識別モデルを用いた英文読み上げ音声中の弱・強勢自動評価*

☆加藤集平, 鈴木雅之, 峯松信明, 広瀬啓吉 (東大), 山内豊 (東京国際大), 西川惠 (東海大)

1 はじめに

学習者が英語の自然な発音を身につけるためには，/l/と/r/の区別に代表される分節的特徴以外にも，強勢やリズム，イントネーションといった韻律的特徴を修得する必要がある．どれも重要な要素ではあるが，本研究では強勢，特に文強勢に注目する．文強勢は自然な英語リズムを形成するほか，強調などの話者の意図が表現されるため，母語話者と十分な意思疎通を行う上で重要な要素だからである [1]．しかし，現場で学習者音声を逐一手動評価するのは膨大な労力を要するため，音声情報処理技術を用いた自動評価の研究が行われてきた [2–4]．評価対象の音声は研究によって様々であるが，[2] は，「日本人学生による読み上げ英語音声データベース」(ERJ (English Read by Japanese) データベース) [5] の「文強勢，文リズムに関する文」(以下リズム文とする) に対する読み上げ音声を用いている．この音声の特徴は，英語の強勢・リズムに着目して構成された文を，原稿に記載された強勢記号に従って学習者自身が読み上げたもので，学習者自身は正しく読めたと考えている音声となっている点である．読み上げ文が英語の強勢・リズムに着目して作られている点，学習者自身が正しく読めたと考えている音声である点で教育的な観点から評価対象としてふさわしく，読み上げ音声という点で技術的にも扱いやすい．ただし，[2] では HMM を使って評価しているため，特徴量を増やしたりすることに限界がある．

本研究では ERJ のリズム文の読み上げ音声を評価対象として，識別モデルを用いてより精度の高い文強勢の自動評価を目指す．具体的には，評価対象音声に対して音節ごとに強勢の度合いを高精度に自動評価することを目的とする．ただし，ERJ には音節ごとに強勢の度合いを評価したラベルは付与されていないため，手動評価という形でラベルを付与した．また，本研究では読み上げ音声を使用するため，評価対象音声の発話内容は既知である．そこで，音響特徴量以外の，テキストから得られる特徴量も積極的に使って評価精度を上げることを目指した．

2 音声試料

評価対象の音声試料は，ERJ のリズム文である．ERJ は，日本語を母語とする大学生・大学院生相当の学生による読み上げ音声データベースである．リズム文の読み上げ原稿には，音節ごとに強勢の度合いを示す記号が 3 段階（弱勢，強勢，核強勢）で付けられている．ERJ の音声収録にあたっては，発声者は事前の発声練習が許されており，収録も発声者自身が「正しい」と判断できる発声が得られるまで繰り返されている．

ERJ のリズム文は全部で 120 文あるが，本研究では，そのうち第五著者（英語教師）が「発声者によって出来に差がつきやすい」と判断し選んだ 22 文を使用した．この 22 文について，すべての発声を評価対象とした．結果，評価対象となったのは 792 発声（8,494 音節），話者 202 名（男性 100 名，女性 102 名）である．

以下の実験においては，全 8,494 音節を話者が重複しないようにほぼ 5 等分に分割し，4 つを学習データ，残りの 1 つをテストデータとした．

3 手動評価による強勢ラベリング

3.1 評価者

評価者は 2 名である．英語音声学を専攻している大学院博士課程の学生（以下評価者 A）と第六著者（英語教師，以下評価者 B）である．両者で事前に評価基準の統一をとることなく，それぞれ独立に評価を行った．

3.2 手動評価タスク

実験で用いる発声を聴取して，音節ごとに強勢の度合いを評価した．音声収録時の強勢記号は 3 段階（弱勢，強勢，核強勢）であるので，各々に対して，より弱い／適切／より強い，の 3 段階を考慮し，合計 9 段階とし（1 が最も弱く，9 が最も強い），また，「発声されていない」ことを示す 0 を加えた．本稿ではスペースの関係上，9 段階を 3 段階に丸めて扱うことにする（1 から 3 は “弱勢” としてひとくくりにする．以下同様）．

全 8,494 音節に対する手動評価の結果を Table 1 に示す．参考のため，収録時に読み上げ原稿に呈示された強勢記号の個数も併記しておく．

また，手動評価の安定性を検証するために，評

*Automatic evaluation of (un)stressed syllables in read sentences using discriminative models. by Shuhei KATO, Masayuki SUZUKI, Nobuaki MINEMATSU, and Keikichi HIROSE (Univ. of Tokyo), Yutaka YAMAUCHI (Tokyo International Univ.), Megumi NISHIKAWA (Tokai Univ.)

Table 1　手動評価の結果

評価	0（発声されていない）	1–3（弱勢）	4–6（強勢）	7–9（核強勢）
評価者 A	26	4,861	3,101	506
評価者 B	31	5,022	2,158	1,283
（原稿の強勢記号）	0	5,504	1,395	1,595

Table 2　1 回目の評価と再評価の一致率および相関

評価者	一致率	相関
評価者 A	0.871	0.802
評価者 B	0.868	0.874

Table 3　評価者間の評価の一致率および相関

一致率	相関
0.773	0.776

価対象である 792 発声のおよそ 1 割にあたる 79 発声を選び，数週間の時間を置いて再評価を行った．再評価対象発声の選定は，1 回目の評価順（792 回の評価作業で何番目に評価を行ったか）のバランスを考慮して，準ランダムに行った．評価条件は 1 回目と同様であるが，1 回目に自分が下した判定は参照させていない．1 回目の評価と再評価の一致率および相関を Table 2 に示す．

両評価者は事前に評価基準の打ち合わせ無しに評価を行ったが，評価者間の（1 回目の）評価の一致率および相関を Table 3 に示す．識別モデルを用いた自動評価を試みるが，その精度評価は，Table 2 や Table 3 が比較対象となる．

4　先行研究の手法を用いた実験

第 1 節で，先行研究として HMM を用いた手法を紹介した [2]．本研究ではベースラインとして，[2] と同様の手法を実装し，実験を行った．

[2] では，音節の強勢／弱勢を HMM でモデル化し，入力文発声の各音節に対して HMM を用いて強勢／弱勢の判定を行っている．本研究でもこれにならい，弱勢／強勢／核強勢の 3 クラスのモデルを構築した（手動評価が 0 であったものについては，本実験では使用しなかった）．特徴量は SPTK を用いて抽出した [6]．音響分析条件を Table 4，HMM の学習条件を Table 5 に示す．

特徴量も [2] にならった．F_0 は，SPTK によって抽出した F_0 を母音区間だけ残して対数をとったものを線形補間する形で求めた．文頭（末）の無声区間の F_0 は，母音区間の F_0 の対数をとったものから平均 F_0 と標準偏差 σ を求め，母音区間から平均音節時間長前（後）に，平均 $F_0 - 3\sigma$ の値（基本周波数生成過程モデルにおける基底周波数 F_b に相当）にあるものとして線形補間を行った．さらに，F_0 とエネルギーに対しては，当該発声の平均値を引く形で正規化を行ったものを最終的な特徴量として用いた．

Table 4　音響分析条件

標本化周波数	16kHz
量子化ビット数	16bit
抽出周期	8.0ms (96 samples)

Table 5　HMM の学習条件

特徴量	対数 F_0 + Δ対数 F_0, エネルギー+Δ エネルギー, MFCC+ΔMFCC（各 1〜4 次元）
トポロジー	6 状態 4 分布，left-to-right
分散行列	3 種類の特徴量に対して個別に 全共分散行列を算出

Table 6　先行研究の手法を用いた実験の結果

学習データ	テストデータ	正解率
評価者 A	評価者 A	0.451
評価者 B	評価者 B	0.446

5　先行研究の手法を用いた実験の結果

実験結果を Table 6 に示す．評価者 A，評価者 B いずれのデータを用いた場合でも，正解率はほぼ等しくなった．Table 1 に示すように，強勢，弱勢の出現には数的偏りがある．先行研究ではこれらの事前分布を使わずに評価を行っており，Table 6 の結果は，常に “弱勢である” と答えた場合よりも，正解率が劣る結果となった．

6　提案手法を用いた実験

提案手法では，SVM を用いて強勢の度合いを判定する．具体的には，R の kernlab パッケージに含まれる ksvm を用いた [7]．カーネルは RBF カーネル，ハイパーパラメータは ksvm のデフォルトのものを用いた．手動評価が 0 のデータはここでも使用しなかった．使用した特徴量セットを Table 7 に示す．

特徴量セット a に含まれるのは音響特徴量で，音節の母音部の平均 F_0，継続長，平均エネルギー，平均 MFCC（12 次元），およびそれらの差または比を前後 2 音節に対して計算している．このような特徴量を用いたのは，1) 強勢が，高さ，長さ，強さ，音質の複合的な変化により実現されることと，2) 強勢とは，他の音節よりも際立って聞こえることを指すからである [8, 9]．

特徴量セット b に含まれるのは，音節が持つコンテキストに関するシンボリックな特徴量である．これらは，どのような音節に文強勢が置かれやすいかという観点から選んでいる [8–10]．

特徴量セット c に含まれるのは，ERJ の原稿に記載されていた強勢記号に関する特徴量である．発声者は原稿に記載されていた強勢記号の

Table 7　提案手法で用いた特徴量セット

特徴量セット	特徴量
a（音響特徴量）	母音部平均対数 F_0 ／母音部継続長／母音部平均エネルギー／母音部平均 MFCC（12 次元）（いずれも発話ごとに正規化），母音部平均対数 F_0 ／平均エネルギーの差（前後 2 音節まで），母音部継続長の比（前後 2 音節まで）
b（コンテキスト）	音節中の母音が単母音か否か，フレーズ頭の単語（最初の 1 単語）／フレーズ中の単語／フレーズ末の単語（最後の 1 単語）中の音節である，内容語中／機能語中の音節である，どの品詞中の音節か（名詞／代名詞／動詞／形容詞／副詞／前置詞／接続詞／間投詞），単語を単独発声したときに第 1 強勢／第 2 強勢／弱勢で発声される
c（原稿の強勢記号）	収録時に参照した強勢記号（弱勢／強勢／核強勢）
a-b	（a と b を合わせたもの）
a-c	（a と c を合わせたもの）
b-c	（b と c を合わせたもの）
a-b-c	（a，b，c すべてを合わせたもの）

とおりに発声できていると考えているため，有用な情報であることが期待される．

特徴量セット b，c には，音響的な情報が含まれておらず，ある読み上げテキストのある音節に対しては，どの発声者に対しても同じ特徴量となることに注意されたい．なお，特徴量セット b，c に関しては，各特徴量を 0/1 の二値で表現した（“X である” という特徴量は，X である場合は 1，X でない場合は 0 をとる）．最終的に，a，b，c のうち 2 つを組み合わせた特徴量セット，すべてを組み合わせた特徴量セットを用意した．

評価は，手動評価の評価者クローズドの場合と，評価者オープンの場合の両方を行った．すなわち，評価者クローズドの場合は評価対象の全 8,494 音節を発声者が重複しないように 5 分割したもののうち 4 つで学習，残り 1 つでテストするという作業を 5 回繰り返し平均をとった．評価者オープンの場合は一方の評価者のデータで学習し，他方の評価者のデータでテストを行った．

7　提案手法を用いた実験の結果

提案手法を用いた実験の結果を Table 8 および Table 9 に示す．

まず，各特徴量セットに対する結果を比較する．学習データとテストデータの組み合わせがいずれの場合でも，最も正解率および相関が低かったのは特徴量セット a の場合で，最も高かったのは特徴量セット a-b-c の場合であった．また，Table 6 に示した先行研究の精度を大幅に上回る結果となった．

特徴量セット a，b，c をそれぞれ単独で用いた場合の結果については，特徴量セット a のみを用いた場合が正解率，相関とも最も低かった．特徴量セット b と c に関しては，正解率では特徴量セット b が上回ったものの，相関では学習データとテストデータの組み合わせによって結果がばらついた．2 つの特徴量セットを組み合わせた場合 (a-b, a-c, b-c) の結果については，特徴量セット a，b，c をそれぞれ単独で用いた場合の結果と比べて正解率，相関ともおおむね向上しており，a-b の組み合わせが最も精度が高かった．

次に，すべての特徴量セットの中で最も高精度であった特徴量セット a-b-c を用いた場合の結果と，手動評価の結果を比較する．評価者クローズドの結果については Table 2 と比較することになるが，正解率について，どちらの評価者のデータを用いた場合でも，手動評価の一致率を少し下回る結果となった．相関についても，手動評価の結果には及ばなかった．評価者オープンの結果については Table 3 と比較することになるが，正解率について，どちらの評価者のデータを学習データとした場合でも，手動評価の一致率を若干上回る結果となった．相関については平均すると 0.771 で，手動評価の結果に近い結果が得られた．

8　考察

特徴量セット a，b，c をそれぞれ単独で用いた場合は，特徴量セット a（音響特徴量）を用いた場合が最も精度が低かった．入力音声に関する音響的情報が含まれているのが特徴量セット a のみであることを考えると，(ERJ の発声者である) “日本語を母語とする大学生・大学院生相当の学生” という集団に対しては，発声者がどのような発声をしたかにかかわらず，音節の持つコンテキストあるいは原稿の強勢記号から推定した評価結果を呈示したほうが，発声の情報のみから推定するよりも精度が高いということになる．すなわち，強勢記号が記載された原稿を渡されて，発声者が「正しい」と思って読み上げた場合，多くの発声者が似たようなパターンで読み上げることになることが予想される．これは逆に，非日本語を母語とする話者の読み上げ音声を対象にした場合は，ミスマッチが起こることも予想される．

また，本実験では発話者オープンにデータセットを分割したが，文オープンとはなっていないことに注意したい．音節が持つコンテキスト情報

Table 8　提案手法による実験の結果（正解率）

学習データ	テストデータ	a	b	c	a-b	a-c	b-c	a-b-c
評価者 A	評価者 A	0.735	0.821	0.815	0.827	0.814	0.824	0.831
評価者 B	評価者 B	0.702	0.782	0.781	0.805	0.807	0.794	0.813
評価者 A	評価者 B	0.705	0.768	0.765	0.775	0.770	0.771	0.776
評価者 B	評価者 A	0.729	0.771	0.760	0.782	0.775	0.764	0.780

Table 9　提案手法による実験の結果（相関）

学習データ	テストデータ	a	b	c	a-b	a-c	b-c	a-b-c
評価者 A	評価者 A	0.557	0.717	0.708	0.739	0.705	0.729	0.748
評価者 B	評価者 B	0.594	0.765	0.782	0.789	0.790	0.784	0.801
評価者 A	評価者 B	0.659	0.790	0.788	0.807	0.793	0.802	0.810
評価者 B	評価者 A	0.622	0.707	0.734	0.735	0.713	0.702	0.732

を特徴量とする特徴量セット b を用いた場合に精度が高かったのは，文クローズドであることが影響した可能性がある．

2 つの特徴量セットを組み合わせた場合 (a-b, a-c, b-c) の結果については，a-b の組み合わせが最も高精度であった．音声の音響情報を加えることで，精度が向上したものと考えられる．一方，b-c の組み合わせについては，特徴量セット b，c をそれぞれ単独で用いた場合と比べて，さほど精度が向上しなかった．特徴量セット c で用いている収録時に参照した原稿に記載されていた強勢記号は，ある音声学者によって付与されたものであるが，そのラベリングの際に，特徴量セット b に含まれるようなコンテキスト情報が用いられたことが関係していると考えられる．すなわち，特徴量セット b から得られる情報と特徴量セット c から得られる情報は相関が高く，両者を組み合わせても，精度向上に寄与しなかったと考えられる．

全ての特徴量を用いた場合（a-b-c）は最も高精度であったが，評価者クローズドの手動評価の精度には及ばなかった。これは各々の評価者が，今回検討した特徴量以外の，独自の特徴量に着眼している可能性がある．なお，評価者クローズドの結果と評価者オープンの結果を比べると，一部，評価者オープンのほうがクローズドの場合よりも高精度となっている．Table 2 と Table 3 からも，クローズドの結果が一般に高精度となるはずである．逆の傾向を示した結果に関する十分な考察はできていない．

9　おわりに

本研究では ERJ データベースのリズム文の読み上げ音声を評価対象として，識別モデルを用いて高精度に文強勢の自動評価を行うことを目指し，音節ごとに強勢の度合いを 3 段階で自動評価する実験を行った．その過程で，評価対象音声に対して，手動評価による強勢の度合いを 9 段階でラベリングする作業も行ったが，本稿では 3 段階に丸めて用いた．特徴量セットに関して様々な検討を行い，最高で 83.1%の正解率を得ることができたが，評価者クローズドの手動評価の精度には及ばなかった．

今後の課題としては，特徴量セットや識別器などの改良による精度向上を考えている．また，音節が持つコンテキスト情報を特徴量とする特徴量セット b を用いた場合に精度が高かったのは，文クローズドなデータセットの分割の仕方をしたことが影響した可能性がある．文オープンにした場合にどのように結果が変わるか，実験的に検討する必要がある．さらに，9 段階の手動評価結果を 3 段階に丸めずにそのまま用いた実験を行い検討すること，手動評価の評価基準をより明確なものとした上で評価者を増やし，より信頼のおける手動評価データを構築することを考えている．

参考文献

[1] 根間弘海他，英語とリズムをマスターする英語音声学，英宝社，2000.

[2] N. Minematsu *et al.*, “Acoustic modeling of sentence stress using differential features between syllables for English rhythm learning system development,” *Proc. ICSLP*, 745–748, 2002.

[3] K. Imoto *et al.*, “Modeling and automatic detection of English sentence stress for computer-assisted English prosody learning system,” *Proc. ICSLP*, 749–752, 2002.

[4] C. Li *et al.*, “English sentence stress detection system based on HMM framework,” *Applied Mathematics and Computation*, 185 (2), 759–768, 2007.

[5] 峯松信明他，“英語 CALL 構築を目的とした日本人および米国人による読み上げ英語音声データベースの構築,” 日本教育工学会論文誌，27 (3), 259–272，2003.

[6] Speech Signal Processing Toolkit (SPTK) Version 3.5, http://sp-tk.sourceforge.net/, 2011.

[7] Package ‘kernlab’ Version 0.9-15, http://cran.r-project.org/web/packages/kernlab/index.html, 2012.

[8] A. C. Gimson, *An introduction to the pronunciation of English*, 1962 （竹林滋訳，ギムスン英語音声学入門，金星堂，1983）.

[9] 榎本正嗣，日英語 話し言葉の音声学，玉川大学出版部，2000.

[10] 窪薗晴夫，音声学・音韻論，くろしお出版，1998.