Vol.2019-MUS-123 No.27
Vol.2019-SLP-127 No.27
2019/6/22

調音運動及び音響特徴量に着目した日本語単語における発声しやすさの自動評価の検討

佐藤 禎哉[1]　齋藤 大輔[2]　峯松 信明[2]

概要：本稿では調音運動から抽出した特徴量および音響特徴量を用いた，日本語単語の発声しやすさに関する定量的指標を検討する．日本語単語に対する発声のしやすさは，これまで放送にあるべき言葉遣いの手がかりとして議論がなされる一方，その指標は定性的なものにとどまっていた．これに対して定量的指標の導入は工学的，教育的観点から有用と考えられる．本研究では発声のしやすさに着眼して，発声速度も考慮した上で単語データセットを作成した．一方で調音運動および音響特徴量に着目した定量的指標を作成し，前述のデータセットを用いた識別実験を通じてその有用性を検討した．

キーワード：発声しやすさ，調音運動，音響特徴量

1. 序論

日本語の発声の仕方，特に発声しにくい語句の特徴は，放送に携わる専門家たちの間で放送におけるあるべき言葉遣いの 1 つの手がかりとして議論がなされてきた．また，単語認知過程研究の主要なテーマとして視覚提示された単語の音韻特性の影響が検討されており，その中でもその単語がどの程度発声しやすいと感じるかが重要視されてきた．湊・木村による調査では，アナウンサー 282 名が常日頃発声しにくいと思っている語句を記述式で収集し，回答数が多かった語句を元に定性的な指標を提示している [1]．鈴木らの研究では，早口言葉を資料として，言い誤りや言い淀みをどの音素で起こしやすいかについて分析を行っている [2]．また，川上は視覚提示された単語の音韻特性として，文字列が「どの程度発音が容易か」の主観的評定値である発音容易性のデータベース作成を行っている [3]．

これらの研究で提示されている指標は，対象の語句に対してどのような音素が続いた場合に発声しにくいと考えられるか，といった定性的な分析によるものにとどまっている．そのため，任意の発声しやすい単語と発声にくい単語にこれらの研究で提示された指標が適用可能である保証はない．したがって，発声しにくい単語を発声しやすい単語に入れ替えることで発声しやすい文章を提示する，といった応用に対して今までに提示されている指標を用いることが可能かどうかは検討する必要がある．宋らの研究では，発声しにくさの指標として調音運動という生体の観測と直接関係した指標を用いることで，定性的な評価ではなく定量的な評価を行うことを検討している [4]．しかし，検討している語句が 10 個のみであり，発声しにくい語句の包括的な指標となっているかどうかは検討の余地がある．

本研究では，発声しやすい単語と発声しにくい単語を識別できる定量的な指標の提案を目的とする．指標の提案に利用する特徴量は宋らの研究でも検討がなされていた調音運動や [4]，音響特徴量から作成する．そして，提案した特徴量を用いて発声しやすい単語と発声しにくい単語の分類や発声しやすさの推定を行い，特徴量の有効性を検討する．

またこれまでに発声しやすさに着目したデータセットの作成も行われているが，実際に発声した場合の発声速度は考慮されていない．本研究では評価者に実際に発声させるプロセスを追加することで，発声の速度を考慮した，声しやすさを定量的に評価可能なデータセットの作成を検討した．

2. 先行研究

2.1 発声しにくさの定性的な分析

湊らは，NHK 放送文化研究所番組研究部において行われたアンケートに基づいて，発声しにくい語句の分析を行った [1]．この調査では，全国のアナウンサー 282 名を対象にして，4 名以上が発声しにくいと回答した語句の音声的構造を調べ，発声しにくさの要因と考えられる特徴を取り出している．語句の音声的構造を調査する際には音素表記法を用いている．この研究で提示された発声しにくさの要

[1] 東京大学大学院情報理工学系研究科電子情報学専攻
[2] 東京大学大学院工学系研究科電気系工学専攻

表 1: 発声しにくさの要因と考えられる語句の特徴とその例 [1]

番号	特徴	例
1	音節主音として/i/あるいは/u/を持つ音節の連続	ひしひしと
2	音節主音として/a/を持つ音節の連続	あたたかい
3	それぞれ別々の音節に属している/-ei/の/e/と/i/，/-ii/の/i/と/i/などの母音音素の並列	委員会
4	子音音素の次に半母音音素/j/が位置している音節の連続	
5	子音音素の次に半母音音素/j/あるいは/w/を間に立てての母音音素/a/の並列	見誤る
6	モーラ音素/N/の連続	問題懇談会
7	いわゆる硬音的な無声子音音素/s/，/t/，/c/，/k/ などを音節副音として持つ音節の連続	人たち
8	摩擦音・破擦音系列の子音音素を音節副音として持つ音節の連続	7 時
9	両唇音の子音音素を音素副音として持つ音節の連続	見守られる
10	歯茎音の系列の子音音素を音節副音として持つ音節の連続	バナナなど

因と考えられる音声的構造の特徴とその例を表 1 に示す．

2.2 調音運動に基づく発声しにくさの定量的評価

宋らは調音運動に基づいたテキストの発声しにくさの指標の作成手法を提案した [4]．まず，テキストを調音運動に変換するために，隠れマルコフモデル（Hidden Malkov Model; HMM）に基づく音声合成法を応用して，テキストから調音位置 12 次元の系列を生成するシステム（Text to Articulation）の構築を行った．続いて，Text to Articulation を用いて，湊らの調査 [1] で発声しにくい語句として挙げられている語句の調音運動を出力し，観察した．そして，12 次元の調音運動のうち，水平方向の舌の往復，垂直方向の唇の停滞の 2 つに着目し，この 2 つの特徴を基にして発声しにくさの指標を作成した．

2.3 発声しにくさに着目したデータベースの構築

川上はカタカナ 3 文字表記語に対して発音容易性を調査し，データベース作成を行なった [3]．発音容易性とは，文字列が「どの程度発音が容易か」の主観的評定値である．被験者は日本福祉大学に所属する大学生 123 名であり，各被験者は配られた質問紙に記載されたそれぞれのカタカナ 3 文字表記語について「非常に発音しにくい（1）」から「非常に発音しやすい（5）」までの 5 段階のいずれかを評定した．また，発音容易性の評定値としては被験者が回答した 5 段階の評定値の平均値を用いた．

調査の結果，カタカナ 3 文字表記語のうち，もっとも発音容易性が高いと評定されたのは「タンク」及び「ダンス」であり，評定値は 4.66 であった．また，最も発音容易性が低いと評定されたのは「パナマ」であり，主観的評定値は 1.97 であった．この調査で得られた単語の発音容易性と，単語を日常でどの程度目にするかの主観的評定値である主観的出現頻度には正の相関が見られた．

3. 発声しやすさに着目した単語データセットの作成

3.1 概要

本研究では，発声しやすさに着目した単語データセットの作成を目的として被験者に対して単語を提示し，実際に発声させた上でどの程度発声しにくいか 5 段階で評価させる主観実験を行った．

3 人の被験者に対して実験を行った．3 人のうち 2 人が東京都出身の大学生であり，1 人が神奈川県出身の大学生である．また，早口で何度も発生させることで発声しやすい単語と発声しにくい単語に差が生まれると仮定し，早口で複数回発声させた後に評価させる形で実験を行った．データセット作成の対象とする単語は，単語親密度データベース [5] の中で，文字音声単語親密度が 5.0 以上の 7 モーラ単語 633 個とした．

3.2 実験手順

実験は以下の手順で行った．

(1) Web ページ上に白抜きの単語が表示される．
(2) 「録音」ボタンを押すと，1 秒後からアニメーションが動き出し図 1 のように徐々に黒くなっていく．それに合わせて被験者に表示されている単語を読み上げさせる．読み上げは 3 秒に 5 回のペースで行う．
(3) 読み上げが終わったら，発声のしにくさについて，「とても発声しやすい」から「とても発声しにくい」の 5 段階で評価させる．その後，「次の単語」ボタンをクリックしてもらい，次の単語のページに遷移する．また，このときに録音した読み上げ音声と単語の発声しにくさの評価を取得する．
(4) （1）～（3）を終了の表示が出るまで繰り返させる．

3.3 実験結果

主観実験で得られたデータの確認として，特徴的な単語の確認および被験者どうしのラベリングの比較を行った．

まず，被験者全員がとても発音しやすい（1）と回答した単語及び被験者全員がとても発音しにくい（5）と回答した単語を表 2 に示す．

次に，被験者どうしのラベリングの比較を行った．各単語についてそれぞれの被験者が，とても発音しやすい（1）もしくは発音しやすい（2）と回答した単語，どちらとも言

Vol.2019-MUS-123 No.27
Vol.2019-SLP-127 No.27
2019/6/22

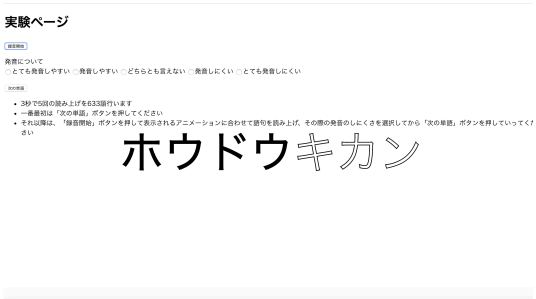

図 1: 主観実験のページ．開始ボタンを押すとアニメーションが動き始める．そのアニメーションに合わせて被験者に発声させた後，単語の発声しやすさについて 5 段階評価させる．

表 2: 主観実験において，被験者全員がとても発音しやすい（1）と回答した単語および，被験者全員がとても発声しにくい（5）と回答した単語の例

とても発声しやすい	何と言っても，アンダーライン，オーバーネット，リーダーシップ
とても発声しにくい	50 歩 100 歩，持ちつ持たれつ，今の今まで，物珍しげ，直接選挙，追跡調査，推理小説，兎にも角にも，プロセスチーズ，南無阿弥陀仏

えない（3），発音しにくい（4）もしくはとても発音しにくい（5）と回答した単語の数は表 3 の通りであり，被験者間の回答の傾向の一致率は表 4 の通りである．表 4 の結果より，被験者間で高い一致率を確認することができる．

3.4 考察

表 4 より被験者間の回答の傾向の一致率が高いことから，主観実験で得られたデータは多少の個人差はあるものの有効なラベル付きデータであるとみなし，以降はこのデータセットを用いて検討を行う．

4. 調音運動及び音響特徴量に着目した特徴量の提案

単語どうしを比較する際に発話の長さがそれぞれ異なる場合，フレームに分割した音声から特徴量を抽出し，それを結合すると特徴量の次元が合わずそのまま比較することができないという問題が生じる．このような問題への対処として，例えば感情認識においては時系列データに対して統計量を計算して特徴量として扱っている [6]．そこで，本研究では感情認識に用いられる openSMILE[*1] における特徴量の抽出方法に着想を得て新たに抽出する特徴量を検討した．今回検討を行った 4 つの特徴量を表 5 に示す．

調音データを時系列データとして扱い，平均およびケプストラムを抽出した特徴量を articualtion-based1 とする．また，調音データから平均や分散といった統計量を抽出することで得る特徴量を articulation-based2 とする．さらに，ケプストラムおよび MFCC の統計量を抽出して得る特徴量をそれぞれ cepstrum-based，mfcc-based とする．調音データは先行研究で述べた Text to Articulation を用いて合成し，音声データは Open JTalk[*2] を用いて合成している．実際の調音運動や実音声を用いた場合，発話者が発声しにくいと感じる単語の傾向が調音運動や音声に現れる可能性を考慮し，本研究では調音運動や音声を合成している．音声合成の際に使用する音響モデルは，HTS-2.3[*3] のデモスクリプトを用いて日本語の話者依存学習モデルを作成し，それを用いた。また，音声データからケプストラムやメル周波数ケプストラム係数（mel frequency cepstral coefficient; MFCC），Δ 特徴量，ΔΔ 特徴量を抽出する際

表 3: 各被験者どうしの主観評価値の比較．それぞれがとても発音しやすい（1），発音しやすい（2）と回答した単語，どちらとも言えない（3）と回答した単語，発音しにくい（4），とても発音しにくい（5）と回答した単語の個数を比較している．

(a) 被験者 1 と被験者 2 の比較

被験者 2 \ 被験者 1	1，2	3	4，5
1，2	253	52	15
3	23	10	3
4，5	107	70	100

(b) 被験者 2 と被験者 3 の比較

被験者 3 \ 被験者 2	1，2	3	4，5
1，2	196	20	52
3	83	9	64
4，5	41	7	161

(c) 被験者 3 と被験者 1 の比較

被験者 1 \ 被験者 3	1，2	3	4,5
1，2	223	103	57
3	35	37	60
4，5	10	16	92

表 4: どちらかの被験者がどちらとも言えない（3）と回答した単語を除いた単語における回答の一致率

被験者 1vs 被験者 2	74.3%
被験者 2vs 被験者 3	79.3%
被験者 3vs 被験者 1	82.5%

*1 https://www.audeering.com/technology/opensmile/

*2 http://open-jtalk.sp.nitech.ac.jp

*3 http://hts.sp.nitech.ac.jp

表 5: 本研究で検討を行う発声しにくさの指標として機能することを期待する特徴量

タグ	抽出する特徴量	次元数
articulation-based1	12 次元の調音データの各次元の時系列データに対する平均，ケプストラムの先頭 12 次元	156
articulation-based2	12 次元の調音データの各次元の時系列データ，時系列データの一次微分，二次微分それぞれに対する平均，分散，最大値，最小値，最小二乗法によって一次関数，二次関数に近似した際の係数および近似誤差	396
cepstrum-based	音声データのケプストラム，Δ 特徴量，ΔΔ 特徴量各 13 次元に対する平均，分散，最大値，最小値，最小二乗法によって一次関数，二次関数に近似した際の係数および近似誤差	429
mfcc-based	音声データの MFCC，Δ 特徴量，ΔΔ 特徴量各 13 次元に対する平均，分散，最大値，最小値，最小二乗法によって一次関数，二次関数に近似した際の係数および近似誤差	429

には SPTK[*4] を用いた．フレーム抽出におけるフレーム長は 32.0ms，シフト長は 5.0ms とした．

5. 作成データセットを用いた発声しやすさの識別

5.1 概要

主観実験で得られたデータセットに対して前章で検討した特徴量を用いてサポートベクターマシンによる分類を行った．

5.2 実験条件

主観実験で得られたデータは単語 633 語とそれに対する各被験者がつけた発声しやすさの 5 段階評価である．本項では，各単語に対して被験者がつけた 5 段階評価の平均を単語の発声しやすさのスコアとし，各単語から抽出する特徴量は表 5 の通りとした．特徴量を抽出する調音データは Text to Articualtion を用いて得，音声データは Open J Talk を用いて得た．サポートベクターマシンのカーネルは rbf を選択した．また，パラメータに関して，C は 10^{-3} から 10^{3} の範囲で対数軸で 10 点，γ も同様に 10^{-3} から 10^{3} の範囲で対数軸で 10 点選びグリッドサーチを行った．サポートベクターマシンは scikit-learn[*5] の実装を使用した．データの前処理として，各単語に対応する多次元データの最大値を 1，最小値を 0 として線形変換を行った．そして，データの 8 割を学習データ，2 割をテストデータとして分割し，学習データに対して 5 分割交差検証を行ってパラメータを決定した．そして，決定したパラメータを用いたサポートベクターマシンによるテストデータに対する正解率を確認した．

[*4] http://sp-tk.sourceforge.net
[*5] https://scikit-learn.org/stable/

表 6: 主観実験から作成したデータセットを用いた SVM による分類．

ラベリング	特徴量	正解率	適合率	再現率	F 値
w/ mid	articulation-based1	62.9%	38.7%	33.3%	35.8%
	articulation-based2	75.0%	59.5%	61.1%	60.3%
	cepstrum-based	72.4%	56.3%	50.0%	52.9%
	mfcc-based	70.7%	52.8%	52.8%	52.8%
w/o mid	articulation-based1	75.4%	47.1%	53.3%	50.0%
	articulation-based2	75.4%	47.3%	60.0%	52.9%
	cepstrum-based	**90.8%**	76.5%	86.7%	81.3%
	mfcc-based	81.5%	57.9%	73.3%	64.7%

発声しやすさの個人差を考慮して，各単語に付与されたスコアをもとに 2 種類の方式で 2 値分類のためのラベルを作成した．1 つは，スコア 3.0 を境界として 3.0 より小さい単語 375 語を発声しやすい，3.0 より大きい単語 204 単語を発声しにくいとする方式である．この中間スコアを含むラベリング方式を with intermediate(w/ mid) とする．もう一方はスコアが 2.0 以下の単語 219 語を発声しやすい，4.0 以上の単語 106 語を発声しやすいとする方式である．この中間スコアを含まないラベリング方式を without intermediate(w/o mid) とする．後者の方式は主観評価者のスコアがより安定したもののみを識別に用いているといえる．

5.3 実験結果

表 6 に主観実験から作成したデータセットに対するサポートベクターマシンの正解率，適合率，再現率，および F 値を示す．表 6 より，w/ mid では正解率は 70% 前後となり，特に articulation-based2，cepstrum-based，mfcc-based による分類の正解率には大きな差がなかった．一方，w/o mid では検討を行っている全ての特徴量において正解率が上がる結果となり，音響特徴量に基づいた cepstrum-based，mfcc-based，特に mfcc-based が高い正解率を示した．

5.4 考察

w/ mid と比較して w/o mid の正解率が高いことから，被験者間で意見が分かれたような単語についてはサポート

ベクターマシンによる分類では適切に分類できないと考えられる．また，articulation-based1，srticulation-based2 と比較して，cepstrum-based，mfcc-based は正解率が全体的に高く，特に w/o mid においては 80% 以上の正解率が確認できることから，音声データから特徴量を抽出する方法は一定の有効性が確認できた．

一方，表 4 のにおける被験者間の回答の傾向の一致率と表 6 における w/ mid の正解率を比較すると被験者間の回答の一致率が高い．表 4 における一致率はいずれかの被験者がどちらとも言えない（3）と回答した単語を除く操作を加えているため，表 6 との単純な比較はできないが，本研究で提案した特徴量は人間どうしでの評価と比較すると十分有効とはいえないと考えられる．

6. 作成データセットを用いた追加検討

6.1 概要

前項で特に発声しやすい単語と特に発声にくい単語の識別に関して本研究で提案した特徴量は一定の性能を示したが，スコア自体への回帰や，多クラス分類が可能なのかどうかは検討していない．そこで，本項では，サポートベクターマシンによる多クラス分類，サポートベクター回帰を主観実験によって得られたデータセットに適用可能かどうか検討する．

6.2 実験条件

多クラス分類および回帰について，2 値分類の時と同様にデータの 8 割を学習データ，データの 2 割をテストデータとして分割する．また，カーネルは rbf を選択し，パラメータに関して C は 10^{-3} から 10^{3} の範囲で対数軸で 10 点，γ も同様に 10^{-3} から 10^{3} の範囲で対数軸で 10 点選びグリッドサーチを行った．そして，学習データに対して 5 分割交差検証を行ってパラメータを決定したのち，そのパラメータを用いてテストデータに対する正解率を確認した．

多クラス分類を行う際のラベリングは，各単語のスコアが 2.0 以下の単語 219 語を 0，すなわち発声しやすい単語，スコアが 2.0 より大きく 3.0 より小さい，または 3.0 より大きく 4.0 より小さい単語 254 語を 1，すなわちどちらとも言えない単語，4.0 以上の単語 106 語を 2，すなわち発声しにくい単語の 3 クラスとした．また，回帰の際は各単語のスコアを正解データとし，ϵ は 0.1 で固定した．

6.3 実験結果

表 7 は，サポートベクターマシンによる多クラス分類の結果である．どの特徴量を用いた場合も正解率は 50% 前後であり，十分な正解率は確認できなかった．

表 8 はサポートベクターマシンを用いた回帰の結果である．こちらも，どの特徴量を用いた場合でも，決定係数は低かった．

表 7: 主観実験から作成したデータセットを用いたサポートベクターマシンによる多クラス分類．特徴量は表 5 の通りに抽出した．

特徴量	テストデータへの正解率
articulation-based1	46.6%
articulation-based2	52.6%
cepstrum-based	49.1%
mfcc-based	54.3%

表 8: 主観実験から作成したデータセットを用いたサポートベクター回帰．特徴量は表 5 の通りに抽出した．

特徴量	テストデータへの決定係数
articulation-based1	0.157
articulation-based2	0.234
cepstrum-based	0.252
mfcc-based	0.324

6.4 考察

実験結果の表 7，8 のテストデータへの正解率，テストデータへの決定係数は十分大きい値ではない．したがって，本研究で検討した特徴量では，大まかな単語の発声しやすさの分類は可能であっても，多クラス分類や回帰による点数の推定を行う特徴量としては不十分だと考えられる．これは表 6 から，複数人が発声しやすいと認識できるような単語と複数人が発声しにくいと認識できるような単語の分類には提案した特徴量が有効である一方，被験者によって判断が分かれる単語が含まれると正解率が下がることが確認されたことも一致する．

7. 結論

本研究では，発声しやすい単語と発声しにくい単語を識別する特徴量として調音データから抽出した特徴量と音響特徴量を検討した．その結果，調音データから抽出した特徴量に比べて，音響特徴量は識別率が高く，また，識別を特に発声しやすい単語と特に発声しにくい単語に限定した場合，音響特徴量による識別は 80% 以上の正解率が確認できたことから，特徴量として有効であると考えられる．一方，主観実験における被験者間の回答の一致率と比較して本研究で提案した特徴量を用いた識別の正解率が低いことから，本研究で提案した特徴量は人間どうしの評価以上の精度を出すのは難しいと考えられる．

また，本研究では発声しやすい単語と発声しにくい単語のデータセットの作成を行った．表 4 から被験者間の回答の傾向が一致していることが確認できるように，被験者が行った 5 段階評価は有効であり，データセットとしての利

用が期待できる．主観実験を行う人数や評価を行う単語数を増やすことでより信頼性の高いデータセットを作成することが可能であると考えられる．

参考文献

[1] 湊 吉正，木村圭子:「発音しにくいことば」の調査とその分析，NHK 放送文化研究所年報第 9 集，191–206 ，1964.
[2] 鈴木誠史，臼杵秀範，島村徹也: 日本語早口言葉の構造と性質，放送教育開発センター研究紀要第 12 号．131–149．1995.
[3] 川上 正浩: カタカナ 3 文字表記語 449 語の発音容易性調査，大阪樟蔭女子大学 人間科学 研究紀要 第 1 号，43–52，2002.
[4] 宋健智，齋藤大輔，峯松信明: 調音運動に基づく発声難易度の指標化と歌詞の歌いやすさ評価への応用の検討，研究報告音声言語情報処理，2018-SLP-122，40，1–6，2018.
[5] 天野成昭，近藤公久: NTT データベースシリーズ「日本語の語彙特性」第 1 巻 単語親密度，1999.
[6] 鈴木基之: 音声に含まれる感情の認識，日本音響学会誌，71 巻，9 号，484–489，2015.