少量な不明瞭音声を用いた適応による舌亜全摘出者の話者性を持つテキスト音声合成の検討

Text-to-speech Synthesis for Glossectomy Patients by Speaker Adaptation Using Small Amounts of Indistinct Speech

岡村 優頼
Masayori Okamura

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

概要 本報告では，舌亜全摘出者の話者性を再現可能な音声合成方式を提案する．提案方式では，健常者の音声と舌亜全摘出者の音声を混ぜた学習データで音声合成モデルを適応させることにより，明瞭かつ話者類似性の高い音声を合成する．評価実験の結果，健常者の明瞭音声のデータ数を 200 文から 50 文程度に減らすことに従い，話者類似性のみならず音韻明瞭度の評価も向上することが示された．

1 はじめに

舌亜全摘出者とは，手術などにより舌の半分以上を切除した患者のことである．舌が短いために舌亜全摘出者の発声は不明瞭であり，本人の発声による会話に支障がある．本研究の目的はこうした舌亜全摘出者をはじめとする構音障碍者を対象に，会話を通じたコミュニケーションを支援することにある．

構音障碍者の会話を支援する研究として，声質変換を利用して構音障碍者の音韻明瞭度を改善させる研究がある [1, 2]．構音障碍者の音声を健常者の音声に変換する声質変換モデルを学習することで，構音障碍者の不明瞭な音声を健常者の音韻明瞭度が高い音声に変換することができる．声質変換を用いる手法はキーボードなどによる入力を介する必要がなく，またリアルタイム性を高めることも可能である．一方でこの方式は，声質変換モデルの学習のために構音障碍者の音声を多く収録する必要があるため身体的負担が大きいこと，構音障碍者が健常者として発音したコーパスを収録できることを前提とする場合は利用要件が厳しいことなどの課題がある．また，テキスト音声合成を利用して構音障碍者の音声を合成する研究もある [3, 4]．この方式は健常者の音声で学習された複数話者テキスト音声合成モデルを構音障碍者に適応させることで，構音障碍者の話者性を再現した明瞭度の高い音声を合成することを目標としている．テキスト音声合成モデルを用いる手法は，健常者の音声で学習済みの音声合成モデルを利用することができるため，声質変換を用いる手法と比較して音韻明瞭度を改善しやすい．しかしこの方式は目標話者とは異なる健常者の音声で学習済みのモデルを目標話者に適応させる方式のため，話者性と明瞭性がトレードオフになる傾向がある．また話者類似性を向上させるには，声質変換を用いる手法と同様に多くの構音障碍者の音声データを用いる必要がある．

本研究では，舌亜全摘出者のデータが少ない条件でも明瞭かつ話者類似性の高い音声を合成することを目的として，自己教師あり学習モデルを話者埋め込みに用いたテキスト音声合成方式を用いる．この方式は学習済みの自己教師あり学習モデルを通して得られる音声特徴表現から話者埋め込み表現を作ることにより，少量の目標話者音声データから舌亜全摘出者の音声を合成することを目指す．また本報告では話者適応時に不明瞭な目標話者音声データと明瞭な健常者音声データを混ぜて学習することで，目標話者の話者性を持ちながら音韻明瞭度の高い音声を合成することを目指す．

2 疑似舌摘出音声によるパラレルコーパス

声質変換モデルやテキスト音声合成モデルの学習には舌亜全摘出者の音声データが必要である．また合成した音声の話者類似性を評価するためには，同一の舌亜全摘出者が舌を摘出する前に収録した健常者としての音声データが必要である．この舌亜全摘出者音声と健常者音声のパラレル音声データを実際の患者から収録するためには，舌切除前の患者に明瞭な健常者音声を収録してもらう必要がある．しかし患者への負担を考慮すると，このようなパラレル音声データを収録することは困難である．そこで本研究では先行研究 [5] で収録された，健常音声と疑似舌摘出音声のパラレルコーパスを用いる．疑似舌摘出音声とは，健常者が舌摘出者のような音声を発声できるように健常者の舌を固定する器具を用いて収録された音声である．このパラレルコーパスを用いることで，合成された音声の品質を健常者の音声と比較する評価を行うことが可能となる．

3 提案方式

3.1 自己教師あり学習モデルを用いた話者埋め込みによるテキスト音声合成

本研究で用いる音声合成アーキテクチャの概要を図 1 に示す．提案方式は End-to-End 音声合成アーキテクチャである VITS に対して話者埋め込みを用いた条件付けを行う方式をとる．話者埋め込み生成部には，藤田らによる自己教師あり学習モデルを用いた話者埋め込みへの変換手法 [6] を用いる．話者埋め込み生成部

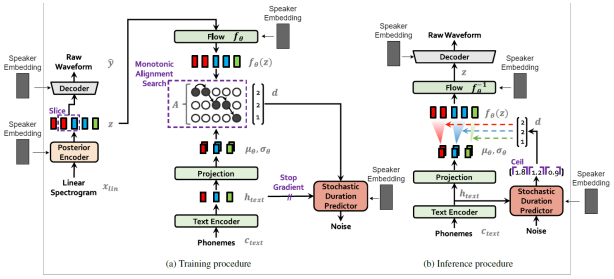

図 1: 提案方式のアーキテクチャ．図は論文 [7] の図をもとに作成.

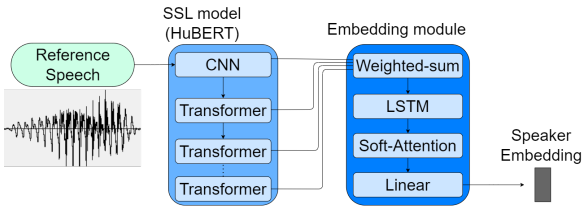

図 2: 話者埋め込み生成部.

のアーキテクチャを図 2 に示す．この手法は，自己教師あり学習モデルの出力として得られる音声特徴表現を LSTM+SoftAttention により固定長の話者埋め込みに変換する．本研究では自己教師あり学習モデルとして rinna 社が公開している事前学習済みの HuBERT-base [*1] を用いる．なお TTS モデルの学習時には HuBERT の重みは固定し，HuBERT 以降の層で重みの更新を行う．得られた話者埋め込みを用いて話者性の制御を行い，目標話者の音声を合成する．

3.2 話者適応における学習方針

音声合成モデルの学習フローを図 3 に示す．提案方式では，複数話者の音声で事前学習した音声合成モデルを不明瞭な目標話者の音声で話者適応させる．話者適応時に目標話者の不明瞭な音声のみを用いて学習する場合，話者類似性は高いが音韻明瞭度が低い音声が合成される．そこで提案方式では，話者適応時に不明瞭な目標話者音声と明瞭な健常者の音声データを一定の割合で混ぜた学習データを用いる．これにより，目標話者への話者類似性を高めながら，かつ明瞭な音声を合成する能力を保持することを期待する．また不明瞭な音声を学習することによる各音素の音韻明瞭度への悪影響を抑えるために，不明瞭な音声に対しては通常の音素ラベルとは異なる不明瞭音声用の音素ラベルを用いる．話者適応時には不明瞭な目標話者音声に対して不明瞭音声用の音素ラベルを用いて学習を行う．そして推論時には不明瞭な目標話者音声に対して通常の音素ラベルを用いて音声合成を行うことで，明瞭な音声を合成する．

[*1] https://huggingface.co/rinna/japanese-hubert-base

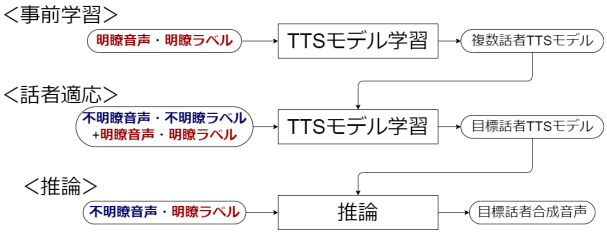

図 3: 音声合成モデルの学習フロー.

表 1: 使用するデータセット

	JSUT コーパス	*JVS* コーパス	研究室収録音声
データ数	7,396 文	13,000 文	各 53 文
発話時間	10 時間	26 時間	各 4 分
話者数	1 名	100 名	4 名
性別	女性	男性 49 名，女性 51 名	男性 3 名，女性 1 名
言語	日本語	日本語	日本語
発話内容	読み上げ	読み上げ	読み上げ

表 2: 話者適応の実験条件

条件	*JVS* コーパスのデータ量	話者数
目標話者のみ	0 文	0 名
目標話者 +2 文 100 名	各 2 文	100 名
目標話者 +1 文 100 名	各 1 文	100 名
目標話者 +1 文 50 名	各 1 文	男性 49 名 or 女性 51 名

4 評価実験

4.1 実験条件

4.1.1 使用するデータセット

使用するデータセットの詳細を表 1 に示す．音声合成モデルの事前学習には JSUT コーパス [8] と JVS コーパス [9] を用いる．話者適応時には JVS コーパスと，2 章で述べたパラレルコーパスのうち不明瞭な疑似舌摘出音声を入力音声および参照音声として用いる．音声のサンプリング周波数はそれぞれ VITS の入力は 22050Hz，参照音声の入力は HuBERT の事前学習時のサンプリング周波数に合わせて 16000Hz にリサンプリングする．

4.1.2 話者適応時の訓練データ

話者適応には，不明瞭な目標話者音声と JVS コーパスの一部を混ぜた訓練データを用いる．不明瞭な目標話者音声は約 4 秒× 10 文の合計約 40 秒の音声を訓練データとして用いる．JVS コーパスの訓練データの量および話者数は実験条件により変化させ，最適な学習データ量を探索する．実験条件を表 2 に示す．話者適応時の JVS コーパスのデータは事前学習段階で学習済みのデータを用い，parallel100 から各話者ごとに異なる発話内容の文を選択する．目標話者 +1 文 50 名の条件では，目標話者が男性の場合は男性 49 名，女性の場合は女性 51 名の音声を用いる．

表 3: 学習条件

条件	学習エポック数	学習率
事前学習 (JSUT)	1000	2.0×10^4
事前学習 (JVS)	500	2.0×10^4
話者適応 (目標話者のみ)	300	1.0×10^4
話者適応 (目標話者 +2 文 100 名)	1000	1.0×10^4
話者適応 (目標話者 +1 文 100 名)	1000	1.0×10^4
話者適応 (目標話者 +1 文 50 名)	600	1.0×10^4

4.1.3 学習条件

音声合成モデルの学習条件を表 3 に示す．事前学習時には JSUT コーパスと JVS コーパスを用いて学習を行う．話者適応時にはそれぞれ設計した訓練データを用いて学習を行う．学習エポック数は，各実験条件において訓練データ量が異なるため，損失が収束するまでのエポック数を参考に設定した．

4.2 評価方法

4.2.1 話者類似性の客観評価

話者類似性の客観評価には，話者埋め込みの Cosine 類似度を用いる．Cosine 類似度は以下の式で定義される．

$$CosineSimilarity(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{||\boldsymbol{a}||||\boldsymbol{b}||} \tag{1}$$

ここで $\boldsymbol{a}$，$\boldsymbol{b}$ はそれぞれ目標音声と合成音声から作成した話者埋め込みのベクトルである．Cosine 類似度は -1 から 1 までの値を取り，1 に近いほど 2 つのベクトルが類似していることを示す．客観評価のための話者埋め込み抽出には，SpeechBrain[*2]が公開している ECAPA-TDNN アーキテクチャ [10] を VoxCeleb データセット [11] で事前学習したモデルを用いる．

4.2.2 音韻明瞭度の客観評価

音韻明瞭度の客観評価には，音声認識器による文字誤り率 (Character Error Rate; CER) を用いる．文字誤り率は以下の式で定義される．

$$CER = \frac{S + D + I}{N} \tag{2}$$

ここで S は置換誤り，D は削除誤り，I は挿入誤り，N は正解文字列の総文字数である．音声認識器には OpenAI が公開している Whisper の large-v3 モデル[*3]を用いる．

4.3 実験結果

客観評価実験の結果を表 4 から表 7 に示す．

4.4 話者類似性の結果に関する考察

話者埋め込みの Cosine 類似度の結果を見ると，JVS コーパスのデータ量を減らし目標話者データの割合を大きくすることで Cosine 類似度が向上する傾向が見られる．どの目標話者に対しても，目標話者＋ 1 文 50 名の条件で最も Cosine 類似度が高くなっていることが分かる．一方，目標話者のみの条件では 4 話者中 2 話者で Cosine 類似度が最も低くなっている．これは目標話者データの割合を大きくするほど話者類似性が向上するという仮説に反する結果である．この結果は，今回客観評価に使用した話者埋め込みでは，話者性の違いよりも音韻明瞭度の違いが大きく影響している可能性を示唆している．従って話者類似性の評価としては，話者埋め込みの Cosine 類似度だけでなく，主観評価実験をあわせて行うことが望ましいと考えられる．

表 4: 実験結果 (女性 1)

条件	*Cosine* 類似度	*CER*
GT	-	0.090
目標話者のみ	0.699 ± 0.057	0.411
目標話者 +2 文 100 名	0.714 ± 0.051	0.210
目標話者 +1 文 100 名	0.768 ± 0.041	0.221
目標話者 +1 文 50 名	0.772 ± 0.040	0.143

表 5: 実験結果 (男性 1)

条件	*Cosine* 類似度	*CER*
GT	-	0.091
目標話者のみ	0.547 ± 0.072	0.554
目標話者 +2 文 100 名	0.527 ± 0.058	0.185
目標話者 +1 文 100 名	0.536 ± 0.063	0.157
目標話者 +1 文 50 名	0.558 ± 0.058	0.156

表 6: 実験結果 (男性 2)

条件	*Cosine* 類似度	*CER*
GT	-	0.086
目標話者のみ	0.525 ± 0.072	0.549
目標話者 +2 文 100 名	0.694 ± 0.056	0.129
目標話者 +1 文 100 名	0.692 ± 0.067	0.146
目標話者 +1 文 50 名	0.719 ± 0.058	0.119

表 7: 実験結果 (男性 3)

条件	*Cosine* 類似度	*CER*
GT	-	0.084
目標話者のみ	0.613 ± 0.060	0.471
目標話者 +2 文 100 名	0.564 ± 0.057	0.172
目標話者 +1 文 100 名	0.656 ± 0.053	0.136
目標話者 +1 文 50 名	0.687 ± 0.053	0.127

[*2] https://speechbrain.github.io/

[*3] https://github.com/openai/whisper

4.5 音韻明瞭度の結果に関する考察

文字誤り率の結果を見ると，どの目標話者に対しても目標話者＋ 1 文 50 名の条件で最も文字誤り率が低くなっていることが分かる．これは話者類似性の結果と一致しており，本実験条件の中では目標話者＋ 1 文 50 名の条件が最も話者類似性と音韻明瞭度の評価が高い条件であることが示された．また明瞭な JVS コーパスのデータ量を減らし，不明瞭な目標話者データの割合を大きくするにつれて文字誤り率が低下する傾向がみられる．これは各実験条件ごとの学習エポック数の影響が考えられる．各実験における学習エポック数は，音声合成モデルの損失が収束するまでのエポック数を目安に設定している．学習データが少ない条件では損失が収束するまでのエポック数が少なくなるため，目標音声の不明瞭さを十分に学習する前に目標話者の話者性を学習できていることが考えられる．この仮定をもとに考えると，目標話者のみの条件でも，目標話者の話者性をもちつつ音韻明瞭度の高い音声を合成することができる音声合成モデルが少ないエポック数で実現できている可能性が示唆される．

4.6 話者間の結果の違いによる影響

本実験では 4 名の話者に対して実験を行ったが，話者によって客観評価の結果に違いが見られる．話者埋め込みの Cosine 類似度の結果のうち目標話者 +1 文 50 名の条件で最も Cosine 類似度が高かった話者は女性 1 で 0.772 ± 0.040 であった．一方で最も Cosine 類似度が低かった話者は男性 1 で 0.558 ± 0.058 であり，両者には 0.214 の差がある．また文字誤り率の結果のうち目標話者 +1 文 50 名の条件で最も文字誤り率が低かった話者は男性 2 で 0.119 であった．一方で最も文字誤り率が高かった話者は男性 1 で 0.156 であり，両者には 0.037 の差がある．このように話者によって客観評価の結果に比較的大きな違いが見られることから，話者によって音声合成モデルの学習に必要なデータ量や学習エポック数が異なる可能性が示唆される．

5 まとめと今後の課題

本研究では舌亜全摘出者の音声を合成するための音声合成方式および話者適応のための学習方針を検討した．評価実験の結果，目標話者 +1 文 50 名の条件が本研究の実験条件の中では最もバランスの良い条件であることが示唆された．ただし話者類似性の評価においては話者埋め込みの Cosine 類似度だけでなく主観評価実験をあわせて行うことが望ましいと考えられる．また話者によって音声合成モデルの学習に必要なデータ量や学習エポック数が異なる可能性が示唆された．今後は主観評価実験によって音声合成モデルを評価するとともに，最適な学習データの設計や学習エポック数のさらなる調整を行うことが課題である．

参考文献

[1] K. Tanaka, S. Hara, M. Abe, M. Sato, and S. Minagi, “ Speaker Dependent Approach for Enhancing a Glossectomy Patient ’ s Speech via GMM-based Voice Conversion, ”Proc. INTERSPEECH, pp.3383–3388, Aug. 2017.

[2] H. Murakami, S. Hara, M. Abe, M. Sato, and S. Minagi, “ Naturalness Improvement Algorithm for Reconstructed Glossectomy Patient ’s Speech Using Spectral Differntial Modification in Voice Conversion, ” Proc. INTERSPEECH, pp.2464–2468, Sept. 2018.

[3] 吉本拓真, 高島遼一, 佐々木千穂, 滝口哲也, “ 音響モデルの話者適応に基づく脊髄性筋萎縮症者の音声明瞭化の検討, ” 日本音響学会 2021 年秋季研究発表会講演論文集, pp. 1053–1056, 2021.

[4] 吉本拓真, 松原圭亮, 高島遼一, 佐々木千穂, 滝口哲也, “ 複数話者 TTS を利用した脊髄性筋萎縮症者音声明瞭化の検討, ” 日本音響学会 2022 年春季研究発表会講演論文集, pp. 1045–1048, 2022.

[5] H. Murakami, S. Hara, and M. Abe, “ DNN-based Voice Conversion with Auxiliary Phonemic Information to Improve Intelligibility of Glossectomy Patients ’ Speech, ” Proc. APSIPA Annual Summit and Conference, pp.138—142, Nov. 2019.

[6] 藤田健一, 芦原孝典, 金川裕紀, 森谷崇史, 井島勇祐, “ 自己教師あり学習モデルを用いた zero-shot 音声合成の検討, ” 日本音響学会 2023 年春季研究発表会講演論文集, pp. 681–684, 2023.

[7] J. Kim, J. Kong, and J. Son, “ Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech, ” Proceedings of the 38th International Conference on Machine Learning, PMLR 139:5530–5540, 2021.

[8] R. Sonobe, S. Takamichi and H. Saruwatari, “ JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis, ” arXiv preprint, 1711.00354, 2017.

[9] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, “ JVS corpus: free Japanese multi-speaker voice corpus, ” arXiv preprint, 1908.06248, Aug. 2019.

[10] B. Desplanques, J. Thienpondt, and K. Demuynck, “ ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification, ” in Proceedings of Interspeech 2020, pp. 3830–3834, 2020.

[11] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, “ VoxCeleb: A Large-Scale Speaker Identification Dataset, ” Interspeech 2017, pp. 2616–2620, 2017.