第32回日本音声学会全国大会@沖縄国際大学
2018年9月15日（土）13:30-14:00

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（ATR）

発表の構成

1. 目的・背景・先行研究
2. データの概要と分析方法
3. 結果と考察
4. まとめ

1. 目的・背景・先行研究

目的

音響特徴に基づいて、アクセント型を
自動で判定する方法を検討し、

発音（韻律）の学習・教育・評価

に役立てたい。

既存の音声コーパスにも応用することで、
音声データマイニング手法としての貢献も期待

背景

□ 日本語教育では音声教育の機会が少ない

- ノンネイティブ（NNS）教師：自信がない（磯村 2001）
- ネイティブ（NS）教師：重要度の認識が低い（轟木・山下2009）

□ しかし、音声教育のニーズや必要性はある

- 自然な発音で話したいと思う学習者は多い（佐藤 1998）
- ビジネスや国内長期生活者での問題（小河原 2001）

■ 求められていること

- 自学自習が可能な音声自動評価法の開発（松崎 2016）
- 評価には単音よりも韻律の影響が大きい（佐藤 1995）

先行研究　（音響特徴からアクセント型判定を試みたもの）

	石井・他（2001）	広瀬（2005）	波多野・他（2014）	Hatano, et al（2018）
データ	日本語母語話者1名の朗読発話	日本語母語話者1名の朗読発話	日本語母語話者10名の朗読発話	日中母語話者各10名の朗読発話
方法	モーラ（CV）内のF0から回帰直線を引き、隣接モーラ間の終端値の差分で判定	隣接モーラのピッチレベルの差を多次元正規分布で表現し、未知音声に対して尤度最大の分布を求めて判定	VC単位内のF0から平均値、中央値、終端値を求め、隣接モーラ間の差分で判定	V内のF0から平均値と終端値を求め、隣接モーラ間の差分で判定
精度	78.4 %	75.5 %	平均値 67.2 % 中央値 71.2 % 終端値 66.5 % ※ 10名の平均	（日） 平均値 65.2 % 終端値 72.4 % （中） 平均値 74.7 % 終端値 77.8 %

課題

- 最適な判定閾値が分からない
  - 判定精度が80%を超えない
  - そもそも1つの情報からで良いのか？
- 音響情報の扱い
  - fo抽出の欠損・不安定さ…
  - アクセント知覚と対応しない物理的特徴（遅下がり）

■ fo外形を適切にモデリングする必要

■ 判定の特徴量を増やして機械学習してみる

2. データの概要と分析方法

データの概要

単語発話	NS	Albin（2017）
	NNS	王・他（2018）
朗読発話	NS	『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）』
	NNS	

母語話者／学習者を問わず有効な手法か

イントネーション成分の影響にどれだけ頑健か

- 本発表では3モーラを対象とした分析結果を報告

単語発話データ（NS）

■ Albin（2017）



- 総数：480
  - 3モーラの無意味語160語
    - 例） ぬしみ、ろさち、むぞと、…
  - 女性話者3名が単独で読み上げ
- アクセント型のバランスを統制

- 設定されたアクセント型通りの読み上げを指示
  - アクセント型の情報はこれを利用

単語発話データ（NNS）

■ 王・他（2018）



- 総数：468
  - 2モーラの有意味語 ＋ 助詞「が」
    - 例） 嫁が、鮎が、沼が、…
  - 初級中国人学習者52名（学習歴約6ヶ月）
  - 単独で読上げ ※便宜上「単語発話」とする

- 読み上げ資料にアクセント記号を付与
  - 違うア型で読上げても、そのまま収集

- 3名の評価者が聴取して、アクセント型を決定

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）』

日本語／中国語／韓国語／タイ語母語話者
各10名が、同一内容のテキストを朗読

母語	平均学習期間（月）	平均滞在歴（月）
中国語	67.1（30.9）	53.6（34.2）
韓国語	43.5（32.9）	50.5（49.8）
タイ語	34.8（ 9.2）	無回答

括弧内は標準偏差

朗読発話データ

- 朗読課題から3モーラの文節を抽出
  - ① コマ　（17文 633モーラ）⇒18文節
  - ② タバコ（11文 592モーラ）⇒ 6文節
    例）ぶつけ、しかし、そとで…





- 総数：1,152
  NSデータ　：250（日）
  NNSデータ：902（273（中） 295（韓） 334（泰））
  ※ 言い直しやポーズ挿入による分割を含むため、
  全員が同じ文節数ではない

- 1名の評価者が聴取して、アクセント型を決定

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	NSS	NS	NSS
0型	168	177	89	319
1型	156	143	111	335
2型	156	148	50	248
計	480	468	250	902

- これらが「正解ラベル」となる
- 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。

分析方法

1. 音素のアライメントとfoの抽出
2. fo形状のモデリング
3. 特徴量の抽出
4. 機械学習の実行

1. 音素のアライメントとfoの抽出

■ 音素単位でfoの代表値を決めるための準備

- 音素のアライメント

  単語発話のNSデータはアライメント済なので利用
  それ以外は Juliusの音素セグメンテーションキット使用
  （言語・音響モデルはデフォルトのもの）

- foの抽出

  praatのTo Pitchコマンド使用

  Time step: 0.001sec, Pitch floor: 75 Hz,
  Pitch ceiling: 600 Hz, Kill octave jumps使用

2. fo形状のモデリング

■ 抽出の欠損や不安定さに対処

- fo値と時間情報から回帰直線を計算（Albin 2017）

  母音/a, i, u, e, o/　鼻音/m, n/　わたり音/w, y/　撥音/N/

  ※ fo抽出フレームが音素区間全体の20%以下だと除外

  「第1・3四分位数 * 1.5 以上・以下」のfo値は計算外

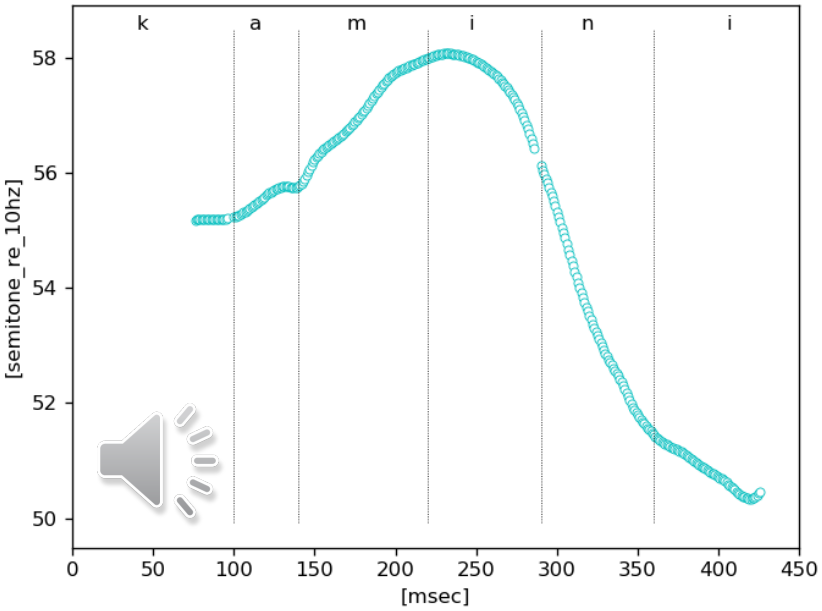
- 制御点の設定と線形補間

  回帰直線の時間方向で25・75%の位置に制御点を設定

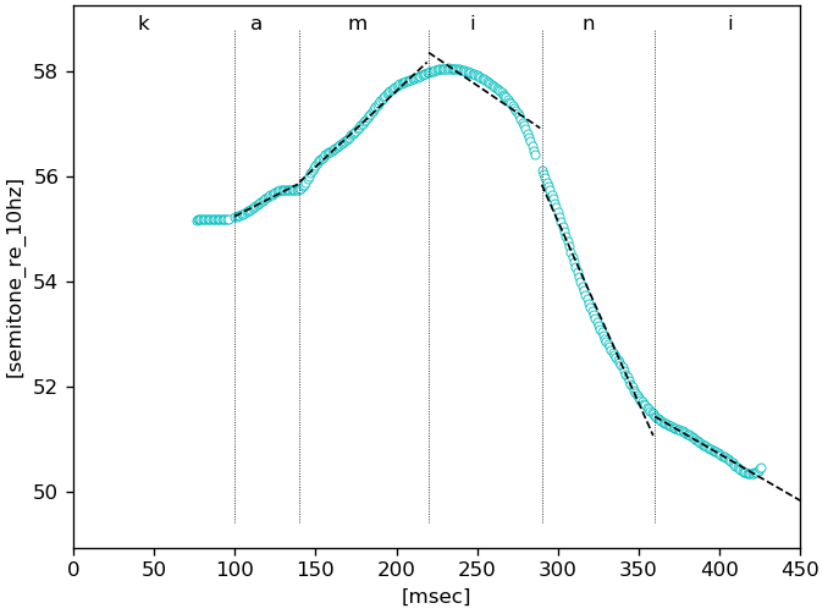
  （先頭音素は0%、末尾音素は100%の位置にも）

  各制御点を直線で連結してfo形状をモデリング

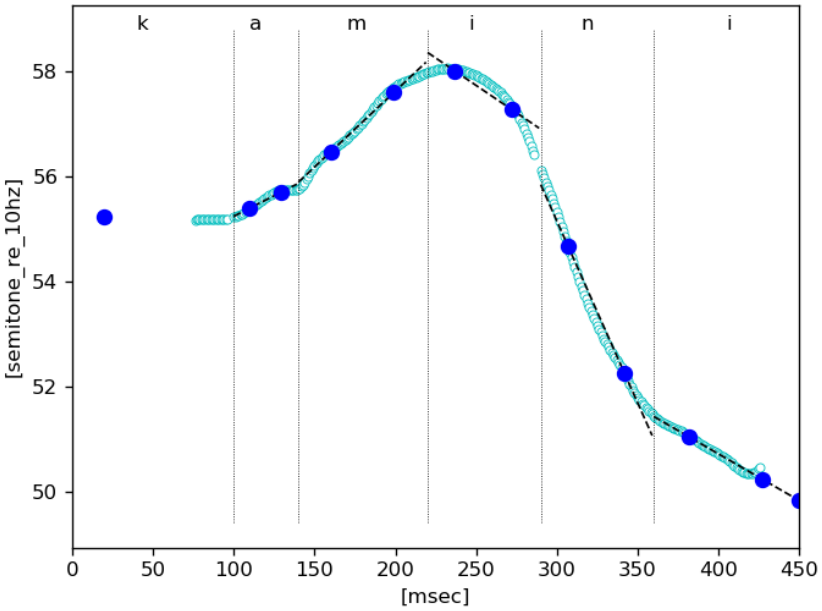
① Juliusとpraatで前処理



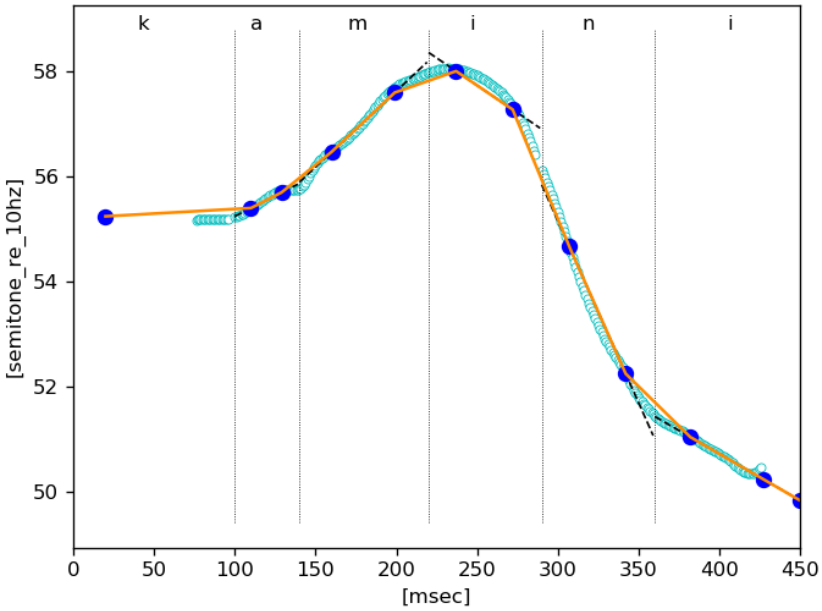
② 回帰直線を計算

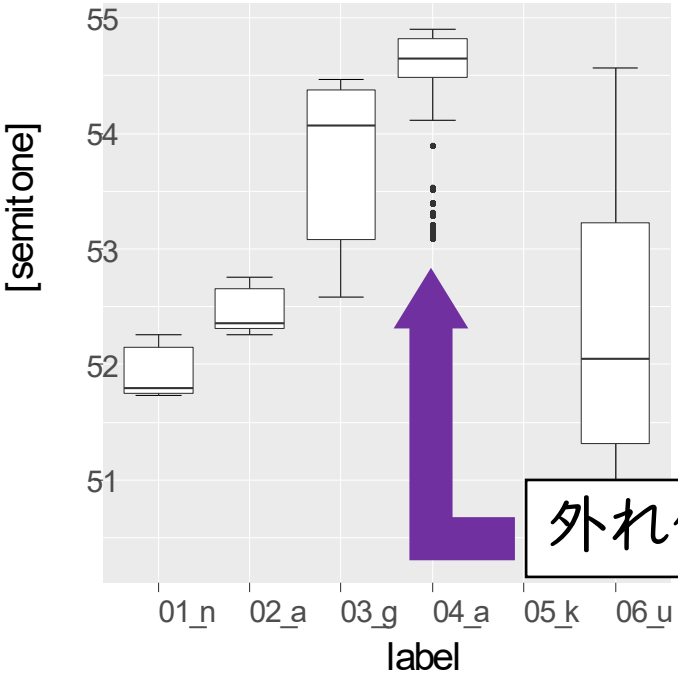


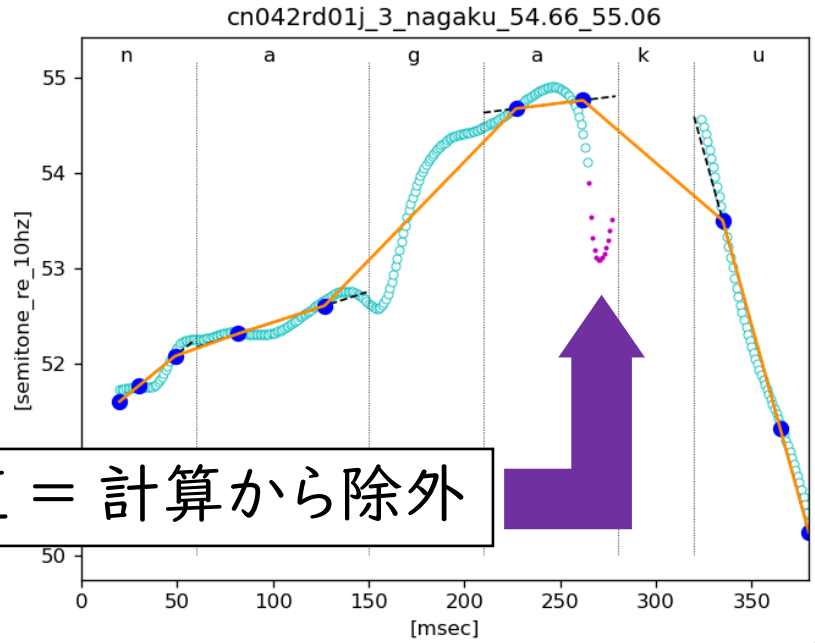
③ 制御点を設定



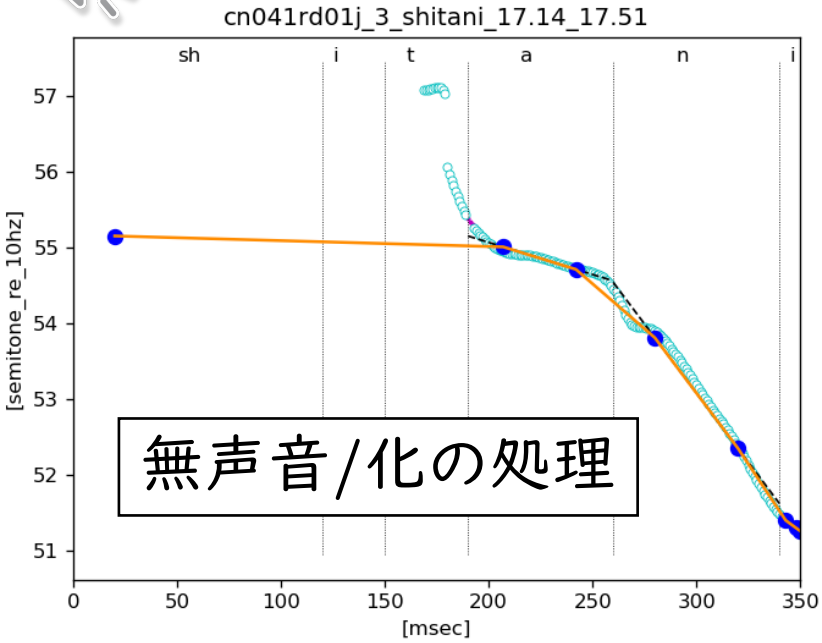
④ 直線で制御点を補間



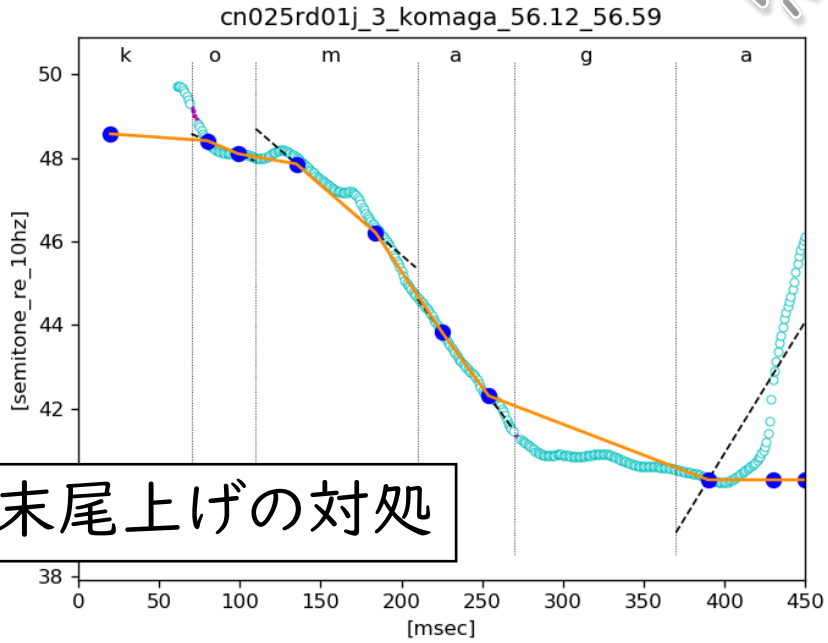




外れ値 ＝ 計算から除外



無声音/化の処理



末尾上げの対処

3. 特徴量の抽出

■ 5種類・7項目の特徴量を抽出

① 母音区間中間地点のfo差分（semitone）

② 母音区間終端地点のfo差分（semitone）

③ 単語・文節冒頭のfo値（Hz）

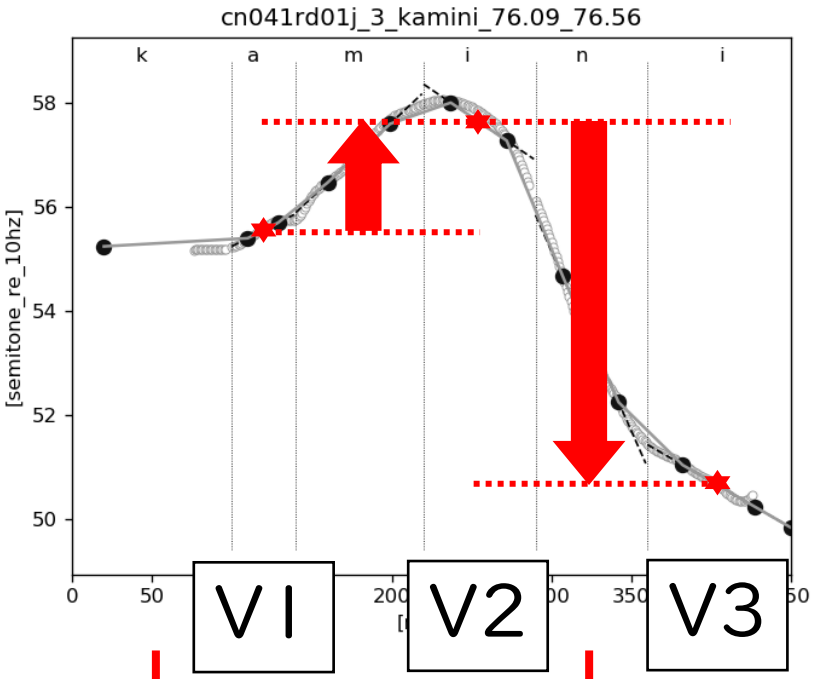
④ 単語・文節全体のfoレンジ（semitone）

⑤ 発話速度（mora/sec）

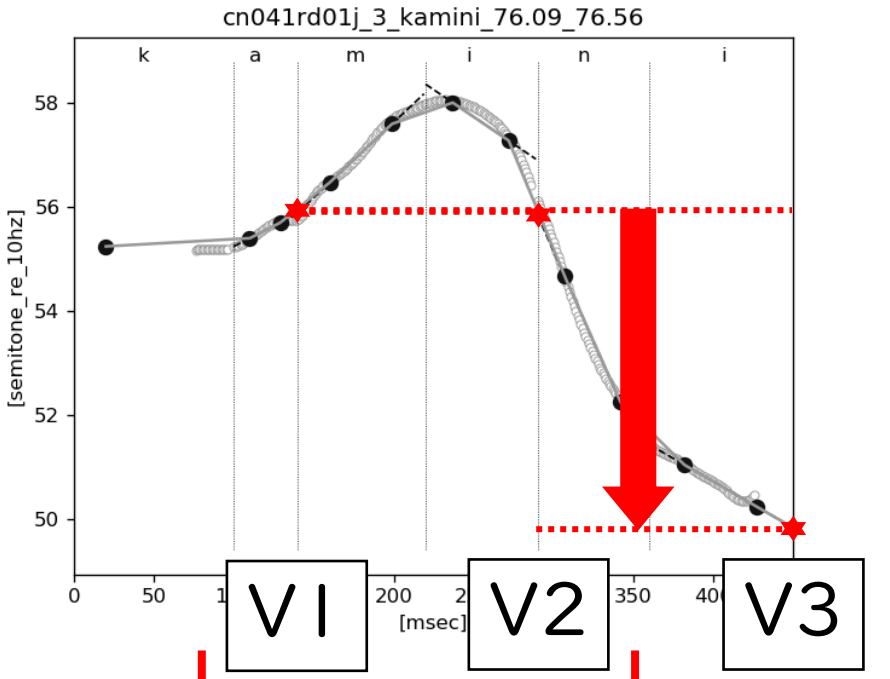
①と②は「V1V2」「V2V3」で計算
①は波多野・他（2014）参考、②は遅下がり対処（石井・他 2001）

母音区間のfo差分の例

①
中間地点の差分



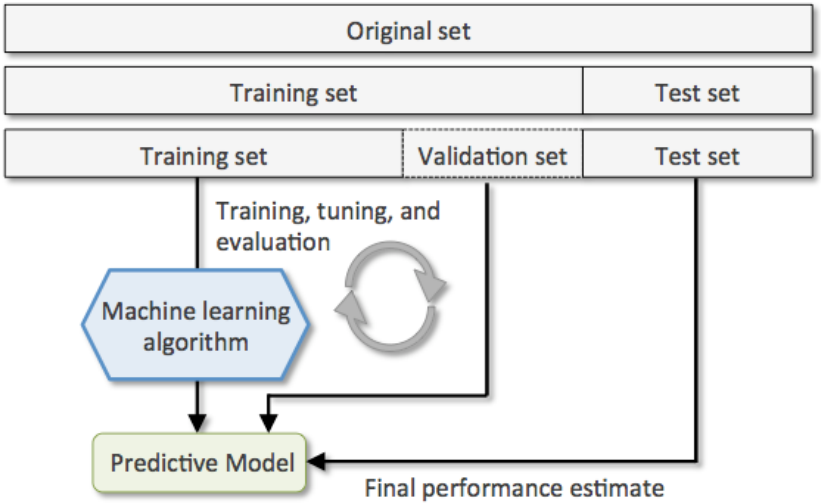
②
終端地点の差分



4. 機械学習の実行

■ SVC (Support Vector Classification)

- PythonのScikit-learnパッケージを使用
- 層化k分割交差検証を実施（今回は k=10）
- 各特徴量は標準化（平均0、標準偏差1）



(Raschka 2015)

※ トレーニングセットとテストセットの比率は8:2に設定

3. 結果と考察

単語発話データの分類結果

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	正解率
NS	480	384	99.7% (0.008)	96	100%
NNS	468	374	94.1% (0.023)	94	92.6%

（トレーニングセットは10回の交差検証における正解率の平均）

朗読発話データの分類結果

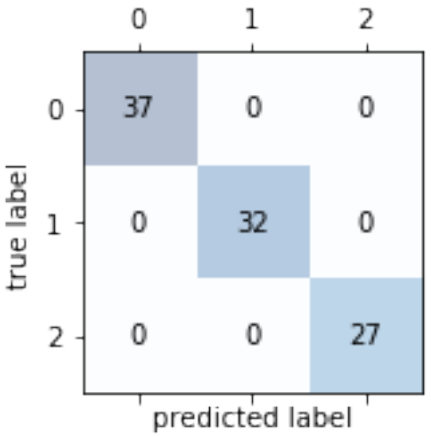
データ	n	トレーニングセット		テストセット	
		n	平均性能 (sd)	n	正解率
NS	250	200	83.0% (0.085)	50	76.0%
NNS	902	721	82.1% (0.029)	181	84.0%

(トレーニングセットは10回の交差検証における正解率の平均)

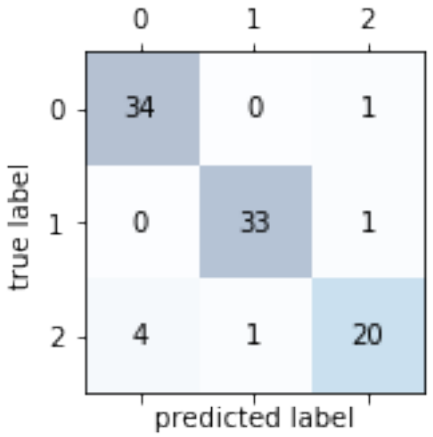
アクセント型ごとの分類性能

単語発話

NS

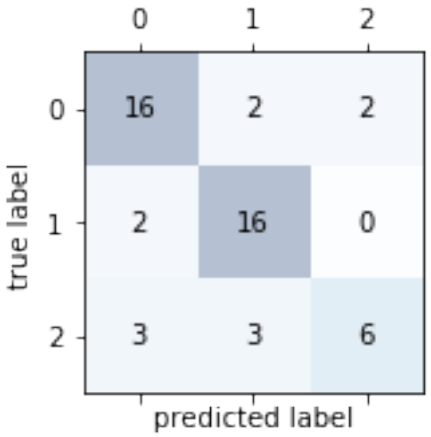


NNS

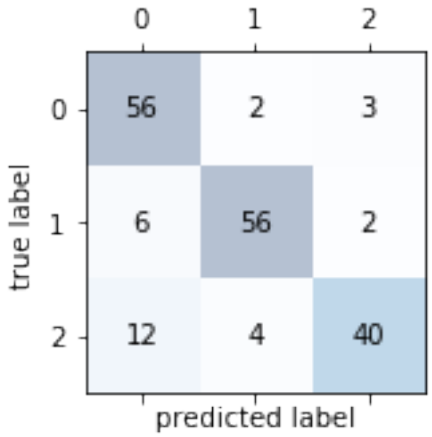


朗読発話

NS



NNS



■ 正解率（正解した予測ア型/その正解ア型の総数）

0型	100%	0型	97%	0型	80%	0型	92%
1型	100%	1型	97%	1型	89%	1型	88%
2型	100%	2型	80%	2型	50%	2型	71%

考察

- 概ね良好な正解率（特に単語発話）
  - 本手法はNS、NNSの発話にも有効
- 朗読発話では性能が落ちる
  - 1名によるアクセント評価の信頼性か
  - NSデータはサンプル数が少なく学習不足？
- 2型の正解率の低さ
  - V2V3のfo差をうまくモデリングできていない？
  - 朗読発話では相対的な出現数が少ない事が影響？

4. まとめ

本研究のまとめ

- 単語発話・朗読発話データを対象に、一定の基準でfo形状のモデリングを行ない、様々な特徴量を選定したうえで、機械学習を用いてアクセント型の自動分類を行った。

- 単語発話レベルでは高い分類精度となったが、朗読発話では改善点あり。

- 今後はサンプル数・モーラ数を拡充し、アクセント型自動判定としての本手法の妥当性を継続して検討したい。

謝辞

本研究はJSPS科研費 JP17H02352の助成を受けたものです。

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一部です。