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

Study of speaker adaptation on end-to-end TTS with recognized text by ASR

井上　勝喜
Katsuki Inoue

岡山大学 阿部研究室
Abe Laboratory, Okayama University

概要 本研究では，目標話者の声を生成可能な音声合成器を低コストで構築することを目指す．具体的には，fine-tuning に基づく話者適応手法を採用し，テキストの書き起こし作業を音声認識器に代替させる．本報告では人手で書き起こされたテキストを用いた理想条件と比較することで，音声認識結果を用いた話者適応方式の性能を評価した．客観評価および主観評価実験より，提案手法は理想条件と同等の性能を示した．

1　はじめに

テキスト音声合成（Text-to-speech synthesis: TTS）は入力テキストから音声を生成する技術である．近年，End-to-End TTS [1, 2, 3, 4] が盛んに研究されている．End-to-End TTS では，文字または音素系列から音響特徴量系列へのマッピングをエンコーダデコーダに代表されるニューラルネットワークで学習する．

End-to-End TTS システムの構築には，音声とテキストの大量のペアデータが学習に必要である．具体的には，数時間から 20 時間に及ぶ単一話者の音声コーパス [5] が必要となる．このため，様々な声質の End-to-End TTS システムの構築には高いコストが必要となる．これを解決するために，数十分の音声データを用いた fine-tuning に基づく話者適応が End-to-End TTS において提案されている [6, 7]．しかし，収録音声に対応したテキストデータを用意する必要があり，これが fine-tuning に基づく話者適応のボトルネックとなっている．

本稿では，少量の非ペア音声データを用いた話者適応を提案する．TTS に必要なテキストデータを生成するために，End-to-End automatic speech recognition（ASR）システム [8, 9] を使用する．まず，十分な量のテキストと音声のペアデータを用いて End-to-End ASR と End-to-End TTS を事前学習する．その後，事前学習された ASR モデルから目標話者の音声データから発話内容のテキストを生成する．最後に，生成テキストと目標話者の音声データを用いて事前学習された TTS モデルを fine-tuning する．学習速度が速く，高い性能を示すことができる [3, 9] ため，ASR と TTS のモデル構造として Transformer [10] を用いる．ASR と TTS に End-to-End 型のモデル構造を利用することで，単純なパイプラインで話者適応を実現することができる．実験結果より，客観的ケプストラム距離と主観的話者類似度の観点において，非ペア音声データを用いた提案方式はペア音声データを用いた話者適応と同等の性能を示した．

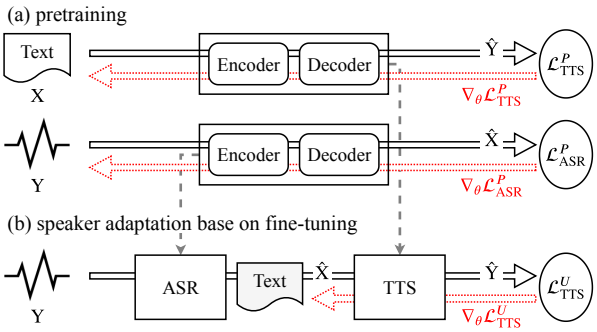

図 1: 非ペアの音声データを用いた話者適応の概要

2　学習済み End-to-End モデルを用いた半教師あり話者適応

図 1 は非ペアの音声データを用いた fine-tuning に基づく話者適応の概要である．提案方式は End-to-End TTS と End-to-End ASR からなる．事前学習済み ASR モデルから生成されたテキストと ASR モデルへの入力音声を用いて，事前学習済み TTS モデルを目標話者の音声データに適応させる．

2.1　ペアデータを用いた音声合成と音声認識の事前学習

End-to-End TTS モデル（図 1 (a) 上部）は C 次元の文字 ID 系列 $\mathbf{X} = \{x_c\}_{c=1}^{C}$ から，T 次元の音響特徴量系列 $\mathbf{Y} = \{\mathbf{y}_t\}_{t=1}^{T}$ へのマッピングを学習する．音響特徴量として，log Mel-filterbank 特徴量を用いる．モデル構造は attention 付きエンコーダ・デコーダである．入力特徴量系列 $\mathbf{X}$ は，エンコーダを通して中間層の特徴量系列 $\mathbf{H}^{\text{tts}} = \{\mathbf{h}_c^{\text{tts}}\}_{c=1}^{C}$ へエンコードされる．エンコーダとして，self-attention mechanism [10] を用いる．中間層の特徴量系列 $\mathbf{H}^{\text{tts}}$ は source-target attention mechanism [12] を用いて，デコーダへ入力される．その後，音響特徴量系列 $\hat{\mathbf{Y}} = \{\hat{\mathbf{y}}_t\}_{t=1}^{T}$ と文の最終トークンとなる確率の系列 $\hat{\mathbf{s}} = \{\hat{s}_t\}_{t=1}^{T}$ がデコーダにより生成される．

ペアデータのデータセット $\mathcal{D}^P$ から，音響特徴系列 $\mathbf{Y}$ と対応する文字列 $\mathbf{X}$ が与えられた時，以下の TTS に関する損失を最小化するように，TTS モデルを最適化する．

$$\mathcal{L}^P_{\text{TTS}}(\mathbf{X}, \mathbf{Y}) = \frac{1}{T}\sum_{t=1}^{T} \text{L1}(\hat{\mathbf{y}}_t, \mathbf{y}_t) + \frac{1}{T}\sum_{t=1}^{T} \text{BCE}(\hat{s}_t, s_t) \tag{1}$$

表 1: モデルの構成（Zero-shot は学習データに含まれない目標話者の単一発話から得られた x-vector を用いた話者適応を示す.）

モデル	学習データ	事前学習モデル	テキスト	適応の種類	目標話者
PT	the LJ speech dataset [5]	-	ground-truth	-	学習しない
PROPOSED	**Unpair** (表 2)	PT	**end-to-end ASR**	fine-tuning	学習する
AD-FT-P	**Pair** (表 2)	PT	ground-truth	fine-tuning	学習する
AD-FT-p	$\mathbf{Pair}_{\mathbf{half}}$ (表 2)	PT	ground-truth	fine-tuning	学習する
AD-EM	LibriTTS [11]	-	ground-truth	feature embedding	学習しない (zero-shot)

ここで，L1(·) は L1 ノルム，BCE(·) は binary cross entropy，s_t は時間 t における音響特徴量が音声の終端であること ($s_t = 1$)，または終端でない ($s_t = 0$) ことを示すラベルである．また，$\hat{\mathbf{y}}_t$ は正解データ（ground-truth）の過去の特徴量系列 $\{\mathbf{y}_{t'}\}_{t'=1}^{t-1}$ により条件付けられる．つまり，teacher-forcing を用いて TTS モデルを学習する．

End-to-End ASR モデル [13, 8] は TTS モデルと同様な構造である．TTS モデルとは対照的に，ASR モデルは音響特徴系列 $\mathbf{Y}$ から文字 ID 系列 $\mathbf{X}$ へのマッピングを学習する．ASR のデコーダは事後確率（posteriror）の系列 $\mathbf{P} = \{p_c\}_{c=1}^{C}$ を出力する．ここで，$p_c = p(x_c|x_1, x_2, ..., x_{c-1}, \mathbf{Y}) = p(x_c|x_{1:c-1}, \mathbf{Y})$ である．音声とテキストのペアのデータセット $\mathcal{D}^P$ が与えられた時，以下の ASR に関する損失を最小化するように，ASR モデルを最適化する．

$$\mathcal{L}_{\mathrm{ASR}}^{P}(\mathbf{Y}, \mathbf{X}) = -\sum_{c=1}^{C} \log p(x_c|x_{1:c-1}, \mathbf{Y}) \quad (2)$$

事後確率 $p(x_c|x_{1:c-1}, \mathbf{Y})$ は文字 ID 系列の正解データに条件付けられ，teacher forcing を用いて ASR モデルを学習する．

2.2 非ペア音声データと音声認識を用いた音声合成の fine-tuning

図 1 (b) における ASR モデルとして，2.1 節で示した学習済み ASR モデルを用いる．非ペア音声のデータセット $\mathcal{D}^U$ から音響特徴量系列 $\mathbf{Y}$ が与えられた時，ASR モデルは長さ C' の文字 ID 系列 $\hat{\mathbf{X}} = \{\hat{x}_c\}_{c=1}^{C'}$ を生成する．

$$\hat{\mathbf{X}} = \underset{\mathbf{X} \in \mathcal{U}^+}{\operatorname{argmax}} \log p(\mathbf{X}|\mathbf{Y}) \quad (3)$$

ここで，$\mathcal{U}^+$ は有限の文字トークンで表現された文章の集合である．その後，TTS モデルは生成された文字 ID 系列 $\hat{\mathbf{X}}$ を用いて，以下の式を最小化することで，新しい話者に適応するように fine-tuning する．

$$\mathcal{L}_{\mathrm{TTS-ASR}}^{U}(\mathbf{Y}) \approx \mathcal{L}_{\mathrm{TTS}}^{P}(\hat{\mathbf{X}}, \mathbf{Y}) \triangleq \mathcal{L}_{\mathrm{TTS}}^{U}(\mathbf{Y}) \quad (4)$$

ここで，$\hat{\mathbf{X}}$ には正解文字列との間の予測誤差が含まれる．

3 実験条件

提案方式の有効性を評価するために，5 種類の End-to-End TTS モデルを用いる．一つ目のモデル (PT) は事前学習モデルであり，話者適応のベースラインとして用いる．二つ目のモデル (PROPOSED) は非ペア音声データを用いた fine-tuning に基づく適応モデルであり，提案方式である．三つ目のモデル (AD-FT-P) は音声とテキストのペアデータを用いた fine-tuning に基づく適応モデルであり，話者適応の上限として用いる．四つ目のモデル (AD-FT-p) は半分のサイズのペアデータを用いた fine-tuning に基づく適応モデルであり，適応データ量の比較に用いる．五つ目のモデル (AD-EM) はペアデータと x-vector [14] を用いた feature embedding に基づく話者適応のモデル [15, 16] であり，話者適応の手法の比較に用いる．

3.1 データセット

ASR の事前学習モデルの構築のために，LibriSpeech [17] を用いる．これは 2,484 名の英語話者からなる約 1,000 時間のコーパスである．TTS の事前学習モデル（PT）の構築のために，the LJ speech dataset [5] を用いる．これは 1 名の女性話者の約 24 時間の英語音声からなる．

Fine-tuning に基づく適応モデル (AD-FT-P, AD-FT-p, PROPOSED) の構築のために，LibriTTS [11] のサブセットを作成した．これは，低ノイズの評価セットに含まれる 3 名の男性話者 ($\mathrm{male_A}$, $\mathrm{male_B}$, $\mathrm{male_C}$) と 3 名の女性話者 ($\mathrm{female_A}$, $\mathrm{female_B}$, $\mathrm{female_C}$) からなる．LibriTTS は，2,456 名の英語話者が発話した約 585 時間の音声コーパスである．表 2 に各話者のサブセットの詳細を示す．**Pair** はテキストと音声のペアデータからなり，AD-FT-P の学習に用いる．$\mathbf{Pair}_{\mathbf{half}}$ は **Pair** の半分のサイズのペアデータからなり，AD-FT-p の学習に用いる．**Unpair** は **Pair** の音声データのみからなり，PROPOSED の学習に用いる．ASR からテキストが生成されなかった音声は **Unpair** には含まれない．学習セットの文字誤り率（character error rate: CER）は最大で 3.3% である．

Feature embedding に基づく適応モデル (AD-EM) の構築のために，LibriTTS [11] を用いる．学習セットは 2 種類の低ノイズのサブセット（50 時間，190 時間），開発セットは低ノイズのサブセット（9 時間）である．

3.2 特徴量とテキストの表現

ASR モデルの音響特徴量として，80 次元の log Mel-filterbank 特徴量と 3 次元のピッチ特徴量を用いた．テキストとして，byte pair encoding (BPE) [18] により表現された単語トークンを用いた．これらのトークンに句読点は含まれない．

単一話者 TTS モデル（PT, PROPOSED, AD-FT-P, AD-FT-p）の教師データとして，22.05 kHz を超える音声データは 22.05 kHz へとダウンサンプリングし

表 2: Fine-tuning に用いるデータセットの構成（Utt.，Dur.，Char.，Del.，Ins.，Sub.，CER はそれぞれ総発話数，総発話長 [minute]，総文字数，削除誤り（deletion error） [%]，挿入誤り（insertion error） [%]，置換誤り（subtraction error） [%]，文字誤り率（character error rate） [%] を示す.）

データセット			学習データ											開発データ
			Pair		$\text{Pair}_{\textbf{half}}$		Unpair							P/U
話者	ID	性別	Utt.	Dur.	Utt.	Dur.	Utt.	Dur.	Char.	Del.	Ins.	Sub.	CER	Utt.
Male_{A}	1089	男性	169	20.08	90	10.05	167	20.06	14786	0.3	0.6	0.3	**1.2**	10
Male_{B}	2300	男性	126	21.71	59	10.15	126	21.71	16543	0.4	0.4	1.0	**1.8**	7
Male_{C}	8230	男性	108	17.19	64	10.04	108	17.19	12181	0.8	0.1	2.4	**3.3**	6
Female_{A}	237	女性	249	21.07	94	10.08	249	21.07	16685	0.5	0.6	0.8	**1.9**	15
Female_{B}	4446	女性	319	20.32	128	10.05	315	20.28	18387	0.4	1.2	0.6	**2.2**	18
Female_{C}	5683	女性	181	20.09	86	10.08	180	20.08	14641	0.6	0.7	1.0	**2.4**	10

た．音響特徴量として，80 次元の log Mel-filterbank 特徴量を用いた．テキストは 33 種類の文字セットへとトークン化した．文字セットは 26 種類のアルファベット（A-Z），5 種類の句読点（',.!?），2 種類の特殊タグ（未知語 ⟨unk⟩，空白 ⟨space⟩）からなる．非ペア音声データの話者適応（PROPOSED）のために，テキストは ASR から生成し，文末にピリオド（.）を追加した．

複数話者 TTS モデル（AD-EM）の教師データとして，24 kHz の音声データを使用した．音響特徴量として，80 次元の log Mel-filterbank 特徴量を用いた．テキストは 76 種類の文字セットへとトークン化した．話者性を制御する特徴量として，x-vector [14] を feature embedding に用いた．

3.3 モデルの条件

ASR モデルとして，Transformer のモデル構造 [10] を用いた．ASR のエンコーダは各層 2048 ユニット，12 層からなり，ASR のデコーダは各層 2048 ユニット，6 層からなる．ASR モデルは学習率のシード 10.0 の Noam [10] を用いて，120 エポックを学習した．デコード時において，ビーム幅 10 のビームサーチを用いた．

TTS モデルとして，Transformer のモデル構造を用いた．TTS のエンコーダは各層 1536 ユニット，6 層からなり，ASR のデコーダは各層 1536 ユニット，6 層からなる．事前学習モデル PT は学習率のシード 1.0 の Noam で，1000 エポックを学習した．Fine-tuning に基づく適応モデル（PROPOSED，AD-FT-P，AD-FT-p）は学習率のシード 0.1 の Noam で，100 エポックを学習した．また，feature embedding に基づく適応モデル（AD-EM）は学習率のシード 1.0 の Noam で，100 エポックを学習した．

波形生成において，80 次元の log Mel-filterbank 特徴量で条件付けられた WaveNet [19] をボコーダ [20] として用いた．Fine-tuning に基づく適応モデルのための WaveNet は，LJ speech dataset [5] を用いて学習した．また，feature embedding に基づく適応モデルのための WaveNet は，LibriTTS [11] を用いて学習した．

表 3: ケプストラムの RMS 誤差 [dB] に関する客観評価結果

話者	PT	AD-FT-P	PROPOSED	AD-FT-p	AD-EM
Male_{A}	29.0	15.5	**15.3**	16.3	15.4
Male_{B}	33.0	17.9	**17.5**	20.1	20.0
Male_{C}	32.2	18.9	19.9	20.0	**17.3**
Female_{A}	29.9	20.2	**19.7**	20.7	20.5
Female_{B}	29.5	**19.8**	21.0	35.5	20.4
Female_{C}	29.6	**23.1**	23.4	26.7	24.2
Male	31.1	**17.2**	17.3	18.5	17.4
Female	29.7	**20.7**	21.2	29.0	21.4
Total	30.2	**19.6**	19.9	25.8	20.1

4 評価実験

評価セットとして，学習に用いていない音声データを用いた．各評価セットは表 2 の開発セットと同一の発話数である．全ての TTS モデルは正解テキストを用いて音声を生成した．AD-EM において，各話者の x-vector は開発セット中の 1 発話からそれぞれ抽出した．

4.1 客観評価

提案方式を評価するために，音響特徴量の予測性能について評価した．客観評価指標として，ケプストラムの root mean squared（RMS）誤差を用いた．RMS 誤差は正解データと生成データの誤差を最小化するように動的時間伸縮（dynamic time warping : DTW）を用いて算出した．ケプストラムは離散コサイン変換を用いて，log Mel-filterbank 特徴量から算出した．客観評価結果は話者ごと，性別ごと，全体の平均値をそれぞれ算出した．

表 3 はケプストラムの RMS 誤差を示している．非ペア音声データの適応モデル（PROPOSED）は事前学習モデル（PT）を上回っている．さらに，PROPOSED はペアデータの適応モデル（AD-FT-P）と同様な性能を示している．これらの結果より，非ペア音声データを用いた提案適応方式はペアデータを用いた適応と同

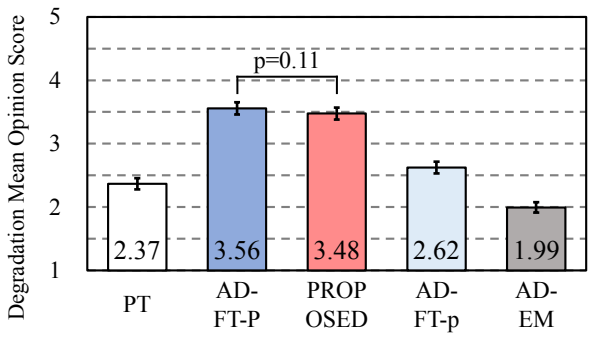

図 2: 話者類似性に関する DMOS の主観評価結果と 95% 信頼区間（$p = 0.11$ は 2 モデル間の two-tailed t-test における p 値である．）

等な話者類似性を示すことが分かる．非ペア音声データとペアデータの両方において，話者性を学習するための音声データの量は同等であり，この結果は妥当である．また，PROPOSED は半量のペアデータを用いた適応モデル（AD-FT-p）を上回る性能を示している．この結果は fine-tuning によって目標話者に適応するためには，音声データの量が重要であることを示している．加えて，feature-embedding モデル（AD-EM）は AD-FT-P や PROPOSED と同様な性能を示している．ケプストラムの評価尺度では，fine-tuning による適応と feature-embedding による適応は同程度の性能である．

4.2 主観評価

提案方式の有効性を評価するために，話者の類似性に関する 5 段階の degradation mean opinion score（DMOS）テストを用いた．合成音声として，30 発話（3 名の女性話者の各 10 発話）を評価した．また，リファレンスとして，目標話者の収録音声を使用した．評価音声は静かなオフィスまたは防音室で 23 名の聴取者によって評価した．評価音声の順番は無作為であり，全ての聴取者で同一である．

図 2 に話者類似性に関する DMOS の結果を示す．非ペア音声データを用いた fine-tuning モデル（PROPOSED）は事前学習モデル（PT）を上回る性能を示している．さらに，非ペア音声データの条件（PROPOSED）とペアデータの条件（AD-FT-P）の間に有意な差は見られない．これらの結果は ASR を用いることで非ペア音声データでの話者適応が実現されていることを示している．加えて，PROPOSED は半量のペアデータを用いた fine-tuning モデル（AD-FT-p）より高い性能である．この結果より，話者性を学習するには音声データの量が重要であることが分かる．また，PROPOSED は feature-embedding モデル（AD-EM）よりも高く評価されている．この結果は，feature-embedding では未知の話者を充分に表現することができない可能性を示唆している．

5 まとめ

本論文では，非ペア音声データを用いた End-to-End TTS のための話者適応方式を提案した．まず，End-to-End ASR モデルが目標話者の音声データからテキストを生成する．その後，End-to-End TTS モデルは生成テキストから目標音声へのマッピングを学習するように fine-tuning する．客観評価結果より，ケプストラムの RMS 誤差において非ペア音声データを用いた fine-tuning モデル（PROPOSED）はペアデータを用いた fine-tuning モデル（AD-FT-P）と同等な性能を示した．主観評価より，話者類似性において PROPOSED と AD-FT-P は同等な性能を示した．これらの結果より，提案方式は非ペア音声データと ASR モデルを用いることで，事前学習された TTS モデルを目標話者の声に適応することが可能であることが示された．

今後の課題として，ASR と TTS の同時最適化の枠組みにおける話者適応についても検討する予定である．

参考文献

[1] Y. Wang *et al.*, *arXiv preprint arXiv:1703.10135*, 2017.

[2] J. Shen *et al.*, in *Proc. of ICASSP*, 2018, pp. 4779–4783.

[3] N. Li *et al.*, *arXiv preprint arXiv:1809.08895*, 2018.

[4] Y. Ren *et al.*, *arXiv preprint arXiv:1905.09263*, 2019.

[5] K. Ito, "The LJ speech dataset," `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[6] S. Arik *et al.*, in *Proc. of NIPS*, 2018, pp. 10019–10029.

[7] Y.-J. Chen *et al.*, in *Proc. of Interspeech*, 2019, pp. 2075–2079.

[8] D. Bahdanau *et al.*, in *Proc. of ICASSP*, 2016, pp. 4945–4949.

[9] S. Karita *et al.*, *arXiv preprint arXiv:1909.06317*, 2019.

[10] A. Vaswani *et al.*, in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[11] H. Zen *et al.*, in *Proc. of Interspeech*, 2019.

[12] D. Bahdanau, K. Cho, Y. Bengio, *arXiv preprint arXiv:1409.0473*, 2014.

[13] A. Graves, N. Jaitly, in *Proc. of ICML*, 2014, pp. 1764–1772.

[14] D. Snyder *et al.*, in *Proc. of ICASSP*, 2018, pp. 5329–5333.

[15] A. Tjandra, S. Sakti, S. Nakamura, in *Proc. of Interspeech*, 2018, pp. 887–891.

[16] Y. Jia *et al.*, in *Proc. of NIPS*, 2018, pp. 4480–4490.

[17] V. Panayotov *et al.*, in *Proc. of ICASSP*, 2015, pp. 5206–5210.

[18] T. Kudo, J. Richardson, in *Proc. of EMNLP*, 2018, pp. 66–71.

[19] A. van den Oord *et al.*, *arXiv preprint arXiv:1609.03499*, 2016.

[20] A. Tamamori *et al.*, in *Proc. of Interspeech*, 2017, pp. 1118–1122.