

## 1. はじめに

- 合成音声の自然性は向上した
  - 対話に用いると単調で無機質
- 相手に合わせ話し方を変えたい
  - **音素単位で話速**を制御できる音声合成を行う



## 2. 先行研究 [1]

“Phonemic-level Duration Control Using Attention Alignment for Natural Speech Synthesis”

- Tacotronを使用
- very slow, slow, fast, very fastの4種類の固定長の話速で音声を合成
- **本研究との違い**
  - 本研究では話速を指定する記述から自動的に音素の継続長を決定する。
  - 本研究では異なる話速を含むデータセットを構築し、学習に用いた。

## 3. 提案手法

### コンテキストラベル

$p_1 \hat{p}_2 - p_3 + p_4 = p_5 @ p_6 - p_7$   
 /A: a<sub>1</sub>-a<sub>2</sub>-a<sub>3</sub> /B: b<sub>1</sub>-b<sub>2</sub>-b<sub>3</sub>@b<sub>4</sub>-b<sub>5</sub>&b<sub>6</sub>-b<sub>7</sub>#b<sub>8</sub>-b<sub>9</sub>\$b<sub>10</sub>-b<sub>11</sub>!b<sub>12</sub>-b<sub>13</sub>;b<sub>14</sub>-b<sub>15</sub>|b<sub>16</sub>  
 /C: c<sub>1</sub>+c<sub>2</sub>+c<sub>3</sub>  
 /D: d<sub>1</sub>-d<sub>2</sub> /E: e<sub>1</sub>+e<sub>2</sub>@e<sub>3</sub>+e<sub>4</sub>&e<sub>5</sub>+e<sub>6</sub> #e<sub>7</sub>+e<sub>8</sub> /F: f<sub>1</sub>-f<sub>2</sub>  
 /G: g<sub>1</sub>-g<sub>2</sub> /H: h<sub>1</sub>=h<sub>2</sub> @h<sub>3</sub>=h<sub>4</sub>|h<sub>5</sub> /I: i<sub>1</sub>=i<sub>2</sub>  
 /J: j<sub>1</sub>+j<sub>2</sub>-j<sub>3</sub> /S: s<sub>1</sub>

### 話速制御のための2つの手法を提案

#### 音素記号の拡張

$p + \begin{matrix} S \\ N \\ F \end{matrix}$

音素記号    話速タグ

S: Slow, N: Normal, F: Fast

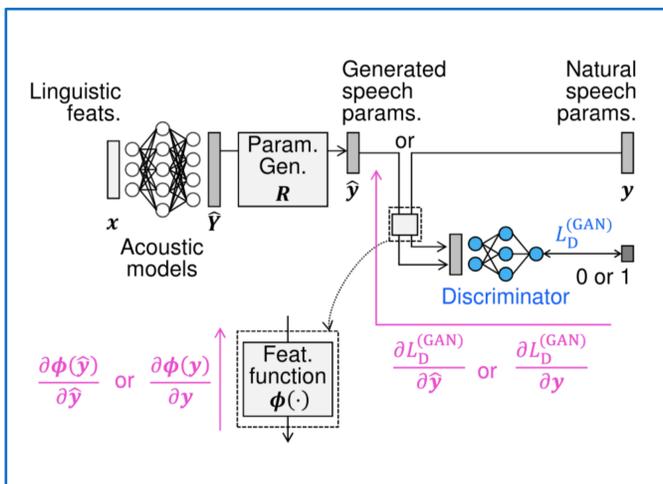
例:  $p_1 F \hat{p}_2 F - p_3 F + p_4 F = p_5 F @ p_6 F - p_7 F / A \dots$

#### 話速の比率を付与

$p / ./ + \begin{matrix} S:100 \end{matrix}$

音素ラベル    100を基準とした話速の比率

例:  $p_1 \hat{p}_2 - p_3 + p_4 = p_5 @ p_6 - p_7 / A \dots / S:100$



GANTTS<sup>[2]</sup>

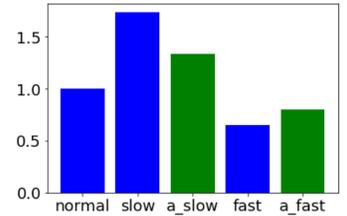
### 人間の自然音声による複数の話速を含むデータセットを構築

- 構築手順
  1. CMU ARCTIC<sup>[3]</sup>の単一話者の1132発話
    - オリジナル音声→Normal
    - 話速を**0.75倍**に変化させた発話→Slow
    - 話速を**1.25倍**に変化させた発話→Fast
  2. 女性1名,男性1名の被験者に聞かせる
  3. 3種類の話速で自然音声を収録 (44100Hz,16bit)

## 4. 実験

### データセットの分析

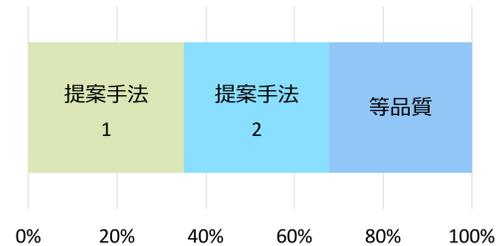
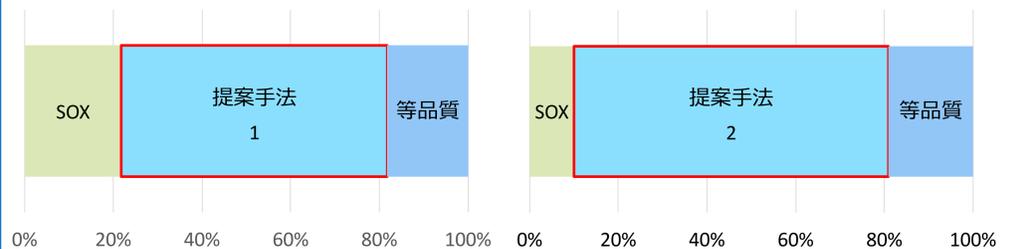
- 自然音声は与えた音声の話速より大きく話速が変化した
- 話速の変化によって母音と子音の平均継続長の変化傾向に違いがない
- 話速の変化によりパワーの変化は起こらない



女声全体発話(3396発話)の平均継続長

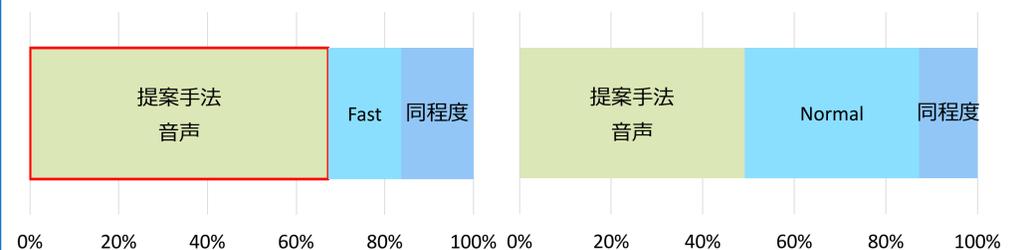
### 音声の自然性に関する主観評価

- 提案手法と通常の合成音声波形を後処理した音声の自然性を比較
- 文全体の話速を操作した音声では、提案手法音声**がより自然**
- 提案手法間での自然性の差は無い



### 話速変化と強調の関係性を調査

- 強調の有無に対し評価
  - 特定フレーズをSlow,他をFast話速で生成した音声
  - 文全体をFast話速で生成した音声
  - 文全体をNormal話速で生成した音声
- Fast話速との比較では強調されている
- 提案手法は正しく**音素単位で話速を制御できている**
- Normal話速とは差がない
  - Normal発話が一般的な話速より遅い



## 5. まとめ

- 発話文内で自由に話速の制御を行うことのできる,音素単位での話速制御を行うGANTTSについて提案
- 提案手法の有用性の検証のためにデータセットの構築および分析を行った
- 提案手法は合成音声の波形を人工的に操作するよりも自然で,かつ音素単位で適切な話速変化を行うことができることを示した

### 参考文献

[1] Park et al., Phonemic-level Duration Control Using Attention Alignment for Natural Speech Synthesis, ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

[2] Saito et al., "Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84-96, Jan. 2018.

[3] CMU ARCTIC, [http://festvox.org/cmu\\_arctic/](http://festvox.org/cmu_arctic/)