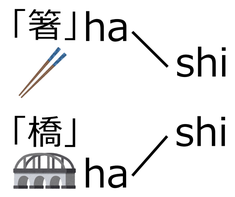
2-Q-36　人間からの誤り訂正フィードバックに基づく適応的なEnd-to-End 日本語音声合成に関する検討

藤井 一貴，齋藤 佑樹，猿渡 洋（東大院・情報理工）

合成音声のアクセント誤りを集合知を用いて修正する Human-In-The-Loop(HITL) フレームワーク

- 音声合成 (text to speech; TTS) [1]
  - テキストを入力し，対応する明瞭で自然な音声を合成する技術
  - パラメトリック音声合成[2]…テキストから音声を予測する問題を複数に分割して解く（パイプライン処理）
    - 自然言語処理など，専門知識が必要となる場面がある
- End-to-End (E2E) 音声合成 [3]
  - テキストからの音声予測を deep neural network; DNN の推論に置換
  - テキストから音声を予測する単一 DNN を学習
    =音声合成システム全体として最適化可能
    - 専門知識が必須ではなくなり，非専門家でも容易に高品質な音声を合成可能
- E2E TTS が抱える問題
  - 誤った音声を合成した時の修正が困難
    - DNN の積み重ねで表現される End-to-End TTS
      …モデルの解釈性低
      非専門家を含む音声合成のユーザが修正することは困難
  - 合成音声の誤りが引き起こす問題
    - 日本語をはじめとするピッチアクセント言語
      …アクセントの違いで言葉の意味を区別している
    - 合成音声のアクセントが誤ると言葉の意味が変化
      =適切なコミュニケーションをとることが困難となる



- E2E TTS に必要となる要素
  - 合成音声を自在に操作するための要素（操作性）
  - 合成音声の誤り訂正を容易にするための要素（適応性）

本発表の主題：合成音声の操作性と適応性を兼ね備えたTTSの提案と集合知を用いた誤り訂正HITLフレームワークの実験的評価

操作性の高いE2E TTS [4]（従来手法）

- 従来法の概要
  - TTSに音素と韻律を表す単一シーケンスを入力
  - 挿入された韻律記号により合成音声のアクセントを操作可能
- 従来法の韻律記号と実際のシーケンス

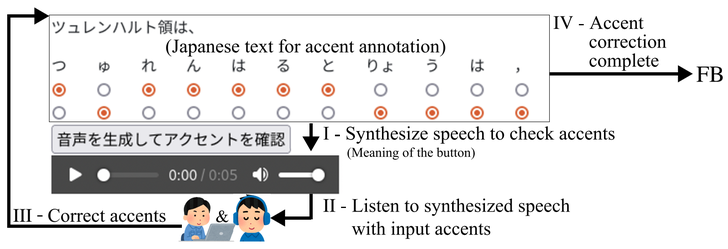
入力テキスト：続いて，ニュースをお伝えします.
↓ 書記素to音素変換
ts u z u i t e ny u u s u o o ts u t a e sh i m a s u
↓ テキスト解析より韻律記号を取得
ts u^z u i t e_ny u!u s u o#o ts u t a e sh i m a!s u(

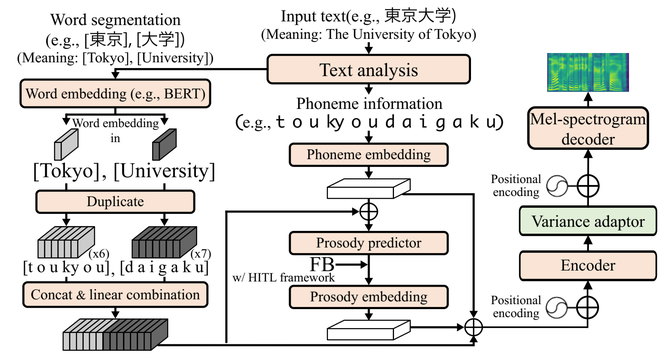
韻律特徴	韻律記号
アクセント立ち上がり	^
アクセント立ち下がり（アクセント核）	!
アクセント句境界	#
文末（非疑問形）	(
文末（疑問形）	?
ポーズ	_

- 問題点
  - 非専門家を含むユーザがアクセントの誤りを訂正するには複雑な記号もある
  - テキスト解析の結果が誤っていた場合は，非専門家を含むユーザが解析辞書を更新して誤りの再発防止を行う必要がある

操作性と適応性を兼ね備えたE2E TTS（提案手法）

- 提案法の概要
  - 入力テキストの文脈を考慮する為，BERT[5] 由来の単語埋め込みを導入
  - テキストからその韻律を予測する韻律予測器を新たに導入
  - アクセント訂正にクラウドソーシングによる集合知を活用するHITLフレームワークの提案





- 韻律予測器
  - 音素埋め込みと単語埋め込みより韻律記号を予測
  - ユーザは正しい韻律をフィードバックするだけで簡単に誤り訂正が可能
- HITLフレームワーク
  - 誤り訂正に人間を取り入れることで適応性が向上
  - クラウドソーシングを用いた文章へのアクセントアノテーション
    - 作業者の殆どは音声合成の非専門家＝誤り訂正の信頼性を高める必要有
      →入力されたアクセントで実際に音声を合成，聴覚的フィードバックにより信頼性確保

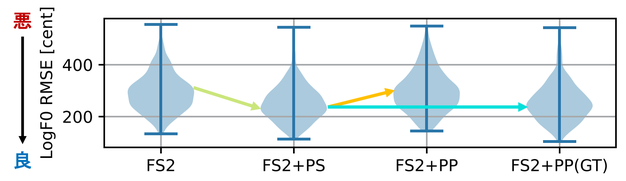
音声の自然性に関する実験的評価

- 実験条件

項目	値・設定
使用コーパス	JSUT コーパス[6]
使用 TTS	FastSpeech2 [7]
サンプリングレート	22,050 Hz
メルスペクトログラム次元	80
F0分析ツール	WORLD [8]
客観評価指標	logF0 Root Mean Squared Error (RMSE) [cent]
主観評価	MOSテスト（TTSモデル） Preference ABテスト（TTSモデル・HITLフレームワーク）

- TTSモデルに関する実験条件
  …韻律情報で条件付けすることの有効性と韻律予測器の性能調査
  - JSUTコーパスのBASIC5000サブセットで学習
    - 学習：4,488文　検証：512文
  - 比較手法
    - FS2 … 韻律情報で条件付けしない手法
    - FS2+PS … テキスト解析由来の韻律情報で条件付けした手法
    - FS2+PP … 提案手法（韻律予測器由来の韻律情報）
    - FS2+PP(GT) … 提案手法（テキスト解析由来の韻律情報）
- HITL フレームワークに関する実験条件
  …HITLフレームワークで得られるアクセントの品質調査
  - JSUTコーパスの VOICEACTRESS100 サブセットで評価
    - 固有名詞や造語などのアクセント推定が困難な文章を多く含む
      =誤り訂正の妥当性を示すのに十分なフィードバックを得ることが可能
    - 評価人数：15人 × 100文 = 1500人
  - 1文あたり15個のアクセント列の統合手法
    - Mode … 各アクセントごとのアクセント高低の最頻値
    - MACE … 期待値最大化法に基づく統合手法
      - 未知の回答とワーカーの能力を交互に推定し，低い能力のワーカーの影響を排除して統合
    - Best, Median, Worst … アクセント列ごとに音声を合成，正解音声との logF0 RMSE をとり，それぞれ
      - 値が最小のサンプル（Best），中央位置のサンプル（Median），最大のサンプル（Worst）

- 実験結果
  - TTSモデルに関する実験結果



- アクセントで条件付けすることの有効性が示唆
- 韻律予測器の予測精度の悪さが示唆
- 辞書由来のアクセントでスコアが改善
  =HITLにより正しいアクセントが得られたら自然な音声を合成できることが示唆

太字は２手法間に有意差有(p<0.05)を示す.

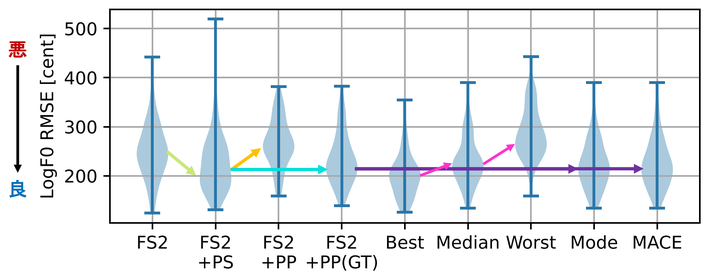
比較手法	Preference score
FS2 vs. FS2+PP	**0.552 vs. 0.448**
FS2 vs. FS2+PP(GT)	**0.268 vs. 0.732**
FS2+PS vs. FS2+PP	**0.768 vs. 0.232**
FS2+PS vs. FS2+PP(GT)	0.520 vs. 0.480
FS2 vs. FS2+PS	**0.248 vs. 0.752**

太字はFS2+PP(GT)と同等値(p<0.05)であることを示す.

手法	MOS ± 95% 信頼区間
JSUT	4.24 ± 0.139
FS2+PP	2.76 ± 0.143
FS2+PP(GT)	3.35 ± 0.163
FS2+PS	**3.45 ± 0.158**
FS2	2.71 ± 0.157

- 主観評価結果も，FS2+PPはFS2+PSよりも性能が低くFS2+PP(GT)はFS2+PSと同程度
  =韻律予測器の予測誤差が合成音声の品質を大きく劣化
  …しかし，正しい韻律があれば自然な韻律を持つ音声を合成することが可能

  - HITL フレームワークに関する実験結果



太字はFS2+PPより高い値であることを示す(p<0.05)

手法	MOS ± 95% 信頼区間
FS2+PP	2.87 ± 0.163
FS2+PP(GT)	**3.54 ± 0.156**
Best	**3.62 ± 0.151**
Median	**3.38 ± 0.156**
Worst	2.84 ± 0.174
Mode	**3.63 ± 0.159**
MACE	**3.49 ± 0.161**

- TTSモデルに関する実験結果と同様結果
- クラウドワーカー間の能力差が示唆
- 統合アクセントで合成した音声スコア＝辞書由来アクセントで合成した音声スコア
  …HITLは韻律予測器の誤りをうまく訂正可能
  =アクセント誤りを容易に訂正可能で品質向上に貢献
- 主観評価結果も同様の傾向を示唆

提案手法は韻律予測器の予測誤差をうまく補正し，合成音声の品質向上に貢献可能

- 今後の方針：フィードバックのインタフェースの検討，韻律列の統合法に関する更なる検討

参考文献

[1] Sagisaka, Proc. ICASSP, 1988.　[2] Zen et al., Speech Communication, 2009.　[3] Wang et al, Proc. INTERSPEECH, 2017.
[4] Kurihara et al., IEICE Transactions on Information and Systems, 2021.　[5] Devlin et al., Proc. NAACL-HLT, 2019.　[6] Takamichi et al., AST, 2020.　[7] Ren et al, Proc. ICLR, 2021.
[8] M. Morise et al., IEICE Transactions on Information and Systems, 2016.

本研究は，JST，ムーンショット型研究開発事業，JPMJMS2011 の支援を受けたものです.