A1

機械学習を用いた日本語アクセント型の分類
－母語話者と学習者による単語発話と朗読発話の比較－

波多野 博顕（国際交流基金日本語国際センター）アルビン エレン・王睿来（神戸大学）
石井カルロス寿憲（株式会社国際電気通信基礎技術研究所）
Hiroaki_Hatano@jpf.go.jp

1. 背景と先行研究

日本語教育では音声評価に自信のない教師が多いことを背景に、学習者が音声教育を受ける機会は少ない。海外のノンネイティブ（NNS）日本語教師の多くは、「自分が知らない」「自信が無い」という理由から、授業で音声を扱っていない（磯村 2001）。しかし、自然な発音で話したいという学習者のニーズは高い（佐藤 1998）。また、ビジネスや国内長期生活では発音が問題となり（小河原 2001）、評価には韻律が与える影響が大きい（佐藤 1995）。ところが、国内のネイティブ（NS）日本語教師であっても、単音等と比較して韻律（アクセント）は音声教育項目としての重要度の認識が相対的に低い（轟木・山下 2009）。このような背景から、自学自習が可能な音声自動評価法の開発が求められている（松崎 2016）。

日本語アクセント（共通語）は単語や文節にかかる韻律特徴であるため、発話音声全体の自然性に大きく影響する。アクセントに注目した研究は多いものの、音響特徴から型の推定を行う研究の多くは 80 %以下の精度に留まっている（石井・他 2001、広瀬 2005、波多野・他 2014、Hatano, et al. 2018）。先行研究では 1 つの特徴量のみを推定に用いているが、実際の発話では「おそ下がり」（杉藤 1969）が生じることもあり、特に連続発話では基本周波数（fo）の動態が不安定な場合が多い。

本研究では、観測した fo に基づいて全体の軌跡を再構築し、そこから抽出した複数の特徴量によって機械学習を行うことで、従来よりも高精度なアクセント型の推定を試みた。

2. データの概要と分析方法

2.1. データの概要

分析には、NS および NNS による単語発話と朗読発話を用いた。NS だけでなく NNS の発話も対象としたのは、音響的に様々な程度で NS と異なる音声に対しても、本手法が頑健かどうか検討するためである。また、対象を単独で発話した単語発話では明瞭な fo が期待できる一方、朗読発話は連続発話中の要素を対象とするためイントネーション成分の影響が相対的に強く表れる。後者も分析することで、本手法の有効性を検証することができると考えた。なお、本研究では分析対象を 3 モーラに限定した。

2.1.1. 単語発話データ

NS による単語発話は、Albin（2017）のデータを用いた。3 名の NS（女性）が 160 語の 3 モーラ無意味語を単独で読み上げている（3 名 * 160 語 = 480 語）。無意味語は 87.5 %が CVCVCV の音節構造（長音、促音、撥音はなし）で、分節音の無声／有声阻害音や共鳴音

の比率および読み上げる際のアクセント型がほぼ均等になるように構成されている。NS は設定されたアクセント型通りに読み上げるよう指示され、異なった場合は再度発話した。

NNS による単語発話は、王・他（2018）のデータを用いた。52 名の初級中国人日本語学習者（男性 9・女性 43、学習歴約 6 ヶ月）が、2 モーラ有意味語+助詞「が」を単独で 9 つ読み上げている（52 名 * 9 語 = 468 語、本研究ではこれを単語発話に含める）。有意味語は母音の無声化や特殊拍を避けたものが選定された。学習者が読み上げる際にはアクセント核の位置に記号が付けられた状態で呈示されたが、それと異なる型で発話されても修正されなかった。各読み上げ音声のアクセント型は、2 名の評価者（NS と NNS）が全て聴取して判定し、一致しなかった場合はもう 1 名の評価者（NS）が聴取して最終的なアクセント型を決定した。全員の評価者が日本語音声学の知識を有した日本語教育の経験者である。

2.1.2. 朗読発話データ

『日本語学習者による日本語／母語発話の対照言語データベース モニター版（2005）』の DVD に収録されている朗読発話を用いた。日本語・中国語・韓国語・タイ語の母語話者各 10 名（うち男性は日 1、中 3、韓 2、泰 1）が朗読する課題のうち、「コマ」（17 文 633 モーラ）と「タバコ」（11 文 592 モーラ）を用いた。各 NNS の平均日本語学習月数（sd）は、中 67.1（30.9）、韓 43.5（32.9）、泰 34.8（9.2）である。テキストから、「1 つの自立語に 0 以上の付属語が後続する」という条件で文節境界を定め、3 モーラの文節を計 24 抽出した（「コマ」18、「タバコ」6）。ただし、言い直しやポーズの挿入による分割などの影響で、全員が同じ文節数ではない。単語発話と揃えるため、特殊拍が含まれる文節は除外した。最終的に、全体で 1,152 文節が得られた（日 250、中 273、韓 295、泰 334）。全文節を 1 名の NS（日本語音声学の知識がある日本語教育経験者）が聴取し、アクセント型を判定した。

全 2,100 データ（単語発話 968、朗読発話 1,152）のアクセント型分布を表 1 にまとめる。なお、朗読発話の NNS には 3 つの母語が含まれるが、全てまとめて分析を行った。

表 1 単語発話と朗読発話における NS と NNS のアクセント型分布

アクセント型	単語発話		朗読発話	
	NS	NNS	NS	NNS
0 型	168	177	89	319
1 型	156	143	111	335
2 型	156	148	50	248
計	480	468	250	902

2.2. 分析方法

音素単位で fo の代表値を求めるため、まず、各音素のアライメントと fo の抽出を行った。その後、fo 抽出の欠損や不安定さに対処するため、fo 形状のモデリング（再構築）を行った。それ基づいてアクセント型を判定するための特徴量を抽出し、機械学習を行った。

2.2.1. 音素のアライメントと fo の抽出

NS の単語発話データは既に音素境界のアライメントが行われていたので、本分析ではそれを利用した。それ以外のデータには、音声認識エンジン Julius（ver. 4.4.2）の音素セグメンテーションキット（ver. 4.3.1）を使用して、音素境界のアライメントを行った。言語モデルと音響モデルは、キットに含まれているデフォルトのものを使用した。朗読音声では、アライメント情報に基づいて、3 モーラ文節部分の切り出しも行った。

全 2,100 の音声ファイルに対し、Praat（ver. 6.0.36）の To Pitch コマンドを用いて fo を抽出した。抽出パラメータは、Time step を 0.001 sec、Pitch floor を 75 Hz、Pitch ceiling を 600 Hz に設定し、Kill octave jumps の機能を使用した。

2.2.2. fo 形状のモデリング

まず、各音素区間の fo と時間情報から、音素ごとに回帰直線を計算した。ただし、fo 抽出の安定性を考慮して、対象は母音/a, i, u, e, o/、鼻音/m, n/、わたり音/w, y/、撥音/N/とした。これらの音素であっても、区間全体で fo 抽出されたフレームが 20 %以下であった場合は、信頼性の観点から回帰直線を計算しなかった。これらの条件は Albin (2017) を参考にした。さらに、区間内の fo から四分位数を求め、「第 1 四分位数 * 1.5 以上」「第 3 四分位数 * 1.5 以下」に該当した fo は外れ値として、回帰直線の計算に含めなかった。

次に、算出した回帰直線の時間方向で 25 % と 75 %の地点に制御点を設定し（先頭音素は 0 %、末尾音素は 100 %の地点にも）、それらを線形補間することで全体の fo 形状を再構築した。なお、末尾の母音音素が無声化などによって回帰直線が計算できなかった場合、直近の回帰直線の 100 %地点の fo を、末尾母音音素の 100 %地点まで延伸させた（図 1 a）。また、朗読発話で頻出した「末尾上げ」に対処するため、最終音素で回帰直線の傾きが正であった場合は、その 25 %地点における fo を 100 %地点まで延伸させた（図 1 b）。このように fo をモデリングすることで、fo 抽出の欠損や不安定さに対処した。

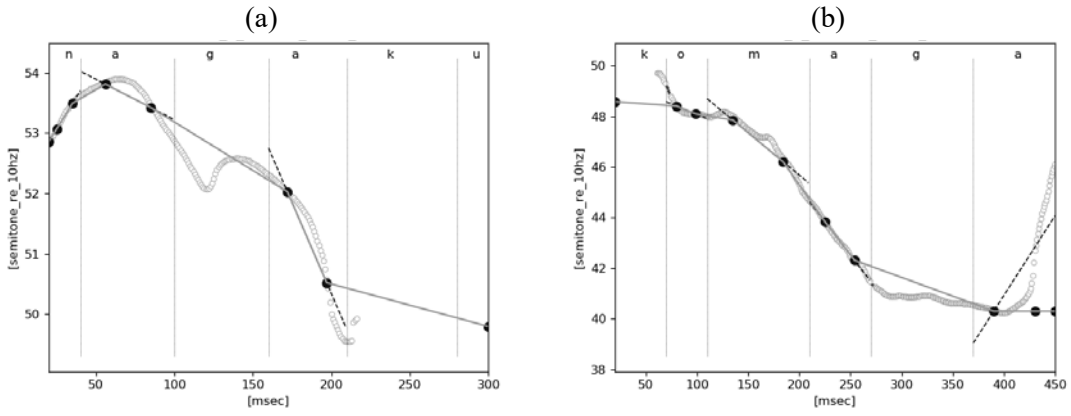

図 1 fo 形状のモデリング概要（いずれも朗読発話より、(a)「長く」と(b)「コマが」）
白抜き円は fo の値を、点線は回帰直線を、黒丸は制御点を、実線は線形補間の結果を示す。

2.2.3. 特徴量の抽出

5 種類・7 項目の特徴量を抽出した。線形補間した fo 軌跡からは、各母音音素の 50 %地点の fo（=中央値）から求めた「V_1V_2」「V_2V_3」の差、および、100 %地点の fo（=終端値）から求めた「V_1V_2」「V_2V_3」差を計算した（V = 母音、semitone)。その他、対象単位冒頭の fo 値（Hz)、対象単位全体の fo レンジ（semitone)、発話速度（msec）も求めた。

母音音素の fo 中央値の差分は、波多野・他（2014）のアクセント型判定において高い推定精度を示したことから、本研究でも特徴量として採用した。fo 終端値の差分は、おそ下がりに対処するための特徴量として採用した。また、冒頭 fo 値は性差を、fo レンジは学習者母語の特性を、発話速度は学習者の習熟度を考慮するため、特徴量に含めた。

2.2.4. 機械学習の実行

機械学習には、Python（ver. 3.6.5）の Scikit-learn（ver.0.19.1）パッケージを使用した。機械学習に用いる推定器は、正解ラベル（今回はアクセント型）に基づいた「分類」であるという点や、データ数を考慮し、SVC（Support Vector Classification）とした。

データ全体を、分類モデルを学習するための「トレーニングセット」と、そのモデルの精度を検証するための「テストセット」に分割し（割合は 8 対 2)、層化 k 分割交差検証（stratified k-fold cross validation）を行った。k 分割交差検証では、トレーニングセット全体を非復元抽出で k 個のサブセットに分割した上で、k - 1 個のサブセットをモデルのトレーニングに使用し､残りのサブセットでそのモデルの検証を行うことを k 回繰り返すことで、最終的な分類モデルを学習する（Raschka 2015)。分析データはアクセント型の分布が不均等なため（表 1)、アクセント型で層化抽出を行うことで、各サブセットにおける型の比率が維持されるようにした。今回は k = 10 とし、10 回の交差検証で推定平均性能を計算した後、トレーニングセットからは独立したテストセットで最終的な分類性能の評価を行った。

なお、SVC を行う際、各特徴量の値は標準化（平均値 0、標準偏差 1）し、ハイパーパラメータはグリッドサーチによって最適な組み合わせを選択した。

3. 結果と考察

3.1.1. 単語発話データの分類結果

表 2 に、NS および NNS による単語発話データの分析結果を示す。いずれもテストセットの分類結果が 90 %を超えており、NS では 100 %となった。これらはアクセントを意識した発話のため、各母音間の fo 差が比較的明瞭であった可能性がある。

表 2 単語発話データにおけるアクセント型の分類結果

データ	n	トレーニングセット		テストセット	
		N	平均性能（sd）	N	分類結果
NS	480	384	99.7 %（0.008）	96	100.0 %
NNS	468	374	94.1 %（0.023）	94	92.6 %

3.1.2. 朗読発話データの分類結果

表 3 に、NS および NNS による朗読発話データの分析結果を示す。NNS ではテストセットの分類結果が 80 %を超えたものの、NS では 76 %に留まった。

単語発話と比較して低い精度に留まったのは、1 名の評価者によるアクセント型判定の信頼性が影響している可能性がある。今後は評価者を増やす等をして正解ラベルの信頼性を高める必要がある。また、NS の分類結果が NNS よりも悪かったのは、過小なサンプル数で十分な分類モデルの学習ができなかったことによると考えられる。

表 3 単語発話データにおけるアクセント型の分類結果

データ	n	トレーニングセット		テストセット	
		N	平均性能（sd）	n	分類結果
NS	250	200	83.0 %（0.085）	50	76.0 %
NNS	902	721	82.1 %（0.029）	181	84.0 %

3.1.3. アクセント型ごとの分類性能

単語・朗読発話のテストセットにおけるアクセント型ごとの分類結果について、正解ラベルとの混同行列を図 2 に示す。図 2 から、朗読発話で 2 型の分類精度が良くないことがわかる（NS 50.0 %、NNS 71.4 %）。

朗読発話では、イントネーション成分の影響で 2 型の特徴である V_2V_3 の fo 差が明瞭に現れなかった可能性がある。また、朗読発話では 2 型の出現数がそもそも少ない（表 1）。トレーニングセットに分割した際には更に少なくなるため、2 型で特に精度が悪かった原因として以上が考えられる。

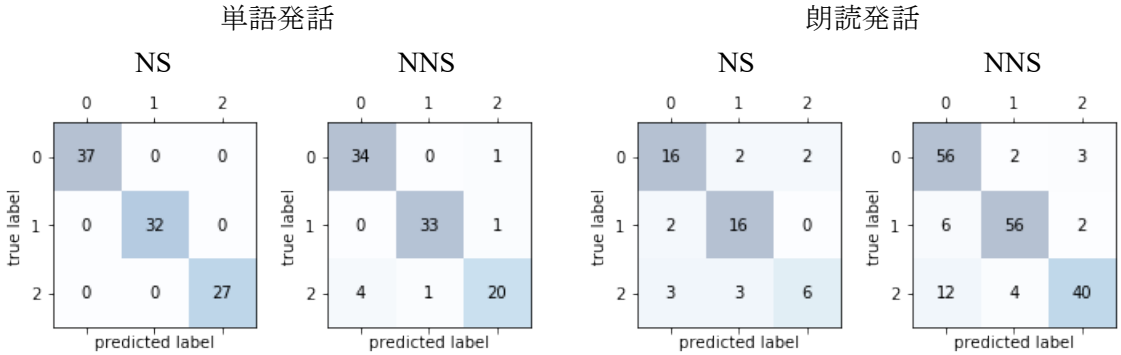

図 2 単語発話と朗読発話のテストセットにおけるアクセント型の混同行列

true label は正解のアクセント型を、predicted label は機械学習によるアクセント型の分類を示す

4. まとめ

機械学習を用いた本研究の手法は、単語発話では極めて高い精度でアクセント型分類が可能であり、朗読発話ではデータ拡充により更なる精度向上が期待できることが示された。今後は更にデータを増やすとともに、本研究で明らかになった問題を改善していきたい。

謝辞

本研究の単語発話データで使用した王・他（2018）は JSPS 科研費 17H02352 の助成を受けたものである。また、朗読発話データで使用した DVD は JSPS 科研費 14380121 の研究成果報告書の付録である。

参考文献

Albin, Aaron（2017）“F0 Contour Parameterization Using Optimal Regression Chains”『第 31 回日本音声学会全国大会予稿集』79-84.

Hatano, Hiroaki., Ishi, Carlos Toshinori., Song, Cheng Chao., Matsuda, Makiko（2018）“Automatic evaluation of accentuation of Japanese read speech,” In M. Ueyama and I. Srdanović (eds.) *Digital resources for learning Japanese*, 55-71, Bologna: Bononia University Press.

Raschka, Sebastian（2015）Python Machine Learning. Packt Publishing.(株式会社クイーブ（訳）（2016）『Python 機械学習プログラミング』株式会社インプレス.)

石井カルロス寿憲・西山隆二・峯松信明・広瀬啓吉（2001）「日本語のアクセント・イントネーションを対象とした発音教育システム構築に関する検討」『信学技報』100(594), 33-40.

磯村一弘（2001）「海外における日本語アクセント教育の現状」『日本語教育学会秋季大会予稿集』211-212.

王睿来・林良子・磯村一弘・新井潤（2018）「中国語母語話者による日本語アクセントの習得－知覚と生成の関係に着目して－」『ことばの科学研究』19, 81-96.

小河原義朗（2001）「日本語非母語話者の話す日本語の発音に対する日本人の評価意識：社会人の場合」『日本語教育方法研究会誌』8(2), 10-11.

佐藤友則（1995）「単音と韻律が日本語音声の評価に与える影響力の比較」『世界の日本語教育』5, 139-154.

佐藤友則（1998）「韓国および台湾の日本語学習者のニーズ調査」『言語科学論集』2, 49-60.

杉藤美代子（1969）「動態測定による日本語アクセントの解明」『言語研究』55, 14-39.

轟木靖子・山下直子（2009）「日本語学習者に対する音声教育についての考え方：教師への質問紙調査より」『香川大学教育実践総合研究』18, 45-51.

波多野博顕・石井カルロス寿憲・松田真希子（2014）「日本語朗読音声を対象にしたアクセント型自動判定方法の検討」『日本音響学会秋季研究発表会講演論文集』363-364.

広瀬啓吉（2005）「音声の韻律と CALL」『音声研究』9(2), 38-46.

松崎寛（2016）「日本語音声教育における韻律指導」『日本音響学会誌』72(4), 213-220.

『日本語学習者による日本語発話と母語発話との対照データベース－開発・応用のための研究－』（平成 14-16 年度科学研究費補助金　基板研究（B）（2）研究成果報告書、研究代表者：宇佐美洋）