

# アクセント制御音声合成と 主観比較評価最適化による 専門家非依存 アクセントアノテーション法

安田裕介（NII, 産総研），井本桂右（京都大学, 産総研），  
深山寛（産総研），戸田智基（名古屋大学）

# 研究動機

- アクセントの正しさは、ピッチアクセント言語である日本語の合成音声の品質を左右する.
- 正確なアクセントの再現にはアクセントラベルが必要.
- しかしアノテーションは言語学の専門家を必要としスケールしない.

→もし専門家に限らず一般の被験者でも、高精度かつ安価なアクセントアノテーションが実現すれば、アクセント教師ラベルの量を大幅に増やすことができるため、日本語の合成音声の品質を大幅に向上できるのでは？

# 提案手法

主たるアイデア：アノテーション問題を、「どちらがより自然なアクセントに聞こえる音声か？」というより単純な比較評価に落とし込む。

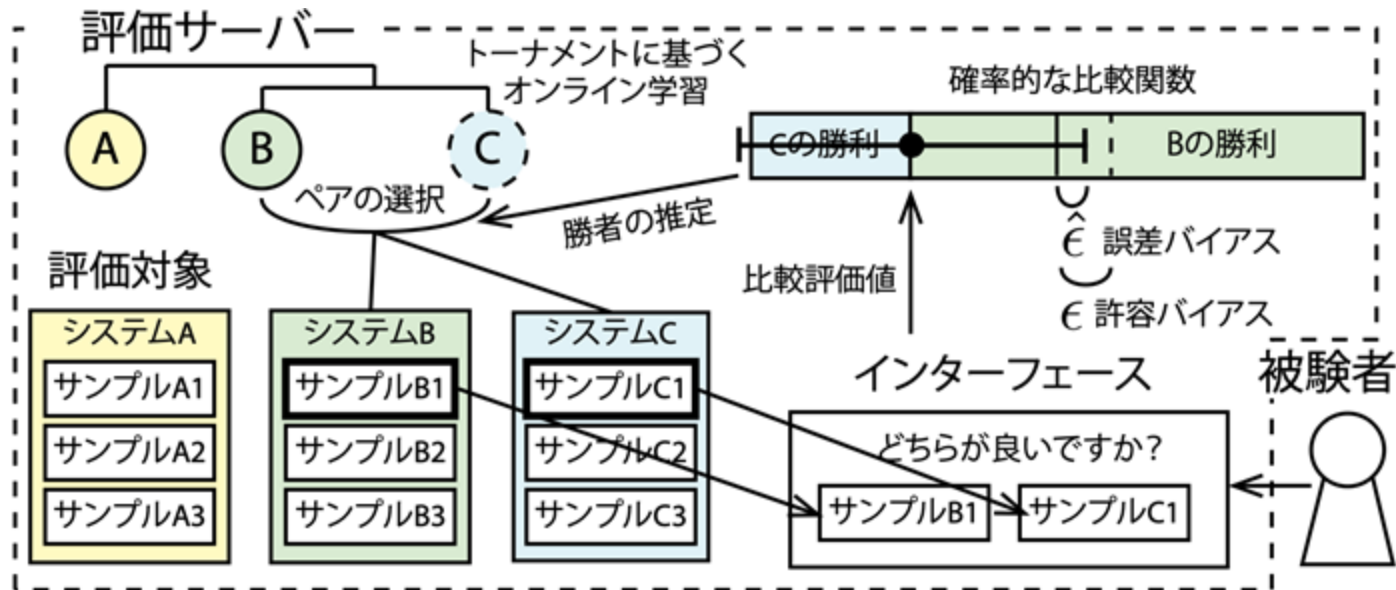
以下の3つの技術要素を組み合わせて、専門家非依存のアノテーションを実現：

- （１）目的のアクセントラベルを反映したアクセント音声をレンダリングできるアクセント制御音声合成器
- （２）複数のアクセント合成音声から最良のものを精度良く推定する主観評価手法
- （３）主観評価を効率的に行う最適化手法

クラウドワーカーといった非専門家による、  
スケールするアクセントアノテーションを、  
低コストかつ高精度で実現できることを期待

# 提案手法：システム構成概要

- 単語の全アクセントパターンを比較評価する
- トーナメントオンライン学習アルゴリズムを用いて、正解アクセントを推定



# オンライン学習アルゴリズム：TOURNAMENT

- 比較評価をもとに、優勝者を推定する
- 優勝者の推定に、トーナメントの勝ち抜き戦を行う
- $n$ 個の評価対象があると、 $n-1$ 回の比較で優勝者が求まる

---

## Algorithm 1 TOURNAMENT

---

**Input:** 集合 $S$ , 許容バイアス $\epsilon$ , 信頼度 $\delta$ .

```
if  $|S| = 1$  :  $s = S[0]$ ,  
else if  $|S| = 2$  :  $s = \text{COMPARE}(S[0], S[1], \epsilon, \delta)$ ,  
else :  
     $E = S[::2]$  (偶数番目の評価対象のみ取り出し)  
     $O = S[1::2]$  (奇数番目の評価対象のみ取り出し)  
     $e = \text{TOURNAMENT}(E, \epsilon, \delta)$   
     $o = \text{TOURNAMENT}(O, \epsilon, \delta)$   
     $s = \text{COMPARE}(e, o, \epsilon, \delta)$ 
```

**Output:**  $s$

---

# オンライン学習アルゴリズム：COMPARE

- 確率的な比較評価値から、勝者を推定する
- 比較平均が0.5よりも大きい方を勝者と決定する
- ある評価誤差を保証しながら最小の評価値を割り当てる

---

## Algorithm 2 COMPARE

---

**Input:** ペア  $(i, j)$ , 許容バイアス  $\epsilon$ , 信頼度  $\delta$ .

**Initialize:**  $\hat{p}_{ij} = \frac{1}{2}, m = \frac{1}{2\epsilon^2} \log \frac{2}{\delta}, r_{ij} = 0, w_{ij} = 0$ .

**Define:**  $\hat{c}(r) = \sqrt{\frac{1}{2r} \log \frac{4r^2}{\delta}}$  if  $r > 0$  else  $\frac{1}{2}$ .

**Define:**  $\hat{e}(r, \hat{p}) = \hat{c}(r) - |\hat{p} - \frac{1}{2}|$ .

**while**  $\epsilon \leq \hat{e}(r_{ij}, \hat{p}_{ij})$  and  $r_{ij} \leq m$  **do**

$(i, j)$ を比較. **if**  $i$ の勝利  $v_{ij} = 1$  **else**  $v_{ij} = 0$ .

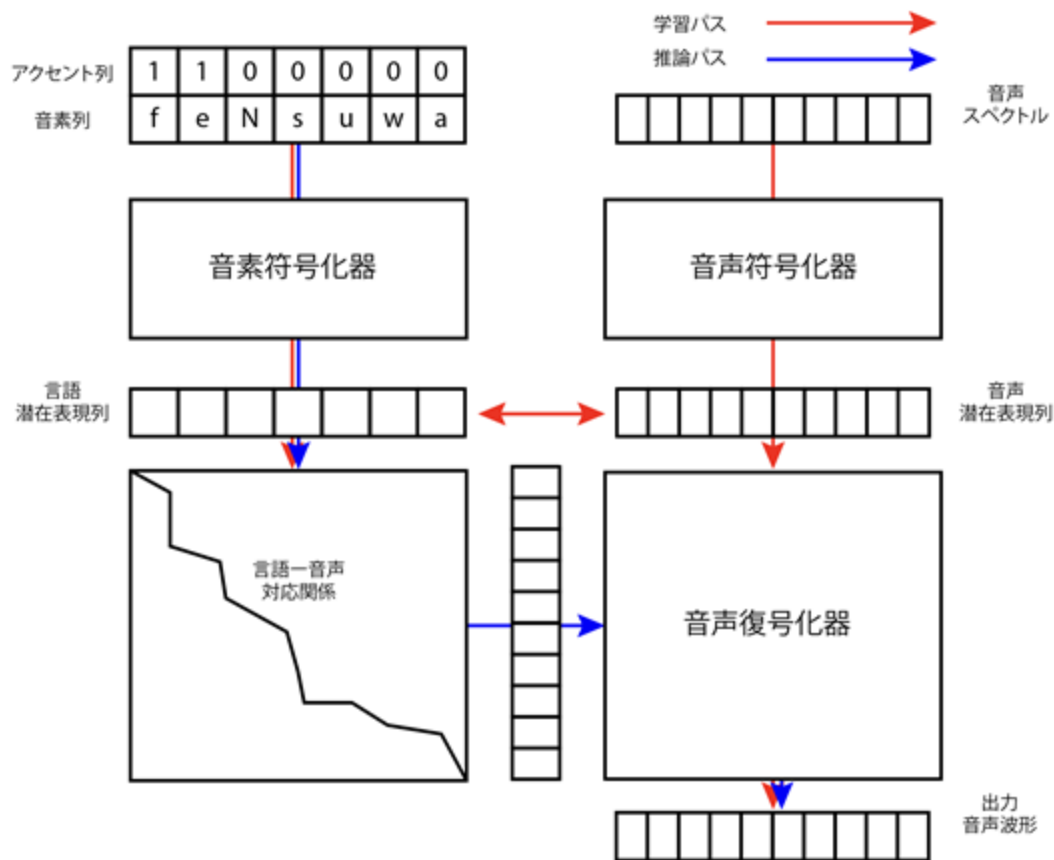
$w_{ij} = w_{ij} + v_{ij}, r_{ij} = r_{ij} + 1, \hat{p}_{ij} = \frac{w_{ij}}{r_{ij}}$ .

**if**  $\hat{p}_{ij} \leq \frac{1}{2}$  **Output:** 勝者  $j$  **else Output:** 勝者  $i$

---

# アクセント制御音声合成器：トーン条件つきVITS

- 変分自己符号化器に基づくエンドツーエンド音声合成手法VITSを用いる
- 入力にアクセント列を追加し、学習する
- アクセント列はトーン列で表現する
- 推論時に全アクセントパターンを合成する



# 実験条件：概要

- アクセント規則が既知な44単語を対象にアノテーション
  - 31個のアクセント規則
  - 2から9のモーラ数
- 音声コーパス中の発話文4文も対象に加え、発話文のアノテーションコストも計測
  - 12から21モーラ数
  - 4から8のモーラ数の12個の単語句
  - ATR Ximeraコーパスから選択
- オンライン学習の各ペアの最大評価数を150回に設定
- アクセント制御音声合成モデルの学習にJSUTコーパスを使用
- 音声合成器から単語の全アクセントパターンのサンプルを合成して評価
- 比較評価のためにクラウドソーシングを利用

# 実験条件：アクセント規則と単語例

| グループ             | 規則名     | 数 | 例       |
|------------------|---------|---|---------|
| 地域差              | -       | 1 | ありがとう   |
| 3 モーラ<br>名詞      | 起伏式     | 3 | 命       |
|                  | 平板式     | 1 | ねずみ     |
| 動詞               | 起伏式     | 2 | 成る      |
|                  | 平板式     | 2 | 鳴る      |
| 形容詞              | 起伏式     | 2 | 青い      |
|                  | 平板式     | 2 | 赤い      |
| -3の<br>規則        | 地名      | 1 | 秋田      |
|                  | 外来語     | 1 | オーストラリア |
|                  | 新語      | 1 | ソゴホジョモド |
|                  | 特殊モーラ   | 1 | 埼玉      |
| 複合語<br>アクセ<br>ント | 規則a     | 1 | 乳飲み子    |
|                  | 規則b     | 2 | 野ねずみ    |
|                  | 後部要素の保存 | 1 | ペルシャねこ  |

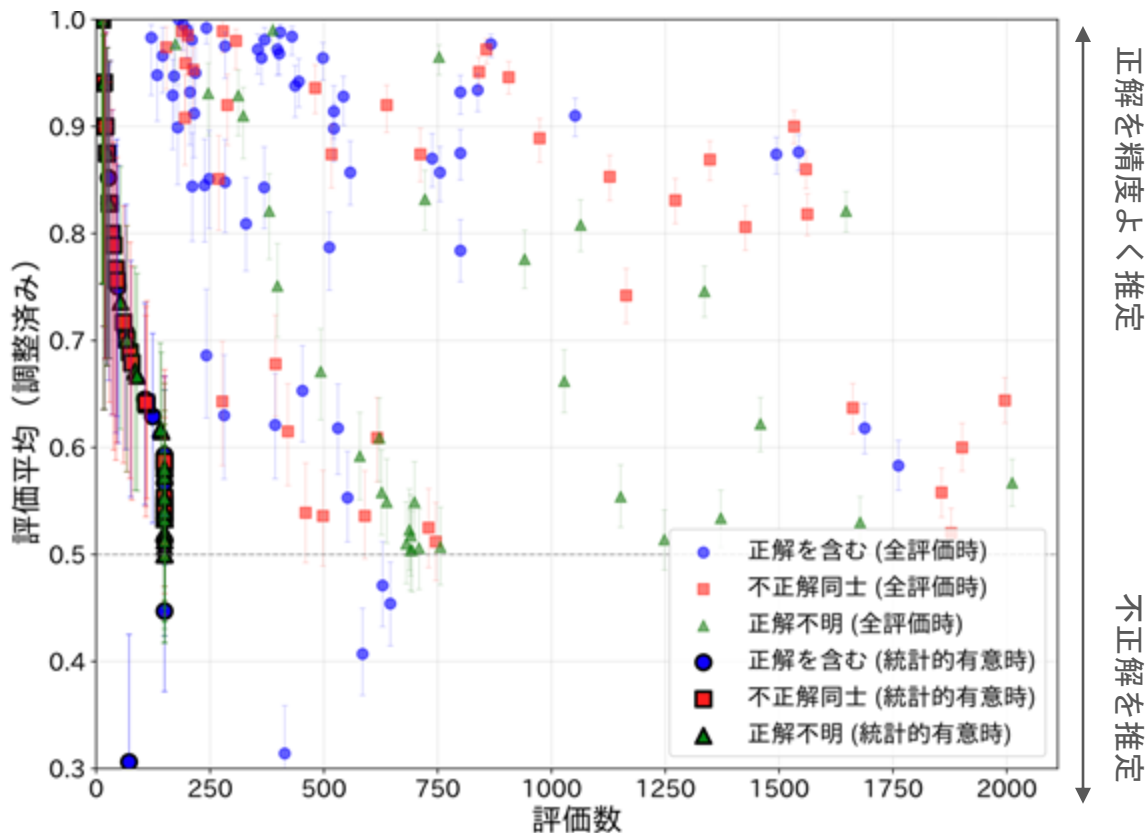
|                                   |            |    |         |
|-----------------------------------|------------|----|---------|
| -3の規則<br>の一致と<br>語末<br>音節<br>パターン | HLL        | 1  | エンジンバラ  |
|                                   | 不一致・LLH    | 1  | カラヤン    |
|                                   | 一致・LLH     | 1  | ピカチュウ   |
|                                   | 両方可・LLH    | 1  | エネルギー   |
|                                   | 不一致・HLH    | 1  | インタビュー  |
|                                   | 一致・HLH     | 1  | クーデター   |
|                                   | 両方可・HLH    | 1  | ミュージシャン |
| 弁別機能                              | 橋・箸・端      | 3  | 橋が      |
| 3 モーラ<br>の名前                      | 起伏式        | 1  | 春子      |
|                                   | 平板式        | 1  | 春美      |
| 一で終わる<br>名前                       | 規則a (起伏式)  | 1  | 勝一      |
|                                   | 規則b (平板式)  | 1  | 勝一      |
| 語種別<br>アクセント                      | 起伏式・和漢     | 3  | 命       |
|                                   | 平板式・和漢外    | 3  | ねずみ     |
| 複合語<br>全体                         | 起伏式        | 2  | 秋田県     |
|                                   | 平板式        | 2  | 秋田犬     |
| 外来語                               | 平板式音形      | 3  | カリフォルニア |
| コーパス                              | ATR Ximera | 12 | 悲しみの... |

# 実験結果：概要

- 言語学の非専門家を用いても、低コストで高精度なアノテーションを実現
  - 音声を聞いてアクセントが自然かどうかを判断することは、音声を解析してアクセント型を同定するよりも、容易
  - 被験者は自然なアクセントを強く選好する
- ペアレベルでの正解アクセント型の推定精度：90%以上の高精度
- 単語レベルでの正解アクセント型推定精度：80%以上の高精度
- ペアレベルでの必要な評価数：10から100
- 単語レベルでの必要な評価数：多くても200程度

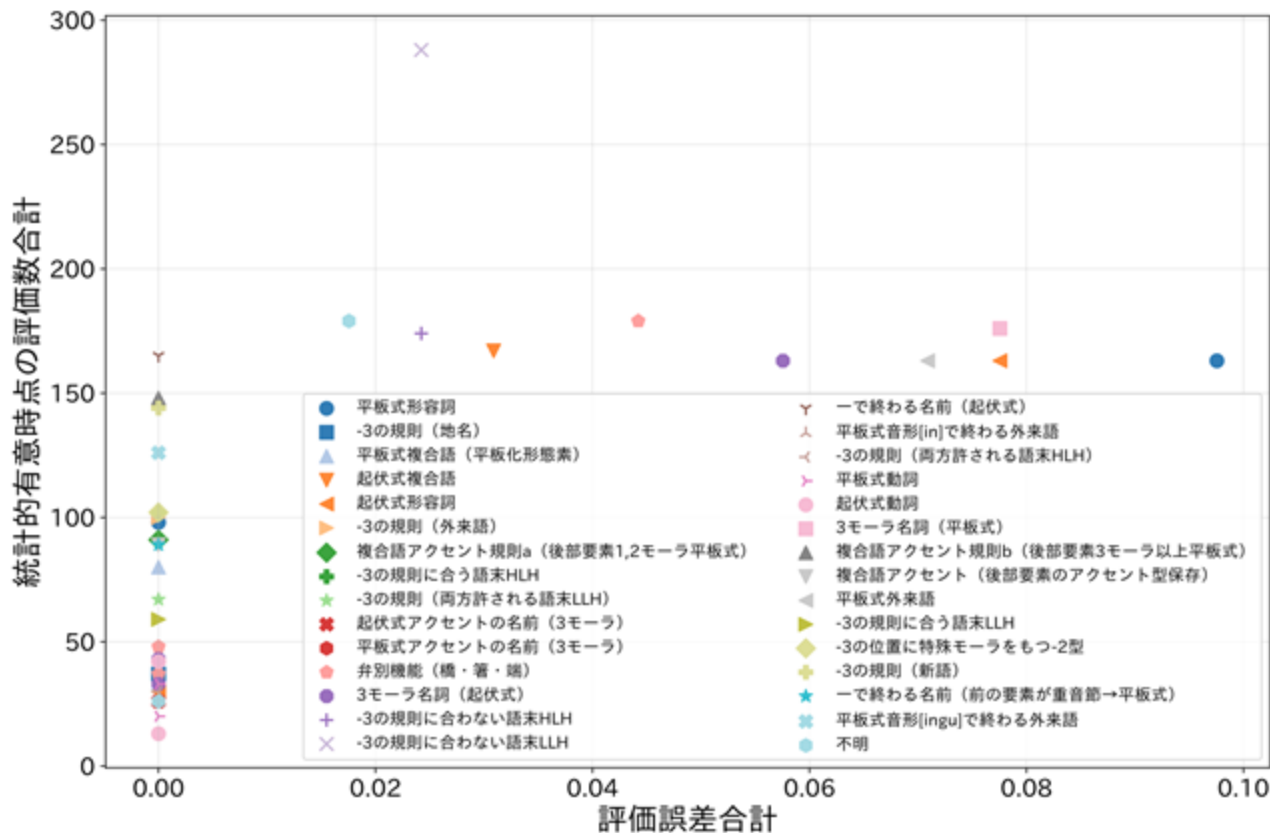
# 実験結果：アクセント型ペアの評価平均と評価数

- ペアの正解アクセント型の推定精度：90%以上の高精度
  - 全評価数を用いた場合：93.1%
  - 統計的有意が確定した評価数の時点：96.6%
- 評価数が少なくても正しいアクセントを推定
  - 最小の評価数の場合：わずか13回の評価数で正解アクセント型を推定
  - 被験者は正解アクセントに強い選好を示した
  - 不正解アクセント同士の比較では、どちらが勝っても推定精度に影響なし
- 評価数を増やしたとしても、推定精度が向上するわけではなかった
  - 信頼区間は縮小するが、推定精度に寄与しない



## 実験結果：単語単位の評価誤差合計と評価数合計

- 高い単語レベルでのアクセント型推定精度
  - 85.2%
- 評価誤差の合計は、多くの単語でゼロ
- 1単語をアノテーションするのに必要な評価数：
  - 多くても200評価数程度
- アクセント規則に対する傾向：なし



# まとめ

- アクセント制御音声合成とトーナメントに基づく動的比較評価を組み合わせることで、専門家非依存のアクセントアノテーション法を提案
- 提案手法が高精度かつ低コストであることを実験で検証
- 専門家に頼らずに一般の被験者でも高精度かつ安価なアクセントアノテーションを実現