2024-03-08 日本音響学会第151回（2024年春季）研究発表会　3-2-4

# F0 に基づいて伸縮された画像文字からの音声合成

☆**大中 緋慧**，宮崎 亮一（徳山高専），
高道　慎之介（東大院・情報理工）

Miyazaki-Lab.

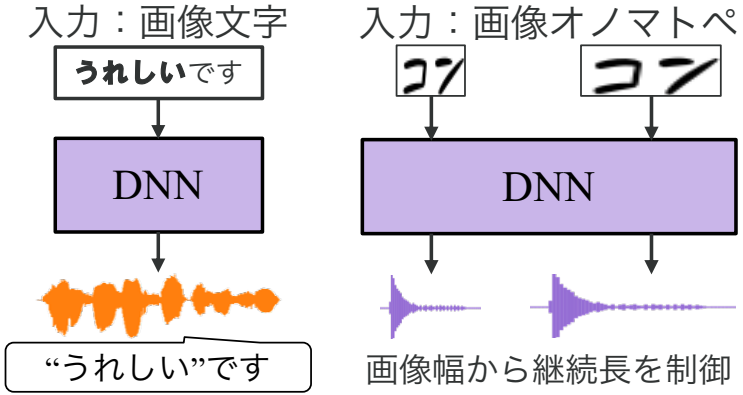
Department of Computer Science and Electronic Engineering, National Institute of Technology, Tokuyama College

# 研究背景

- **近年のテキスト音声合成 (text-to-speech: TTS)**
  - 読み上げ音声の合成の自然さは人に近いレベルまで発展 [Shen+, 2018]
    - ✓ 応用先：発音トレーニングにおける TTS の活用 [Garcia+, 2020]
  - より多様な音声を合成可能とすることを目指す研究の増加
    - ✓ 中間表現として予測される F0 の手動制御による合成音声の操作 [Ren+, 2021]
    - ✓ 自然言語による発話スタイルの制御 [Guo+, 2023] など
    - ✓ 応用先：動画コンテンツ制作などのユーザ自身が音声を制御する（音声デザイン）状況下における TTS

# 研究背景：画像文字を利用した音声システム

- **画像文字を入力とする音合成**
  - visual-text to speech [Nakano+, 2023]
    - ✓ 画像文字の強調属性を音声に反映
  - Visual onoma-to-wave [Ohnaka+, 2023]
    - ✓ 画像文字の伸縮に基づく継続長の制御




- **韻律に基づく変形画像文字を用いた発音指導 [Rude, 2012]**
  - 音声の長さ，高さ，大きさに基づいて文字を変形
  - **変形された画像文字を提示した発音指導により発音が改善**

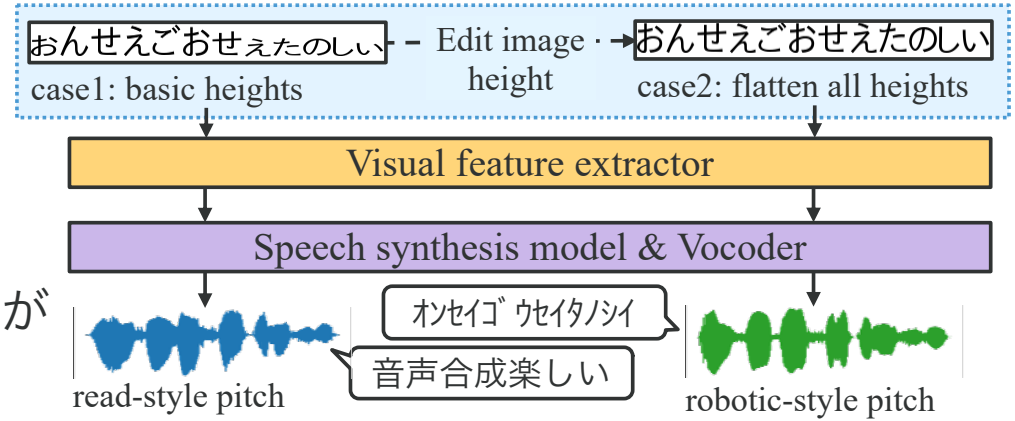

[Rude, 2016]の **Fig. 1** より引用

# 本研究の概要

- **動機**
  - オノマトペの伸縮表現や発音指導における変形画像文字の有効性に着目
    → 画像文字を介した韻律制御が可能な音声合成によって
    より良い音声デザインのための TTS や発音指導が実現できるのでは

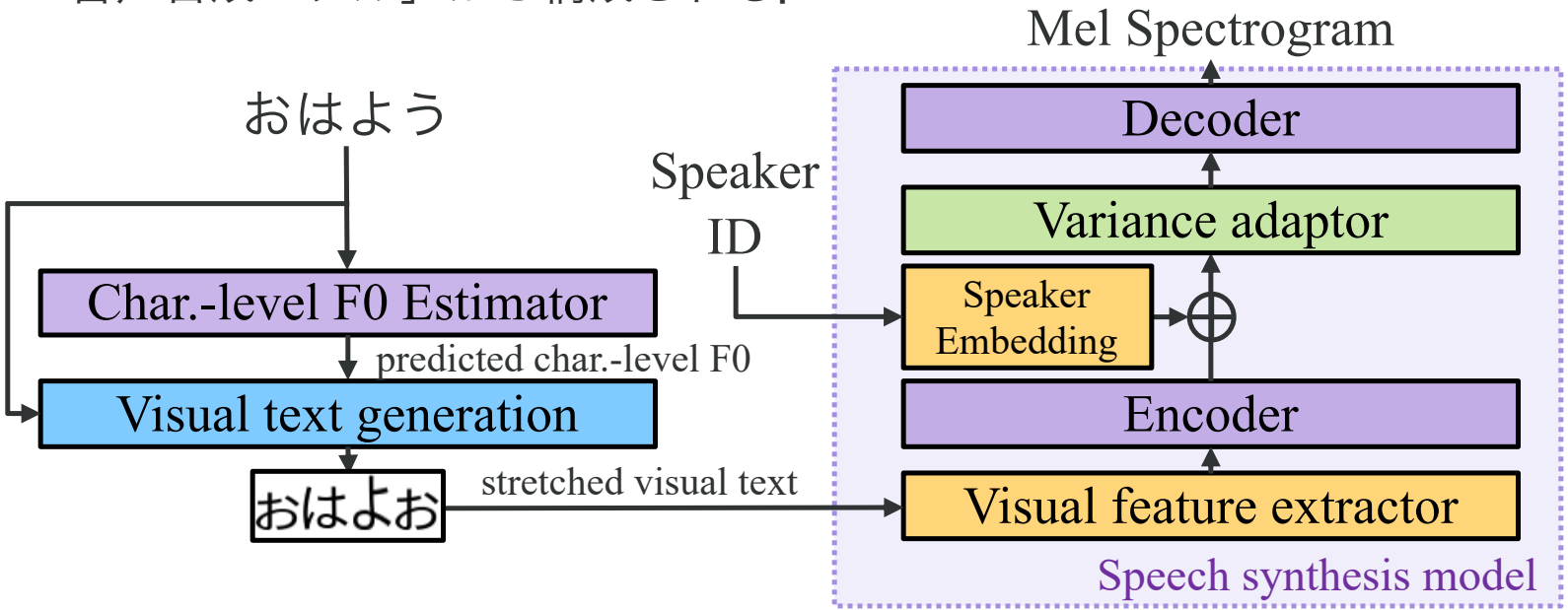
- **提案手法：画像高さの伸縮に基づく F0 制御可能な音声合成**
  - 画像高さの伸縮に基づき合成音声の F0 を視覚的に手動制御可能

- **実験**
  - 既存手法よりも
    優れた F0 制御性能
  - 主観的にも画像文字と合成音声が
    よく対応することを確認



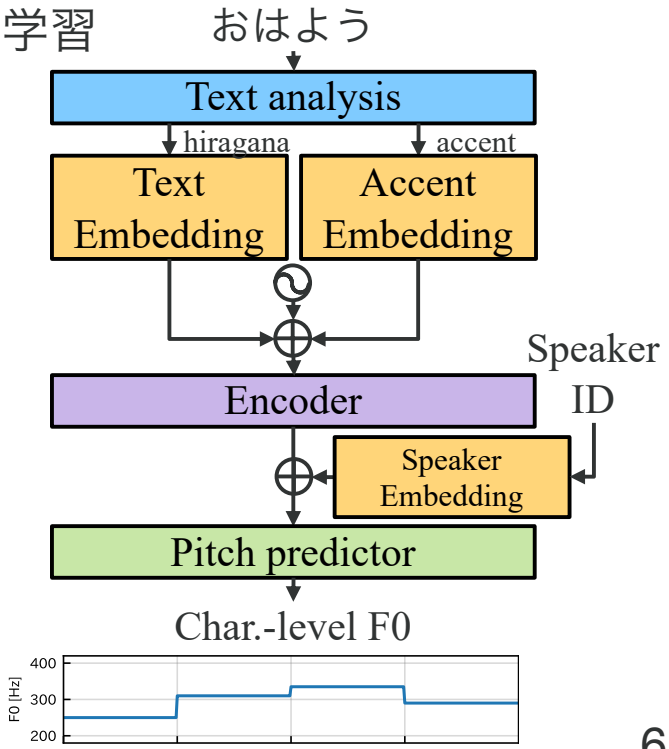
# 提案手法：全体像

- **入力文字を F0 に基づいて伸縮された画像文字に変換しそれを入力することでメルスペクトログラムを出力**
  - モジュールは「文字単位 F0 予測器」「画像高さ伸縮器」「音声合成モデル」から構成される.



# 提案手法：文字単位 F0 予測器

- **与えられたテキストから読み上げ音声に対応するF0 を推定するモジュール**
  - 学習：推定 F0 と真の F0 の平均二乗誤差による学習
  - 推論：読み上げスタイルの F0 を出力
- **モデル構造**
  - 入力：テキストとアクセントラベル
  - FastSpeech2 と同様の encoder と variance adaptor の pitch predictor を使用



# 提案手法：画像高さ伸縮器

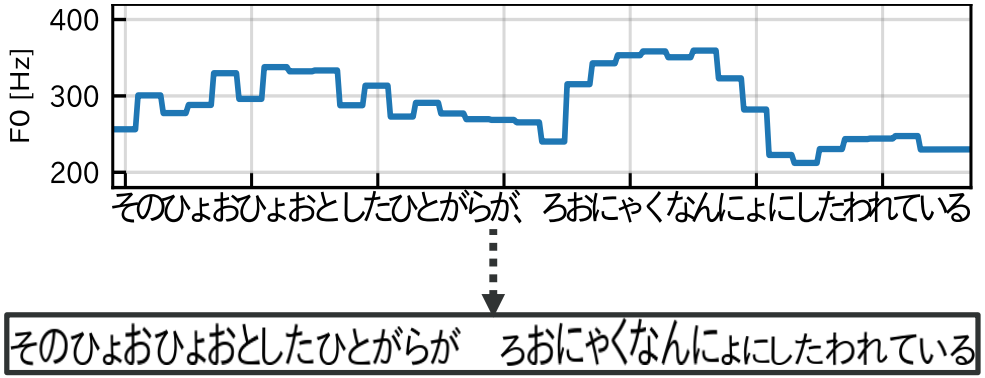
- F0 とテキストを受け取り高さが伸縮された画像文字を生成

- 生成方法
  - データセット内の平均 F0 $f_{\mathrm{mean}}$ を基準に伸縮後の画像高さ $h_i$ を F0 $f_i$ から決定

$$h_i = \frac{\mathcal{M}(f_i)}{\mathcal{M}(f_{\mathrm{mean}})} H$$

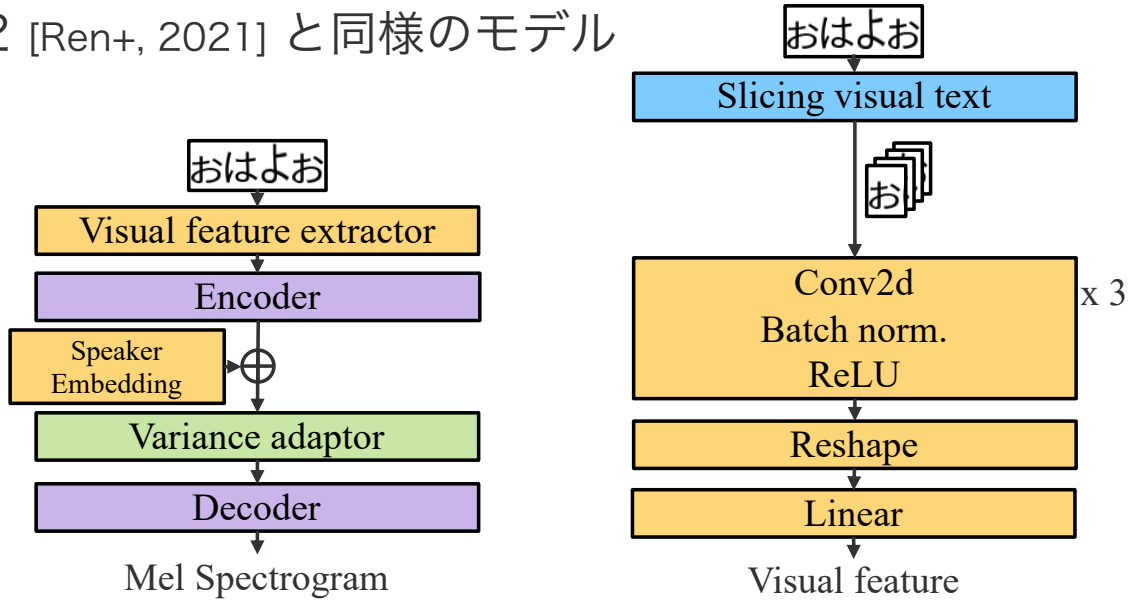
$H$: デフォルトの高さ
$\mathcal{M}(\cdot)$: メルスケール変換

  - F0 が高いほど伸びた画像になる.
  - 学習時：真の F0 から伸縮を決定
  - 推論時：文字単位 F0 予測器から得られた F0 から伸縮を決定



# 提案手法：音声合成モデル

- **Visual-text to speech [Nakano+, 2023]と同様のモデル**
  - 最初段：Visual feature extractor
    - ✓ 文字形状（音韻）と伸縮度合い（F0）の両方を捉えることを期待
  - 後段：FastSpeech2 [Ren+, 2021] と同様のモデル




(a) Speech synthesis model　(b) Visual feature extractor

# 評価実験

- **目的**
  - 合成音声の品質（自然性，F0 制御性）の評価
  - 画像の伸縮の変化と合成音声の F0 の変化の対応度合いの評価

- **三つの実験を実施**
  1. 自然性MOS & CER による基本品質評価
  2. F0 制御性の客観評価
  3. 手動制御における画像高さの伸縮と
     合成音声 F0 の変化の対応度合いの主観評価

# 実験条件

| | | |
|---|---|---|
| 使用データ | データセット | **JVSコーパス [Takamichi+, 2019] 97話者** |
| | データ数 | 学習 12,139 発話，推論 291 発話，テスト 291発話 |
| 画像文字フォント | | 等幅IPAexGothic （フォントサイズ 34, $H$ は 36 px） |
| 特徴量抽出 | アライメント | JVSコーパスで提供された音素アライメントを使用 |
| | F0 抽出 | WORLD [Morise+, 2016] （$f_{\mathrm{mean}}$ : 196.2 [Hz]） |
| | 音響特徴量 | 80 次元のメルスペクトログラム |
| モデル設定 | Visual Feature Extractor | カーネルサイズ (21, 5) の CNN |
| | 波形生成 | 事前学習済み Hifi-GAN [Kong+, 2020] |
| 比較手法 | | **ひらがな入力 Fastspeech2 [Ren+, 2021]<br>（variance adaptor の予測 F0 の手動制御が可能）** |

# 基本品質の評価

- **比較手法と提案手法のスコア表**
  - Naturalness: 日本語母語話者 40 名による自然性 MOS
  - CER: Whisper [Radfold+, 2023] baseモデルを用いたひらがな単位の文字認識率

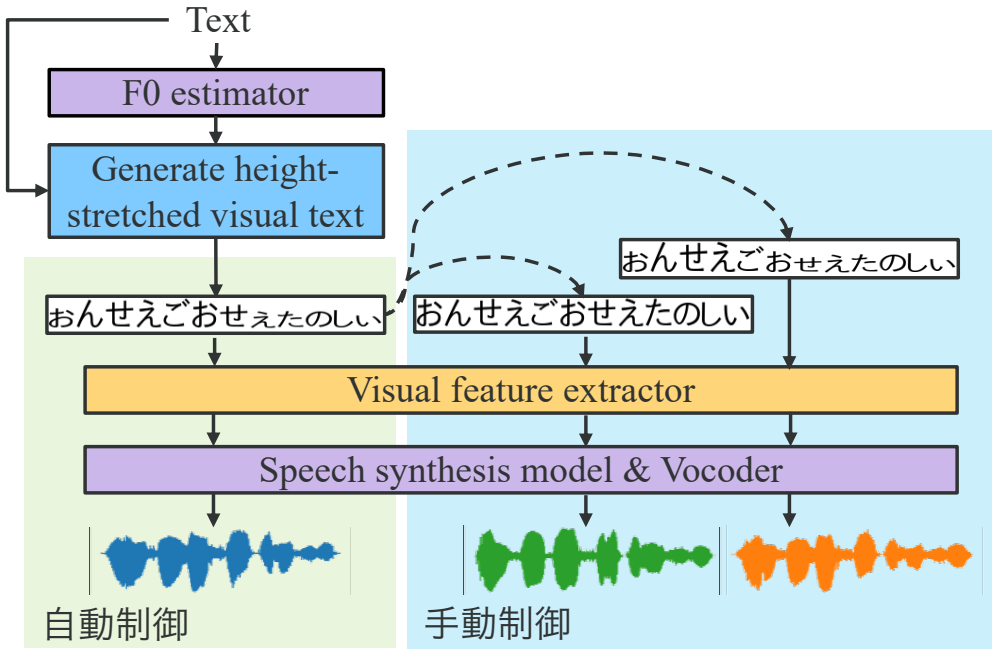
| | Naturalness (↑) | CER [%] (↓) |
|---|---|---|
| Conventional (text-input) | $3.29 \pm 0.105$ | 11.85 |
| Proposed (visual text-input) | $3.16 \pm 0.104$ | 12.87 |

**従来手法と比較して自然性・CER の両方でやや劣る**

→ 画像の伸縮により文字形状が変化するため
テキスト入力と比較して音韻情報がぼやけ
発話内容の明瞭度が低下することが原因と考えられる

# F0 制御性の客観評価

- **二つの観点で評価**
  1. 自動制御性
     - ✓ 自動生成された画像文字に対応する F0 の音声を合成できるか
  2. 手動制御性
     - ✓ 画像文字の高さを加工した際に F0 がその加工に追従するか

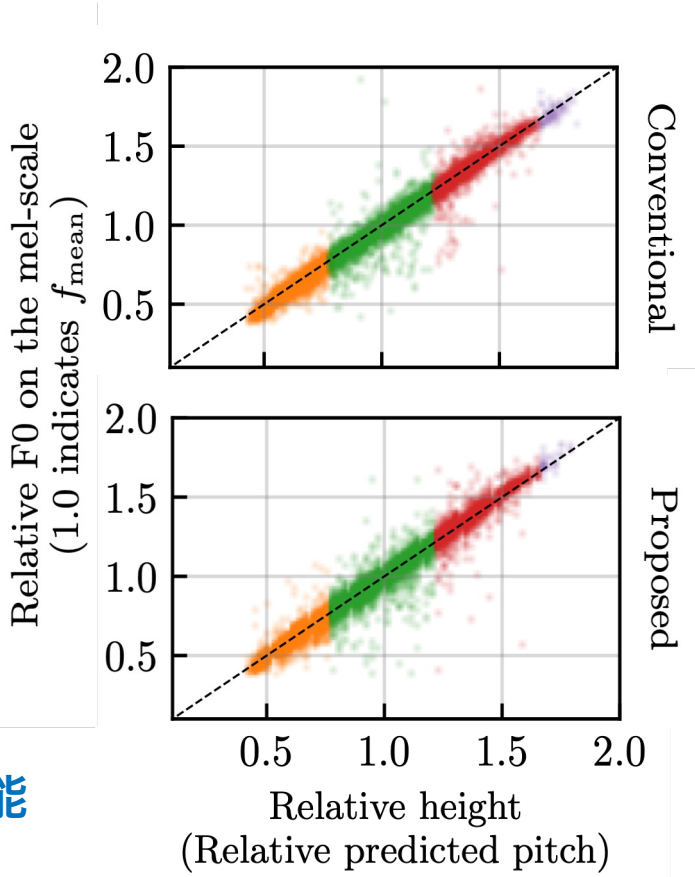



- ➢ 比較手法ではVariance adaptor から予測された F0 系列の操作による F0 制御を用いた.

# F0 制御性の客観評価：自動制御性

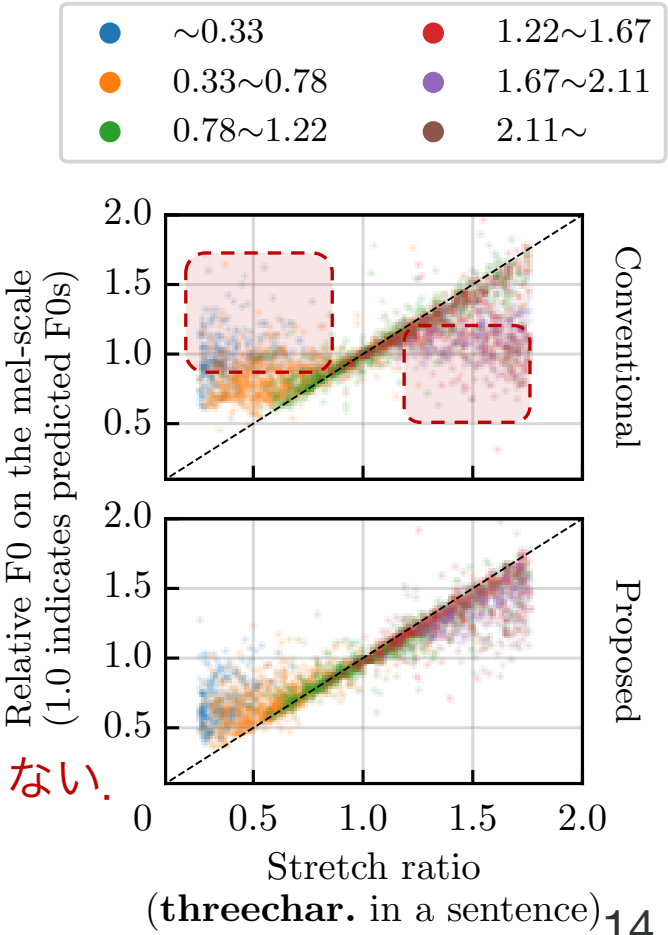
- 各文字を1点とした散布図
  - 横軸
    - ✓ 比較手法：$f_{\mathrm{mean}}$ に対する中間表現 F0 のメルスケール比
    - ✓ 提案手法：デフォルトの高さ $H$ に対する各文字の高さ $h_i$ の比
  - 縦軸：$f_{\mathrm{mean}}$ に対する合成音声の F0 のメルスケール比

- 結果
  - **提案手法，比較手法ともに与えられた F0 情報を合成音声へ適切に反映可能**



# F0 制御性の客観評価：手動制御性

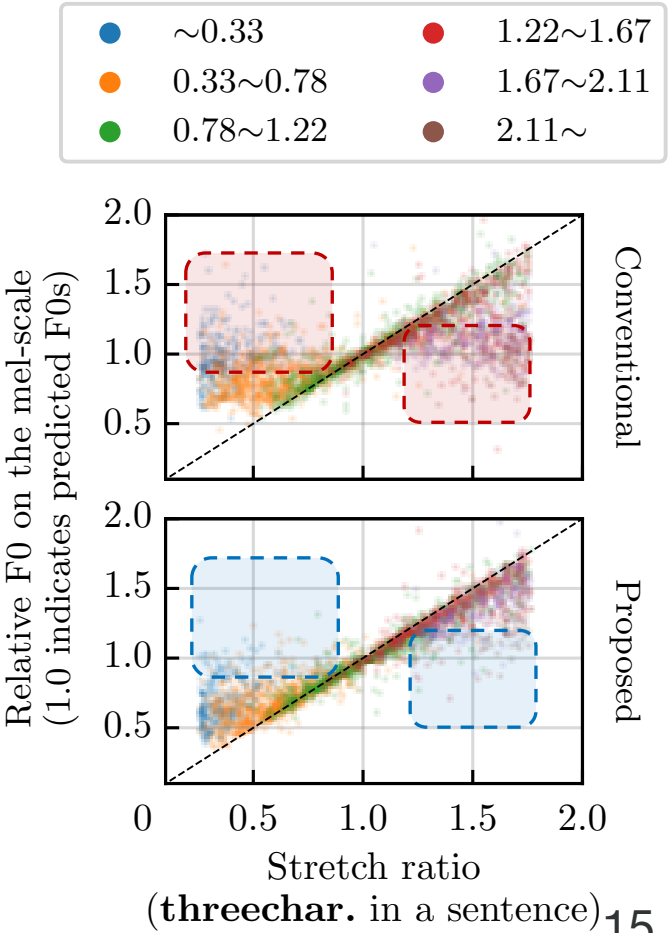
- 横軸
  - ✓ 比較手法：予測された F0 系列のうち
    連続するいずれかの3文字分を乗じた割合
  - ✓ 提案手法：画像文字のうち
    連続するいずれかの3文字を伸縮した割合
- 縦軸：手動制御ありとなしの場合の
  合成音声 F0 のメルスケール比
- ラベル：高さ $H$ と変化後の高さの比




- **結果**
  - 比較手法では変化の度合いが大きく絶対的な
    数値が $f_{\text{mean}}$ から離れた際に合成音声 F0 が追従しない.
    → 低い（高い）F0 をより低く（高く）する際の
    正確性が低いことを意味する.

# F0 制御性の客観評価：手動制御性

- 横軸
  - ✓ 提案手法：画像文字のうち3文字を伸縮（手動制御）した割合
  - ✓ 比較手法：予測された F0 のうち3文字分を乗じた割合
- 縦軸：手動制御ありとなしの場合の合成音声 F0 のメルスケール比
- ラベル：高さ $H$ と変化後の高さの比




- **結果**
  - **提案手法ではこの問題が緩和されよく追従している.**
    **→ F0 制御性の観点で従来手法よりも優れた結果**

# 音声サンプル

- **自動制御による合成音声**

ぁなたのことゎ　ぃろぃろぅかがっております

- **手動制御による合成音声**
  - 文末での上昇

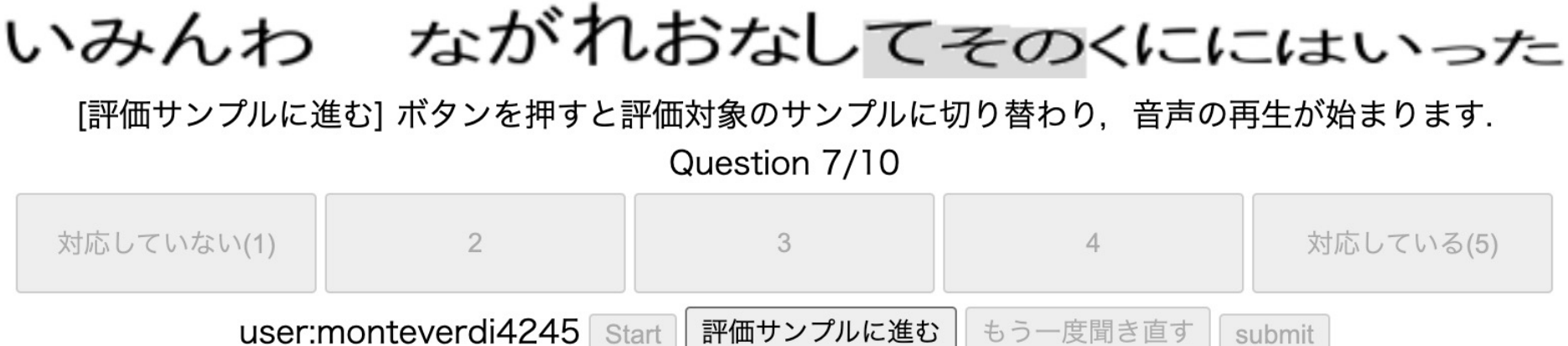
ぁなたのことゎ　ぃろぃろぅかがっております

  - 均一な高さ（robotic-style)

あなたのことわ　いろいろうかがっております

# 画像の変化と合成音声の F0 変化の対応

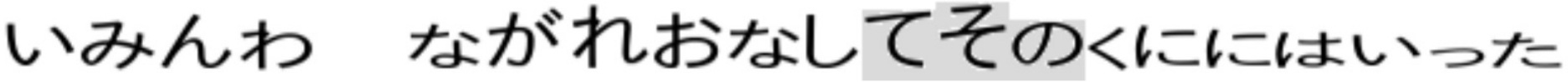
- **目的：画像と合成音声 F0 の変化の対応度の主観的な評価**
  0. 画像高さが音声 F0 に対応する旨を事前に評価者に伝達
  1. 自動推定された画像文字とその合成音声を提示



# 画像の変化と合成音声の F0 変化の対応

- **目的：画像と合成音声 F0 の変化の対応度の主観的な評価**

2. 3文字が手動加工された画像文字と次のいずれかの合成音声を提示
   - ✓ Proposed: 手動加工された画像文字からの合成音声
   - ✓ Control: 1. と同様の合成音声




[1] から [5] を選択してください ([もう一度聞きなおす] ボタンで2つの音声をもう一度再生できます).

Question 7/10

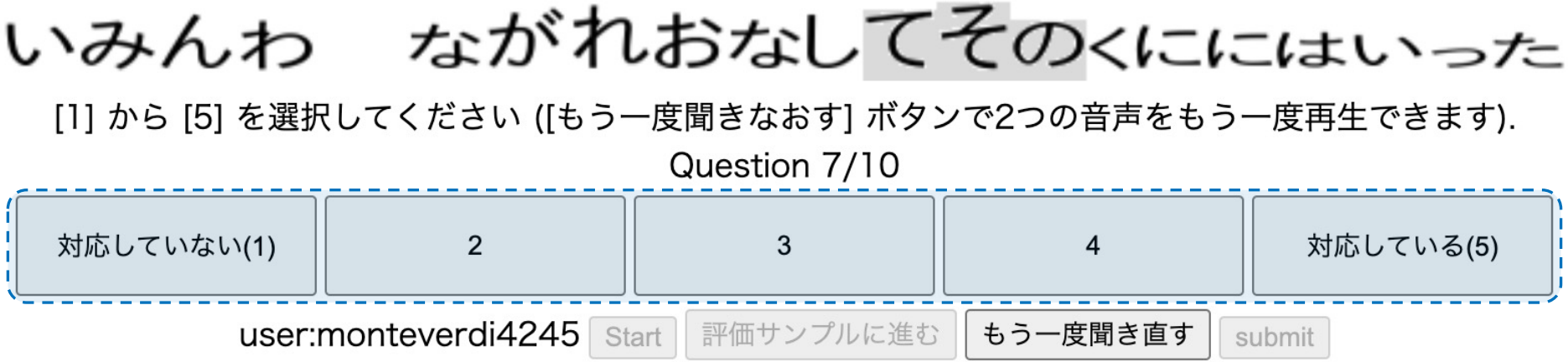
| 対応していない(1) | 2 | 3 | 4 | 対応している(5) |
|---|---|---|---|---|

user:monteverdi4245 Start 評価サンプルに進む もう一度聞き直す submit

# 画像の変化と合成音声の F0 変化の対応

- **目的：画像と合成音声 F0 の変化の対応度の主観的な評価**

  3. 評価者は画像の変化と音声の高さの変化の対応度合いを五段階で評価



# 画像の変化と合成音声の F0 変化の対応

- 評価結果

| | Correspondence |
|---|---|
| Proposed | $\mathbf{3.73 \pm 0.179}$ |
| Control | $3.00 \pm 0.187$ |

**提案手法において高いスコアが得られた.**
**→ 画像の高さの変化と音声の F0 の変化が主観的によく対応することを確認**

# まとめ

- **目的：画像文字を介した韻律制御可能な音声合成の実現**
- **提案手法：画像高さの伸縮に基づく F0 制御可能な音声合成**
- **評価実験：提案手法の基本性能を調査**
  - 従来手法と比較して自然性でやや劣るが F0 制御性で優れた性能
    → **画像文字に基づく合成音声の F0 制御が可能であることを確認**
- **今後の展望**
  - 提案手法の音声デザインにおける有効性を調査するための実験
  - 音量や継続長まで含めた変形画像文字を用いた手法への拡張と
    発音指導における有効性の調査　など

# メモ：画像の変化とF0変化の対応のスコア分布