [7] を実行し，音節核を検出する．また，音声認識結果に対して spaCy[8] を適用し，各単語の品詞を推定する．WhisperX による音声認識，MFA による音素アライメントには誤りが含まれるため，認識結果テキストの各単語の音節数を CMU 発音辞書 [9] から取得し，多音節語については同単語区間で検出された音節核の数が CMU 発音辞書の音節数と一致するかチェックする．一致しない多音節語は誤認識の可能性が高いとして分析対象から除外する．

分析対象とした多音節語については語彙強勢推定のための特徴量として音節毎の F0，インテンシティ，継続時間を求める．具体的には，MFA の音素アライメントをもとに praat を用いて各母音区間の平均 F0，インテンシティ最大値を測り，継続時間は母音区間の時間長とする．各特徴量に対して話者正規化を行う目的で，話者毎に計測した F0，インテンシティ，継続時間の分布を求め，分布に基づき計測値をパーセンタイルに変換する．こうして各音節について，F0，インテンシティ，継続時間のパーセンタイル値が得られる．語彙強勢位置の推定は，各音節について F0，インテンシティ，継続時間のパーセンタイル値を平均した値を音節間で比較し，最大値をもつ音節を強勢位置とする．最後に，推定された強勢位置を CMU 発音辞書の規範的な強勢位置と比較して，強勢位置の正誤を判定する．このとき，CMU 発音辞書に複数の強勢位置がある場合には，推定された品詞により選択して参照することがある．その他，ポーズ位置を分析するために Berkeley Neural Parser[10] を用いた構文解析も行うが，本稿では割愛する．

---

* Automatic estimation and native speakers' evaluation of lexical stress positions in English recitation speech produced by Japanese elementary school children. by KIMURA Tatsuya, COULANGE Sylvain, KATO Tsuneo (Doshisha University)

2.2 英語母語話者評価

暗唱音声を英語を母語とする評価者に聞かせて，語彙強勢位置を判定してもらう．暗唱テキストを印刷した紙を評価者に渡し，評価者は暗唱音声をヘッドフォンで聞きながら，多音節語の第 1 強勢があると感じられた位置にペンでアクセント記号を記入する．多音節語は評価者に分かりやすいように太字で印刷されている．強勢位置が聞き取れないもしくは分からない場合には，単語の上に横棒線をひいてもらう．以下に例を示す．アクセント記号と横棒線は実際には手書きである．

> He had **sérious concérns** about the **fúture** of **Japán** and **réalized** the **impórtance** of **stúdying** in **Wéstern culture**

得られた語彙強勢位置について，自動推定と同様に CMU 発音辞書の規範的な強勢位置と比較し，正誤を判定する．横棒線が引かれて強勢位置を判断できないとされた単語は non と分類する．つまり，多音節語は正，誤，non の 3 種類に判定される．

各暗唱音声の語彙強勢評価の後，以下の 4 項目についても 5 段階で評価してもらう．

- Accuracy of pronunciation (発音の正確さ)
- Fluency (流暢さ)
- Expressiveness (表現の豊かさ)
- Easiness of understanding (聞き取りやすさ)

以上について，評価者によるばらつきを評価するため，複数の評価者に同じ作業をしてもらう．

2.3 母語話者評価と自動推定結果の比較分析

先ず，母語話者評価における強勢位置誤り検出率のばらつきを確認する．強勢位置誤り検出率とは，母語話者評価，自動推定それぞれにおいて強勢位置誤りと評価された単語が全対象単語に占める割合である．次に，両手法の強勢位置誤り検出率を比較する．

続いて，パイプライン評価における第 1 強勢とそれ以外の音節のパーセンタイル値のコントラスト（以下，第 1 強勢コントラストと呼ぶ）と母語話者評価における正解判定の割合の相関を調べる．

パイプライン処理における第 1 強勢コントラストを次式で定義する．

$$c_{\mathrm{E1}} = \frac{z_{\mathrm{E1}}}{z_{\mathrm{E1}} + mean(\{z_{\mathrm{other}}\})} \quad (1)$$

[1]https://gricad-gitlab.univ-grenoble-alpes.fr/lidilem/plspp

ここで，z_{E1}，z_{other} はそれぞれ発音辞書の第 1 強勢音節とそれ以外の音節におけるパーセンタイル値を表す．規範的な第 1 強勢音節に強勢が推定された場合，c_{E1} は 0.5 以上の値を取る．

母語話者評価における正解判定の割合は，評価対象の各単語について non 評価を除外した正解判定の割合である．つまり，誤りと判定した評価者がいなければ 1，正解と判定した評価者がいなければ 0 となる．

正しい強勢が明確な場合には，パイプライン処理における第 1 強勢コントラスト，母語話者評価における正解判定の割合ともに 1 に近づき，誤った強勢が明確な場合には，両者ともに 0 に近づくと考えられる．強勢位置が曖昧な場合には，第 1 強勢コントラストは 0.5 に近づき，母語話者評価も評価者間のばらつきが大きくなるために中間的な値になると考えられる．

2.4 聞き取りやすさに対する諸要因の重回帰分析

母語話者評価における全体的な評価 4 項目のうち，聞き取りやすさは，話し手の意図が聞き手にどれだけ明確に伝わるかを測る総合的な評価項目である．そこで，聞き取りやすさに対して，強勢の正確さと，聞き取りやすさ以外の 3 種類の指標が与える影響の大きさを重回帰分析を用いて分析する．聞き取りやすさの評価値を目的変数とし，強勢の正確さに加えて，発音の正確さ，流暢さ，表現の豊かさの 3 種類の評価値を説明変数とする．強勢の正確さは，各暗唱音声に対する各評価者の語彙強勢評価において non 評価を除外した正解判定の割合を用いる．

3 実験

3.1 日本人小学生による英語暗唱音声データ

分析対象とした英語音声は，英語暗唱大会に向け，京都府内にある私立小学校の 4 年生から 6 年生の中から選抜された児童 6 名による英語暗唱音声である．暗唱文は全員共通で，歴史的人物の伝記の一節であり，6 つの段落，計 302 語から構成される．児童らは十分に練習を積んでおり，概して流暢に暗唱している．

3.2 実験方法と条件

パイプライン処理は，MFA のアライメント誤りを低減するために，各暗唱音声を段落毎に分割して実行した．

母語話者評価は，日本在住の米国人 10 名に行ってもらった．全員が日本に 5 年以上居住して日本人の英語には慣れている．ヘッドフォンを装着し，自身で PC 上のオーディオソフトを操作して音声を再生する．一時停止や巻き戻し再生は簡単に操作でき，自由にしてよいとした．

Table 1: 各暗唱音声における評価対象語数

	Sp1	Sp2	Sp3	Sp4	Sp5	Sp6
(a) パイプラインで検出された多音節語数 (パイプラインの評価対象語数)	55	48	49	34	45	46
(b) うち認識結果が正しかった多音節語数 (母語話者評価との比較対象語数)	51	45	48	33	44	37

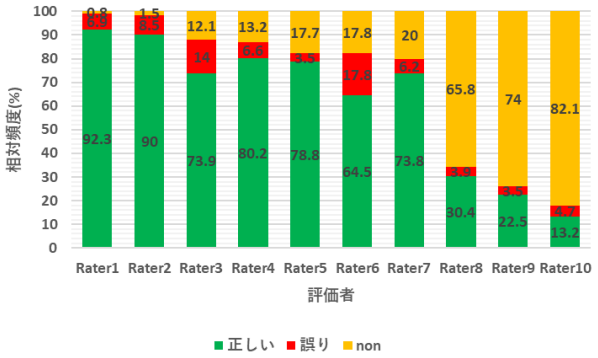

Fig. 1: 母語話者評価における評価者別の語彙強勢判定の相対頻度分布 (緑：正解，赤：誤り，黄：non)

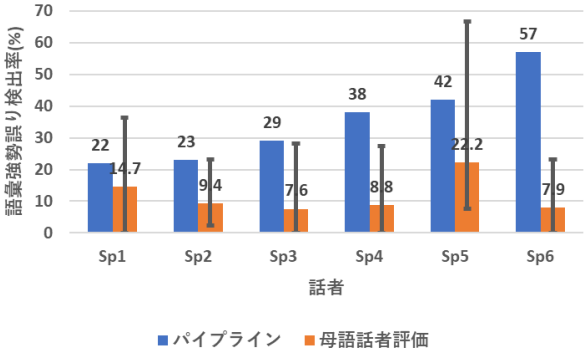

Fig. 2: 暗唱音声別の語彙強勢位置誤り検出率の比較 (%)(青：パイプライン，オレンジ：母語話者評価)

3.3 英語母語話者による語彙強勢位置判定の結果

先ず，パイプライン処理，母語話者評価それぞれにおける評価対象語数を Table 1 に示す．パイプライン処理では，音節数の一致によるスクリーニング処理を行っているが，誤認識単語が少数含まれている．母語話者評価との比較においては，誤認識単語を除外したため，対象語数が減少している．

英語母語話者 10 名の語彙強勢評価における 3 分類の割合を Fig. 1 に示す．緑が正解，赤が誤り，黄色が non を表す．non の割合で昇順にソーティングした．non の割合に大きなばらつきが見られ，non をほとんどつけない評価者 2 名，non を 10〜20%つける評価者 5 名，半数以上に non をつける評価者 3 名の 3 群に分かれた．non を除外すると，各評価者とも誤りの判定は 20%未満で，平均で 11.8%であった．

評価者間の一致度を Fleiss の κ 係数で測ると 0.43 となり中程度の一致を示した．さらに，non の割合が高い評価者 3 名を除くと κ 係数は 0.61 に上昇し，実質的な一致を示した．また，non の割合が低い 2 名の間の Cohen の κ 係数は 0.80 で，高い一致を示した．

3.4 パイプラインによる語彙強勢位置推定の結果

Fig. 2 に，パイプライン処理と母語話者評価の暗唱音声別の語彙強勢誤り検出率を示す．青がパイプライン処理，オレンジが母語話者評価を表し，パイプライン処理の誤り検出率で昇順にソーティングしている．母語話者評価における語彙強勢誤り検出率は，non 評価を除外して算出している．また，評価者間のばらつきを表すために，上端を最大値，下端を最小値とするエラーバーを表示している．

パイプライン処理による語彙強勢位置推定では，Table 1 (a) に該当する計 277 語のうち正解判定の単語が 65.7%，誤り判定の単語が 34.3%を占めた．母語話者評価の強勢誤り検出率と比較するとパイプライン処理の方が高い．パイプライン処理の強勢誤り検出率が高い一因として，音声認識のアライメントエラーの影響は無視できない．例えば，Sp6 は他の暗唱音声に比べて両手法の強勢誤り検出率の差が大きいが，Table 1 を参照するとパイプラインの評価対象語数と母語話者評価との比較対象語数の差が大きく，パイプライン処理においてアライメントの誤りが多数混入していることがわかる．Sp6 の暗唱音声を確認すると，言い直しが多い．実際に発声された単語とは別の単語の規範的な強勢位置と比較することになるため，強勢誤り検出率は高くなってしまう．その他の要因として，母語話者評価では強勢位置を判断できない場合のための non 評価を設けているのに対して，パイプライン処理では必ずいずれかの音節に強勢を推定するため，強勢誤り検出率が高くなる可能性が考えられる．いずれにしてもパイプライン処理の方が強勢位置の判定は厳しいといえる．

3.5 母語話者評価と自動推定評価の相関分析の結果

Table 1 (b) に該当する計 258 語について，式 (1) で算出されるパイプライン処理における第 1 強勢コントラストを横軸，母語話者評価における正解判定の割合を縦軸にした散布図を Fig. 3 に示す．

Pearson の積率相関係数は 0.29 で，弱い正の相関が見られた．単純に，パイプライン処理の第 1 強勢コントラスト，母語話者評価における正解判定の割合の 2 軸において 0.5 に閾値を設定し，閾値以上を正解，

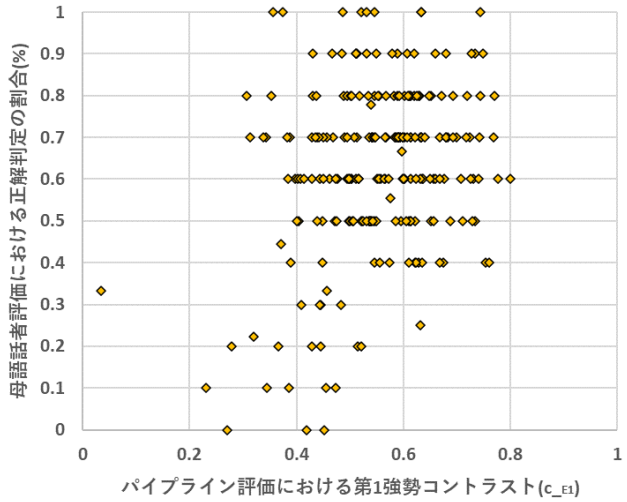

Fig. 3: パイプライン評価における第 1 強勢コントラストと母語話者評価における正解判定の割合の分布

閾値未満を誤りとすると，i) 両手法ともに正解，ii) 両手法ともに誤り，iii) パイプライン処理は正解かつ母語話者評価は誤り，iv) パイプライン処理は誤りかつ母語話者評価は正解，の 4 クラスの度数はそれぞれ，164, 25, 17, 52 となった．両手法の一致率は 73.3%，Cohen の κ 係数は 0.27 であった．

3.6　聞き取りやすさに対する重回帰分析の結果

6 つの暗唱音声に対する英語母語話者 10 名による評価値を用いて聞き取りやすさを目的変数とする重回帰分析を行った．発音の正確さと表現の豊かさの変数間の相関係数が 0.74 と高く，多重共線性が見られたため，表現の豊かさを除いた強勢の正確さ，発音の正確さ，流暢さの 3 種類を説明変数とした．結果を Table 2 に示す．

標準偏回帰係数が高い方から，流暢さ，強勢の正確さ，発音の正確さとなり，流暢さ，強勢の正確さの p 値は有意水準 5%以下であった．流暢さが示すものは曖昧であるが，基本的には適切な位置にポーズを置いて淀みなく発声することと考えられる．次に，強勢の正確さであった．

4　おわりに

本研究では，日本人小学生による英語暗唱音声に対して語彙強勢位置自動推定パイプラインを適用するとともに英語母語話者 10 名に語彙強勢位置を判定してもらい，両者の結果を比較分析した．まず，母語話者評価においては non の割合，即ち語彙強勢の知覚判断に評価者間の大きなばらつきが見られた．比較の結果，パイプライン処理による強勢誤り検出率は平均 34.3%で，母語話者評価の平均 11.8%よりも高かった．パイプライン処理による第 1 強勢コントラスト

Table 2: 聞き取りやすさに関する重回帰分析の結果

変数	標準偏回帰係数	t 値	p 値
流暢さ	0.0630	5.371	1.6E-06*
強勢の正確さ	0.0223	2.423	0.0187*
発音の正確さ	0.0166	1.075	0.2871

* $p \leq 0.05$

と母語話者評価による正解判定の割合の間で Pearson の積率相関係数を測ると，0.29 と弱い正の相関がみられた．単純な閾値処理により，iv) パイプライン処理で誤りかつ母語話者評価で正解とされた 52 単語について，今後まず特徴を分析する予定である．

さらに，母語話者評価における聞き取りやすさの評価値を目的変数とする重回帰分析により，影響を与える要因を分析した結果，流暢さと強勢の正確さの寄与が大きかった．

謝辞　英語暗唱音声データを提供頂いた同志社小学校に感謝致します．本研究は科研費 23H00648 の助成を受けたものである．

参考文献

[1] S. Coulange, 加藤, S. Rossato, M. Masperi, “フランス人学習者による自発英語発話における語彙アクセント自動測定,” 第 37 回日本音声学会全国大会, pp. 126-131, 2023.

[2] S. Coulange, T. Kato, S. Rossato, M. Masperi, “Enhancing Language Learners’ Comprehensibility through Automated Analysis of Pause Positions and Syllable Prominence,” Languages, in press.

[3] M. Sugahara, ”Is the Perception of English Stress by Japanese Listeners Influenced by the Distribution of Accent in their L1? In the Case of Truncated Word Stimuli,” Doshisha Studies in English, Vol. 97, pp. 59-118, 2019.

[4] A. Cutler, D. M. Carter, “The Predominance of Strong Initial Syllables in the English Vocabulary,” Computer Speech and Language, Vol.2, pp. 133-142, 1987.

[5] M. Bain, J. Huh, T. Han, A. Zisserman, “WhisperX: Time-accurate speech transcription of long-form audio,” Interspeech 2023, pp. 4489-4493, 2023.

[6] M. McAuliffe, M. Socolof, S. Mihuc, M, Wagner, M. Sonderegger, “Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi,” Interspeech 2017, pp. 498-502, 2017.

[7] N. H. de Jong, J. Pacilly, W. Heeren, “Praat scripts to measure speed fluency and breakdown fluency in speech automatically,” Assessment in Education: Principles, Policy, and Practice, Vol. 28, pp. 456-476, 2021.

[8] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, “spacy: Industrial-strength natural language processing in python”

[9] The CMU Pronunciation Dictionary, ”http://www.speech.cs.cmu.edu/cgi-bin/cmudict”

[10] N. Kitaev, S. Cao, and D. Klein, “Multilingual constituency parsing with self-attention and pre-training,” ACL 2019, pp. 3499–3505, 2019.