音声認識結果を用いたEnd-to-End音声合成における話者適応の検討

D 井上勝喜（岡山大学　ヒューマンセントリック情報処理研究室）

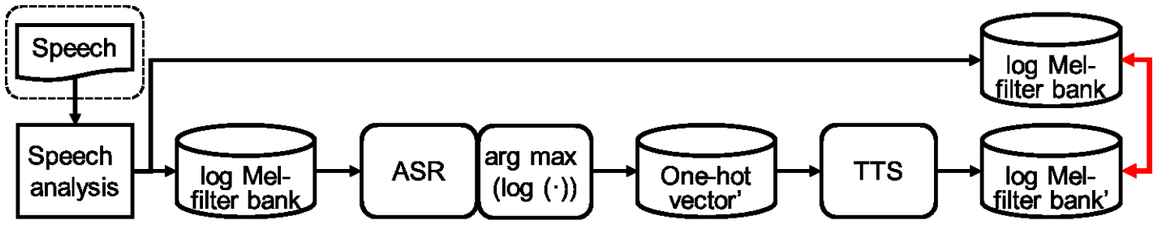
デモページ
はこちら→

1．研究背景

- End-to-End音声合成　→20時間以上の音声，テキストが必要
- Fine-tuningに基づく話者適応　→数十分のデータで構築
- 問題点：「テキストの書き起こしは面倒だが必要」
- 解決法：「End-to-End音声認識の認識結果を利用」
  - 特徴：合成器と類似した構造，専門知識が不要

2．提案手法

- Fine-tuningに基づく半教師あり話者適応



1. End-to-End音声合成，End-to-End音声認識の事前学習
2. **音声認識結果を入力とした音声合成のfine-tuning**

3．実験条件

＜比較する音声合成モデル＞　モデル構造：Transformer-TTS

1. **事前学習**モデル（単一話者，24時間）
2. Fine-tuningに基づく話者適応（**ペアデータ，20分**）
3. Fine-tuningに基づく話者適応（**音声＆認識結果，20分**）
4. Fine-tuningに基づく話者適応（**ペアデータ，10分**）
5. **特徴埋込**に基づく話者適応（複数話者，240時間）

＜データセット＞　言語：英語

LJ speech dataset (1), Test set of LibriTTS (2,3,4), Train set of LibriTTS (5)
Librispeech (End-to-End音声認識，不特定話者モデル)

4．実験結果

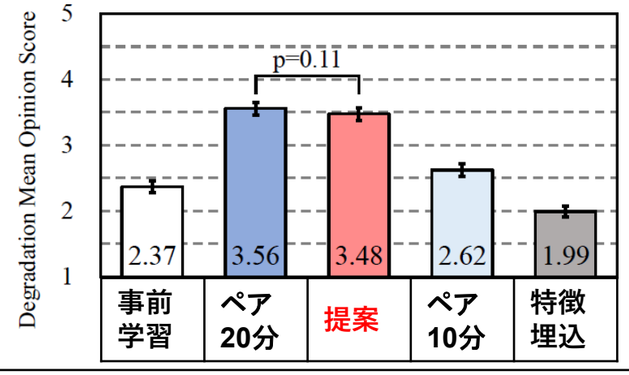
＜客観評価結果＞

話者	事前学習	ペア20分	提案	ペア10分	特徴埋込
女性 A	29.9	20.2	**19.7**	20.7	20.5
女性 B	29.5	**19.8**	21.0	35.5	20.4
女性 C	29.6	**23.1**	23.4	26.7	24.2
男性 A	29.0	15.5	**15.3**	16.3	15.4
男性 B	33.0	17.9	**17.5**	20.1	20.0
男性 C	32.2	18.9	19.9	20.0	**17.3**
全女性	29.7	**20.7**	21.2	29.0	21.4
全男性	31.1	**17.2**	17.3	18.5	17.4
累計	30.2	**19.6**	19.9	25.8	20.1

ケプストラムのRMS誤差
「目標と生成結果の距離」
評価文：6-18文

- 事前学習との比較
  提案手法の誤差が低下
  →目標話者の声に適応
- ペア，20分との比較
  同程度の性能
  →音声データの量は同程度であるため

＜主観評価結果＞



話者の類似性
評価文：10文×女性3名
参加者：23名

- 事前学習との比較
  提案手法が高評価
  →目標話者の声に適応
- ペア，20分との比較
  有意差なし

5．まとめ

音声認識結果を用いたEnd-to-End音声合成の話者適応を提案

＜今後の課題＞
音声認識の性能向上に伴う適応結果の性能評価