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*

☆吉本拓真, 松原圭亮, 高島遼一 (神戸大),
△佐々木千穂 (熊本保健科学大), 滝口哲也 (神戸大)

1 はじめに

内閣府の調査 [1] によると，日本には身体障害者が 436 万人，知的障害者が 109.4 万人，精神障害者が 419.3 万人いるとされている．また，在宅の身体障害者の中では，聴覚・言語障害者は 34.1 万人いるとされている [2]．このような障害はコミュニケーションに対する大きな障壁となりやすく，バリアフリー社会の実現には円滑なコミュニケーションを行うための支援が不可欠である．

言語障害の一つに，ことばは理解しており話したいことは明確であるものの，声を作る器官やその動きに問題がありうまく発話できない構音障害というものがある．近年ではこのような構音障害者のコミュニケーションを支援するために，スマートフォンやタブレットを用いたテキスト音声合成（text-to-speech; TTS）アプリケーションが開発され使われるようになっている．しかし一般的な TTS アプリケーションによって作成される音声は，事前の学習時に用いられた健常者の声をもとに作成されるため，実際の使用者とは異なる声となってしまう．そこで，TTS のモデルを障害者本人の声だけで学習することも考えられるが，それを実現するためには多くの音声データの収録が必要で体への負担が大きくなってしまう上，作成された音声は元音声と同様に不明瞭なものとなる．

そこで我々の先行研究 [3, 4] では，明瞭性のある健常者 TTS モデルを障害者本人のものへ話者適応することで，本人の話者性を維持しつつ聞き取りやすい音声を合成するシステムを検討した．この手法によって明瞭性を保ちつつ話者性のある音声を生成できたが，話者性の面に関しては依然として改善の余地がみられた．そこで本研究では別のアプローチとして，話者埋め込みを利用した複数話者 TTS モデルから明瞭性のある障害者音声を生成することを検討する．さらに，その音声を従来手法の話者適応時に用いる方法についても検討する．

本研究では，構音障害を持つ脊髄性筋萎縮症（spinal muscular atrophy; SMA）者を対象とする．SMA は脊髄の運動神経細胞の病変によって起こる筋萎縮症であり，下位運動ニューロン病の一つとされる [5]．SMA は発症時期や最大獲得運動機能といった臨床的特徴により I 型から IV 型の 4 つに分類される．本研究では I・II 型に該当する重度の脊髄性筋萎縮症者を対象とする．重度の脊髄性筋萎縮症者には嚥下障害や呼吸不全なども見られ，人工呼吸が必要な場合も多い．そのため重度の脊髄性筋萎縮症者の音声は健常者とは異なる発話スタイルとなり，その音声を聞き慣れていない人からすると聞き取りづらいものとなる．音声の特徴としては，(i) 低周波成分と比べて高周波成分のパワーが弱いため子音が聞き取りづらい，(ii) フォルマントの変化があまり見られず母音の判別がしづらい，などが挙げられる．

2 複数話者音声合成

我々の従来研究では，明瞭性と話者性を両立した脊髄性筋萎縮症者音声を生成するために，単一話者 TTS を話者適応するアプローチを考えてきた [3, 4]．しかし，この手法では適応時に本人の音声をそのまま教師データとして使用しているため，明瞭性に影響を及ぼしてしまうという課題があった．そこで，本研究では健常者データのみで学習した複数話者 TTS システムを利用した zero-shot による合成手法について検討する．

2.1 話者埋め込み

単一話者 TTS は学習時に対象話者の音声とそれに対応するラベルの対を用いることで実現できるが，複数話者 TTS は複数人の音声とラベルを用意し，それに発話者の情報をベクトルとして表現した話者埋め込み（speaker embedding）を加えることで，複数の話者を一つのモデルで表現できるようにしたものである．話者埋め込みとして最も単純なものとして one-hot ベクトルが挙げられるが，これでは学習時に使用した話者しか表現することができない．近年ではほかの表現として，話者識別の分野で広く用いられている i-vector [6] や x-vector [7] が挙げられる．x-vector は話者識別モデルにおける中間層の出力をベクトルとして抽出したものであり，発話者を識別するために有用な特徴が含まれていると考えられる．この話者識別モデルを大量の話者で学習しておくことで，one-hot ベクトルでは表現できない未知話者に関しても話者性を表したベクトルを抽出することができる．本稿ではこの x-vector を話者埋め込みとして使用する．

2.2 複数話者 TTS を利用した脊髄性筋萎縮症者音声合成

複数話者 TTS は先述の話者埋め込みを変えることによって様々な話者の音声を生成することができる．本稿ではその点に着目し，健常者データで学習しておいた複数話者 TTS に，脊髄性筋萎縮症者音声から得た話者埋め込みを入れることでどのような音声が生成されるかを検証した．その結果，4.2.1 項でも触れるが，出力された音声は健常者 TTS から得たものであるから明瞭度については問題がないが，話者性については脊髄性筋萎縮症者に近づいてはいるものの不十分であり，また音声によってばらつきが多かった．

*Speech intelligibility for people with spinal muscular atrophy using multi-speaker TTS. by YOSHIMOTO, Takuma, MATSUBARA, Keisuke, TAKASHIMA, Ryoichi (Kobe Univ.), SASAKI, Chiho (Kumamoto Health Science Univ.), TAKIGUCHI, Tetsuya (Kobe Univ.)

3　2段階話者適応

前章によって，健常者複数話者 TTS に脊髄性筋萎縮症者音声から得た x-vector を入れて音声を生成したが，その音声は話者性の面では不十分であった．しかしながら明瞭性については健常者と同等であるから，この合成音声は脊髄性筋萎縮症者により近い健常者音声と考えることができる．そこで，我々が先行研究で行っていた話者適応のアプローチ [4] に今回生成した合成音声を用いることを検討する．

本稿では適応を 2 段階行うことを考え，その流れを Fig. 1 に示す．このシステムでは（単一話者）TTS モデルを，音素レベルの言語特徴量から音素の長さを推定する継続長モデル，フレームレベルの言語特徴量から音響特徴量を出力する音響モデルの 2 つに分けて考え，音響特徴量から音素を推定する音声認識（automatic speech recognition; ASR）モデルと合わせて計 3 つのモデルから構成される．学習では，はじめに健常者音声とそのラベルを使用して 3 つのモデルをそれぞれ学習する．これにより得られた健常者 TTS モデル，ASR モデルはそれぞれ明瞭な音声を合成，認識できる．次に，前章で得られた合成音声を用いて音響モデルの適応を行う．ここで用いる合成音声は聞き取りやすく，かつはじめに用いた健常者音声よりも脊髄性筋萎縮症者音声に近いため，適応によって明瞭性を落とすことなく話者性を向上させた音声を生成できるようになる．最後に，脊髄性筋萎縮症者の収録音声を用いてさらに音響モデルの適応を行う．この際，従来法と同様に音響モデルの出力をはじめに学習しておいた ASR モデルに通して得られる音素と正解ラベルとの損失も考慮して，明瞭性の低下を抑えるようにする．これを式に表すと次の通りである．

$$L = L_{tts} + \alpha \times L_{asr} \tag{1}$$

ただし，L は全体の損失，L_{tts} は音響モデルの出力と実際の音響特徴量との平均二乗誤差，L_{asr} は ASR モデルから出力された音素と実際のラベルとの交差エントロピーを表している．このような 2 段階適応を行うことで，最後の適応で話者性のギャップが小さくなり，脊髄性筋萎縮症者のモデルの適応がしやすくなることが期待される．

4　実験

4.1　実験条件

4.1.1　脊髄性筋萎縮症者音声

脊髄性筋萎縮症者の音声データとして，女性 2 名が ATR デジタル音声データベース（ATR コーパス）[8] に含まれる音素バランス単語 216 語を 1 単語あたり 5 回発話したものを使用している．ただし，一部収録の取りこぼしがあり，ノイズが多く含まれたデータの除去も行った．また，音声に対する音素セグメンテーション（音素とその開始・終了時間の対応付け）は手作業で行った．

4.1.2　複数話者 TTS

複数話者 TTS では，テキストからメルスペクトログラムに変換するモデル（text2mel）として Tacotron 2 [9] を使用した．また，メルスペクトログラムから波形を生成するボコーダ（mel2wav）として HiFi-GAN [10] を使用した．HiFi-GAN は高速に複数話者音声合成が可能なニューラルボコーダの一種である．なお，本実験はすべて ESPnet 2 [11, 12] 上で行った．

Tacotron 2 の学習には JVS コーパス [13] から，通常な発声の読み上げ文章 130 文（parallel100 + nonpara30）の音声を 97 名分使用した．なお，全ての音声はサンプリング周波数 24 kHz とした．また，HiFi-GAN および x-vector 抽出モデルの 2 つに関しては，それぞれ VCTK コーパス [14]，VoxCeleb コーパス [15, 16] で学習されて一般公開されている学習済みモデルを用いた．合成時に用いる話者埋め込みについては，脊髄性筋萎縮症者の収録音声を x-vector 抽出モデルに通して得られたベクトルをそのまま使用し，収録音声と同数の合成音声を生成した．なお，健常者音声での性能を検証するために，Tacotron 2 の学習に使用した JVS コーパスの音声，ATR コーパスに含まれる音素バランス単語 216 語についても x-vector を抽出して合成音声を生成した．

4.1.3　話者適応

単一話者 TTS の話者適応に用いる音響モデル，継続長モデル，ASR モデルはすべて双方向 LSTM（Bidirectional LSTM）[17] を用いている．初めの健常者モデルの作成については ATR コーパスに含まれる音素バランス文 503 文を女性 1 名分使用した．この際の学習率は 1e-3 とした．1 段階目の適応には，脊髄性筋萎縮症者の音声から抽出した x-vector を複数話者 TTS に通して得られた合成音声を使用し，その際に必要な音素セグメンテーションについては強制アライメントを行い獲得した情報をもとに著者が修正を加えて作成した．この適応時には学習率を 1e-4 とした．2 段階目の適応には，4.1.1 項で説明した音声を使用したが，単一話者 TTS であるため実験は話者ごとに行った．この適応時には学習率を 1e-5 とし，式 (1) における α の値は 0.5 とした．音響特徴量から波形を生成するボコーダには WORLD [18, 19] を用いた．なお，すべての音声はサンプリング周波数 16 kHz とした．

ラベルには Open JTalk [20] のフロントエンド部を利用して生成した 38 種類の音素（空白を含む）からなる HTS 形式のフルコンテキストラベルを使用した．言語特徴量の次元数は 975 次元（フレームレベルの場合はフレーム特徴量が追加されて 979 次元）とし，次元ごとに最小が 0，最大が 1 となるように min-max 正規化を行った [21]．音響特徴量は，メルケプストラム 60 次元，帯域非周期性指標，対数基本周波数，有声/無声フラグで構成され，有声/無声フラグ以外に関しては 2 次までの動的特徴量を含んだ計 187 次元とした．音響特徴量については次元ごとに平均 0 分散 1 となるよう正規化（標準化）を行った．

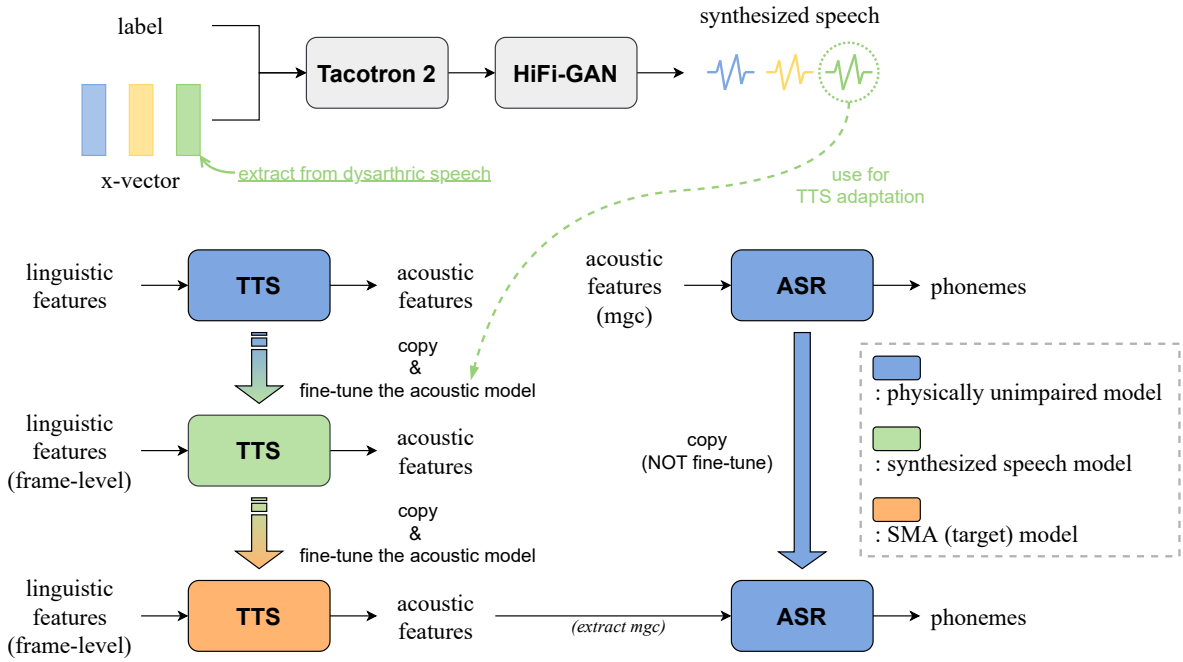

Fig. 1　Training procedure of the proposed TTS system

4.2　実験結果

4.2.1　複数話者 TTS の合成音声と x-vector

複数話者 TTS に脊髄性筋萎縮症者の音声から得た x-vector を入れると，x-vector が話者を切り替え本人らしさのある聞き取りやすい音声が合成されることが期待された．しかし，合成音声は明瞭性のある音声ではあるものの本人らしさは不十分な音声となった．また，音声によっても話者性にばらつきがあり，様々なピッチ・アクセントの音声が生成された．これは健常者音声である ATR コーパスの音声でも同様であり，未知話者に対しては話者性の表現が困難であること，単語の発話は発話長が短く x-vector が安定して出力されない，などが理由として考えられる．なお，Tacotron 2 の学習時に用いた JVS コーパスの音声ではうまく話者の切り替えが行われていたことから，既知話者の TTS は精度よく合成ができていたと考えられる．

512 次元からなる各話者の平均 x-vector を主成分分析（principal component analysis; PCA）によって 2 次元まで次元圧縮を行いプロットしたものを Fig. 2 に示す．図中の数字は JVS コーパスに含まれている話者を，F または M から始まるものは ATR コーパスに含まれている話者を，dys1 および dys2 はそれぞれ脊髄性筋萎縮症者を表している．JVS および ATR の話者の分布を広く見ると，PC1 軸のマイナス側に女性のデータが，プラス側に男性のデータが集中しており，脊髄性筋萎縮症者はそのどちらでもない場所に分布していた．また最後に s がついているものは，未知話者について本実験で生成した合成音声をさらに x-vector 抽出モデルに通して獲得した x-vector をプロットしたものである．話者性が元の音声と合成音声とで同様であれば，この 2 つは同じ場所にプロットされることが期待される．しかし，dys1 と dys1s, dys2 と dys2s はいずれも離れた場所にプロットされ

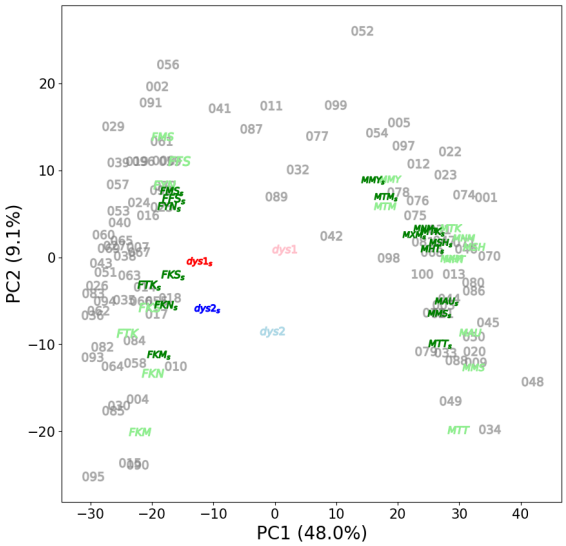

Fig. 2　Plot of the speaker-averaged x-vector

ていることが見て取れる．また，ATR の話者についても全体的に合成音声のほうがばらつきが小さくなっていることが確認される．このことから，未知話者については x-vector だけでは話者性を完全に調節することができず，脊髄性筋萎縮症者の合成音声については既知の健常者女性の音声に近づくようになっていることがわかる．

4.2.2　2 段階話者適応の合成音声

合成音声を用いて 2 段階話者適応を行い生成した dys1 の音声と，合成音声を用いず直接話者適応を行う従来法によって生成した dys1 の音声「勢い」のスペクトログラムを並べたものを Fig. 3 に示す．音声を比較すると，例えば点線で囲われた部分について，3 つ目の音素から 5 つ目の音素までの/i/，/o/，/i/の変化は第 2 フォルマントが大きく変化することで表現

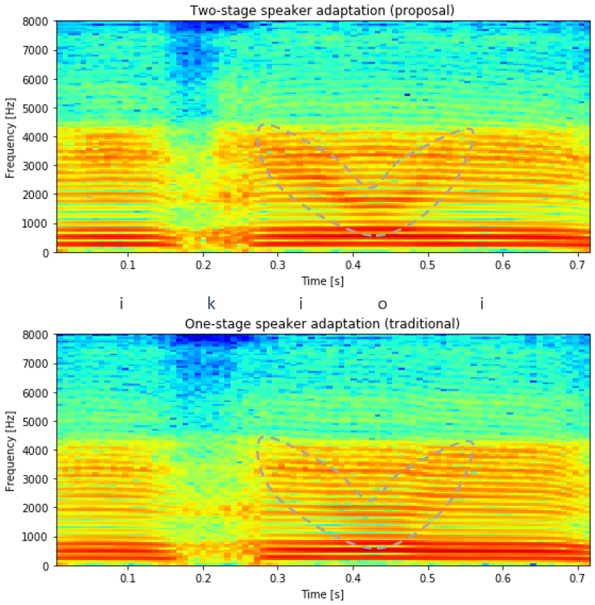

Fig. 3　Spectrogram of synthesized speech

Table 1　Mean f0 of synthesized speech

Method	original	two-stage (proposal)	one-stage (traditional)
F0 [Hz]	247.0	250.2	241.2

される [22] が，2 段階話者適応を行った上の方が U 字状により強くパワーが出ていることで確認できる．すなわち，2 段階話者適応を行った方がより明瞭性のある音声が生成できている．これは一度に行う話者適応の際のギャップが小さくなったことで，うまく適応が進められたと考えられる．また，Fig. 3 の音声について 0.2 秒以降の基本周波数（fundamental frequency; f0）の平均を示したものを Table 1 に示す．2 段階話者適応による合成音声のほうが収録音声の f0 に近い値となっており，話者性についても優れていることがわかる．

5　おわりに

本研究では，健常者データで学習した複数話者 TTS に脊髄性筋萎縮症者音声から得た x-vector を話者埋め込みとして用いた際の音声について検証を行った．健常者 TTS モデルを使用していることによる明瞭性と本人の話者から得た x-vector を用いていることによる話者性を兼ね備えた音声が生成されることが期待されたが，実際は未知話者であることが原因であくまでも本人の声に似た別の健常者の音声が生成されるにとどまった．これを活用し，次に従来の話者適応のアプローチに合成音声を用いることを考えた．2 段階話者適応のアプローチをとることで，従来法よりさらに明瞭性と話者性が両立した音声を生成できた．今後はより話者性を反映できる話者埋め込みを作成する方法や，話者性の評価方法について主観評価実験も行いながら検討する．

参考文献

[1] 内閣府, “令和 3 年版 障害者白書,” 2021.
[2] 厚生労働省, “平成 30 年版 厚生労働白書,” 2019.
[3] 吉本拓真 他, “モデル適応に基づく脊髄性筋萎縮症者の高明瞭度音声合成の検討,” 情報処理学会研究報告, 2021-SLP-137 (33), 1-5, 2021.
[4] 吉本拓真 他, “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,” 音講論 (秋), 1053-1056, 2021.
[5] SMA 診療マニュアル編集委員会, “脊髄性筋萎縮症診療マニュアル,” 金芳堂, 2014.
[6] N. Dehak *et al.*, “Front-End Factor Analysis for Speaker Verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, 19 (4), 788-798, 2011.
[7] D. Snyder *et al.*, “X-Vectors: Robust DNN Embeddings for Speaker Recognition,” *ICASSP*, 5329-5333, 2018.
[8] A. Kurematsu *et al.*, “ATR Japanese speech database as a tool of speech recognition and synthesis,” *Speech Communication*, 9 (4), 357-363, 1990.
[9] J. Shen *et al.*, “Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions,” *ICASSP*, 4779-4783, 2018.
[10] J. Kong *et al.*, “HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis,” *NeurIPS*, 17022–17033, 2020.
[11] T. Hayashi *et al.*, “Espnet-TTS: Unified, Reproducible, and Integratable Open Source End-to-End Text-to-Speech Toolkit,” *ICASSP*, 7654-7658, 2020.
[12] T. Hayashi *et al.*, “ESPnet2-TTS: Extending the Edge of TTS Research,” arXiv, 2021.
[13] S. Takamichi *et al.*, “JVS corpus: free Japanese multi-speaker voice corpus,” arXiv, 2019.
[14] J. Yamagishi *et al.*, “CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92),” The Centre for Speech Technology Research (CSTR), University of Edinburgh, 2019.
[15] A. Nagrani *et al.*, “VoxCeleb: A Large-Scale Speaker Identification Dataset,” *Proc. Interspeech*, 2616-2620, 2017.
[16] J.S. Chung *et al.*, “VoxCeleb2: Deep Speaker Recognition,” *Proc. Interspeech*, 1086-1090, 2018.
[17] M. Schuster, K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE transactions on Signal Processing*, 45 (11), 2673-2681, 1997.
[18] M. Morise *et al.*, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE transactions on information and systems*, E99-D (7), 1877-1884, 2016.
[19] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” *Speech Communication*, 84, 57-65, 2016.
[20] “Open JTalk,” `http://open-jtalk.sourceforge.net/`
[21] 南坂竜翔 他, “構音障害者の少量データを用いた深層学習による音声合成の検討,” 音講論 (秋), 1011-1014, 2019.
[22] T. Hirahara, R. Akahane − Yamada, “Acoustic Characteristics of Japanese Vowels,” *Proc. ICA*, 3287 − 3290, 2004.