脊髄性筋萎縮症者音声合成における明瞭性および話者性を考慮した適応手法の検討*

☆吉本拓真, 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

日本では，国民のおよそ 7.6%が何らかの障害を持っているといわれている [1]。特に身体障害者は 436 万人，なかでも在宅の聴覚・言語障害者は 34.1 万人いるとされている [2]。聴覚・言語障害を持つ人にとって，他の人とコミュニケーションをとることは困難な場合が多く，バリアフリー社会の実現のためには，そのような不自由を解消するためのコミュニケーション支援技術が必要不可欠である。

言語障害の中に構音障害と呼ばれるものがある。これは，ことば自体は正確に理解しており話したいことばは明確であるが，声を出すための発声発語器官やその動きに何らかの異常があるためにうまく発話できない障害である。その構音障害にもいくつか種類があり，例として，構音器官そのものに起こる形態的異常による器質性構音障害，構音器官の運動を制御する神経筋系の異常による運動障害性構音障害などが挙げられる [3]。

本研究では，脊髄性筋萎縮症（spinal muscular atrophy; SMA）者を対象とする。この病気は，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 [4]。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 [5] が，その中でも重症度の高い I 型や II 型の SMA 者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。このような SMA 者は構音障害をもち，その発話は健常者とは異なるスタイルとなるため，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，(iii) 音素ごとの継続長にばらつきがみられる，などが挙げられる。

近年はこのような構音障害者を対象としたコミュニケーション支援のためのアプリケーションが開発されており，様々なテキスト音声合成（text-to-speech; TTS）ツールが使用されている。しかし，一般的な TTS アプリケーションは，学習の際に健常者の音声をもとにしているため，実際の使用者とは異なる声質の合成音声が作られる。使用者である SMA 者にとっては，その音声には自分らしさが感じられず，また複数人が同じアプリケーションを用いた際に誰が発話したものか分かりづらいなどという課題もある。そこで本研究では，SMA 者の声質を残して明瞭な音声を生成することを目指す。

ここで，重度の SMA 者の場合，生まれて間もなく発症するため，健常者のような明瞭な発話は存在しない。また，SMA 者の音声収録では，身体的負担を考慮すると大量データを防音室などの静かな環境で収録することは困難である。そのため，我々の先行研究 [6, 7] では，明瞭性のある健常者 TTS モデルを準備し，それを少量の障害者本人の音声データを用いてモデル適応することで，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。障害者本人のデータを用いてモデルを適応する際に，複数人の健常者データで学習しておいた音声認識（automatic speech recognition; ASR）モデルの損失を考慮することで，明瞭性が適応を進めていく際に失われないようにした。これにより適応の際に明瞭性を担保することができたが，実際に生成された音声は話者性の観点からすると不十分な音声であるといえた。そこで本研究では，話者認識（speaker recognition; SR）モデルから得られる話者埋め込み（speaker embedding）を利用して，合成音声の話者埋め込みが障害者データから得られる話者埋め込みに近づくような損失をさらに加えるアプローチについて検討する。

2 話者識別モデルを導入したモデル適応

2.1 x-vector による話者埋め込み

複数話者 TTS や話者照合（speaker verification）などの分野では，発話者の情報をベクトルとして表現した話者埋め込みが使用されている。話者埋め込みとして最も単純なものとして one-hot ベクトルがあるが，この場合表現できる話者の数はそのベクトルの次元数に一致し，任意の話者を表現することが困難である。そこで近年では，別の優れた話者表現として，GMM と因子分析を利用した i-vector[8]，深層学習モデルを利用した d-vector[9] や x-vector[10] などが提案されている。特に x-vector は優れた性能を持った現在主流の話者埋め込み手法であり，本研究においても x-vector に基づく話者埋め込みを利用する。

x-vector は話者識別の深層学習モデルを利用しており，その概要を Fig. 1 に示す。x-vector では初めにいくつかの層からなる TDNN（time delay neural network），プーリング層，複数の全結合（fully connected; FC）層によって構成される話者識別モデルを学習する。図中の $P(spk_1)$ は，入力として与えられ

*A study of an adaptive method considering intelligibility and individuality on speech synthesis for a person with spinal muscular atrophy. by YOSHIMOTO, Takuma, TAKASHIMA, Ryoichi (Kobe Univ.), SASAKI, Chiho (Kumamoto Health Science Univ.), TAKIGUCHI, Tetsuya (Kobe Univ.)

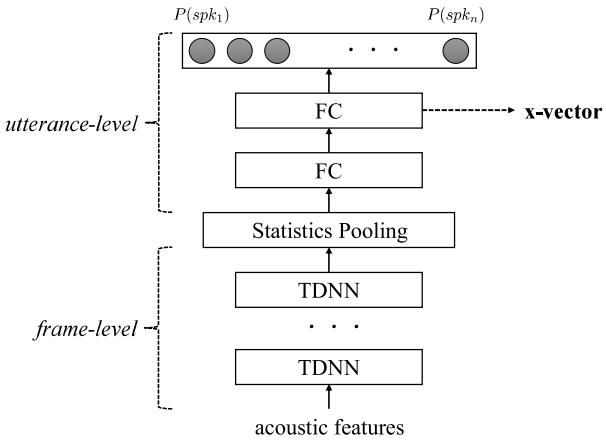

Fig. 1 The structure of DNN model for x-vector extraction.

た音響特徴量が話者 1 のものである確率[1]を表しており，n は学習の際に用いた話者の数を示す。学習後，その話者識別モデルの中間層である FC 層の出力として x-vector が得られる。

2.2 モデル適応による SMA 者音声合成

我々の先行研究 [6, 7] において，健常者データで学習した TTS モデルに対して SMA 者データを用いて適応を行い，明瞭性と話者性が両立した音声を生成することを目指した。特に [7] においては，モデル適応を行う際に，TTS モデルから得られた音響特徴量を複数の健常者データで学習しておいた ASR モデルに通すことで得られる損失を考慮することで，適応の際に SMA 者データを用いることで生じてしまう明瞭性の低下を抑えるようにした。これにより明瞭性のある音声を生成できたが，話者性の観点では改善の余地がみられた。そこで本研究では，適応の際に話者ベクトルの損失を考慮することで，明瞭性に加えて話者性も担保した音声を生成することを目指す。

本研究で提案する適応時のシステムの概要を Fig. 2 に示す。本研究で使用する音声合成システムは，音素レベルの言語特徴量から音素ごと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波形を生成するための音響特徴量を推定する音響モデル，音響特徴量からフレームごとの音素を推定する ASR モデル，音響特徴量から x-vector を抽出する SR モデルの 4 つからなる。以下では継続長モデルと音響モデルを合わせて TTS モデルと表すこととする。

学習では，はじめに健常者音声データとそのラベルを用いて健常者 TTS モデル，健常者 ASR モデルをそれぞれ学習する。ここで，TTS モデルおよび ASR モデルにはどちらも双方向 LSTM（bidirectional long short-term memory）[11] を用いている。これにより明瞭性のある健常者の音声を合成，認識することが出来るようになる。また，多人数の音声データおよびその話者 ID を用いて SR モデルを学習する。この SR モデルには，前節で述べた Fig. 1 のモデルを使用する。これにより任意の話者を表現できる話者埋め込み x-vector が抽出できるようになる。次に，学習された健常者 TTS モデルのうち音響モデルに対して SMA 者音声データとそのラベルを用いてモデル適応を行う。その際，TTS モデルの出力である音響特徴量を健常者 ASR モデルに入力し，それから得られる出力（音素）と正解ラベルとの損失を考慮する。また，TTS モデルの出力である音響特徴量を SR モデルに入力し，合成音声の話者埋め込みを獲得する。それと同時に，適応時に TTS モデルのターゲットとなる SMA 者音声の音響特徴量を SR モデルに入力し，その話者埋め込みを獲得する。そしてその 2 つの話者埋め込みが近づくような制約を加える。これらをまとめると，モデル適応の際の全体の損失 L は次の式のように表せる。

$$L = L_{acoust} + \alpha \cdot L_{recog} + \beta \cdot L_{spk} \tag{1}$$

ここで，L_{acoust} はモデル適応した TTS モデルの出力と実際の SMA 者音声から得られる音響特徴量との平均二乗誤差，L_{recog} は TTS モデルで推定した音響特徴量を健常者 ASR モデルに入力した際の出力と実際の音素ラベルとの交差エントロピー損失，L_{spk} は TTS モデルで推定された音響特徴量を話者認識モデルに入力した際に得られる話者埋め込みと実際の SMA 者音声から得られる音響特徴量を話者認識モデルに入力した際に得られる話者埋め込みとの平均二乗誤差をそれぞれ表している。健常者データで学習した音声認識モデルによる損失を加えることで，音響モデルを SMA 者データで適応する際に明瞭性が失われていくことを抑える効果が期待される。同様に，適応された TTS の出力である音響特徴量から得られる話者埋め込みを実際の SMA 者音声より抽出された音響特徴量から得られる話者埋め込みに近づけることで，適応が進むにつれて話者性を向上させる効果が期待される。

合成では，はじめに健常者データで学習した継続長モデルを用いて，入力されたテキストに対応する音素列の継続長をそれぞれ推定する。ここで継続長モデルは障害者データで適応を行っていないが，これは SMA 者のデータで適応してしまうと，音声の特徴の一つである同じ音素の継続長にばらつきがあることがそのまま反映されてしまい，非流暢で聞き取りづらくなることが懸念されるからである。しかしながらこれでは話速などの観点で話者性が反映されないとも言える。したがって，実際に音素ごとの継続長を推定する際は，継続長モデルから出力される音素継続長について，次の処理を加える。

$$d_{(syn)} = d_{(norm)} \cdot s_{un} + \bar{d}_{dys} \tag{2}$$

ここで，$d_{(norm)}$ は継続長モデルから得られる平均 0，分散 1 で正規化された音素継続長，s_{un} は健常者の音

[1]厳密には，出力層の値に softmax 関数を適用した値が確率となる。

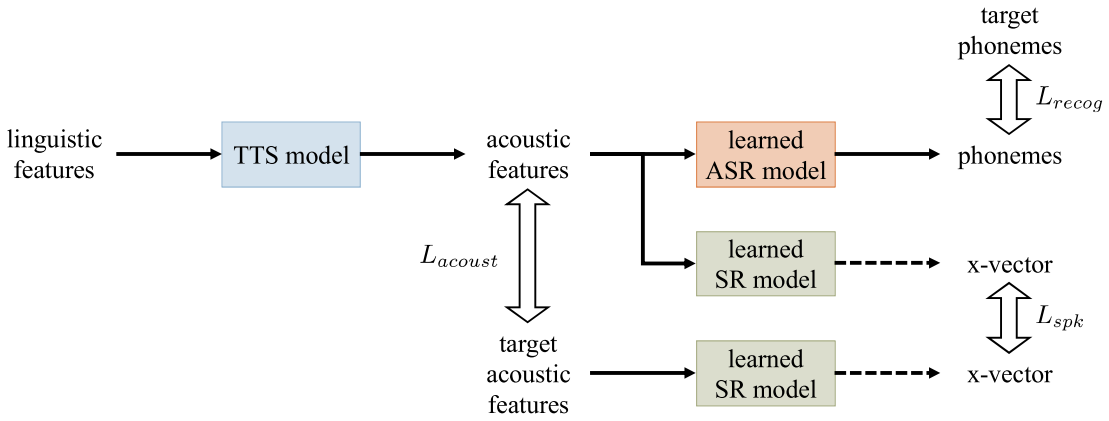

Fig. 2　Adaptation procedure on the proposed TTS system. All model parameters except the acoustic model in the TTS model are fixed during adaptation.

素継続長の標準偏差，$\bar{d}_{dys}$ は SMA 者の音素継続長の平均をそれぞれ表している。式 (2) によって得られた $d_{(syn)}$ を用いてフレームレベルの言語特徴量を作成し，それをモデル適応した音響モデルに入力することで音響特徴量が推定され，その特徴量から音声を合成する。

3　実験

3.1　実験条件

本実験で用いる SMA 者の音声データは，女性 1 名が ATR デジタル音声データベース（ATR コーパス）[12] に含まれる音素バランス 503 文を発話したものを使用する。また，各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，強制アライメントを施したうえで手作業による修正を行い作成した。健常者 TTS モデルおよび健常者 ASR モデルの学習に用いる健常者データは，ATR 音素バランス 503 文を，TTS モデルでは女性 1 名分，ASR モデルでは男女合わせて 10 名分，それぞれ使用する。SR モデルの学習には，LibriSpeech コーパス [13] に含まれる 360 時間分のクリーンな学習セット（train-clean-360）を使用する。なお，このデータセットには男女合わせて 921 人の音声が含まれている。音声のサンプリング周波数は 16 kHz，フレームシフトは 5 ms である。

ラベルには Open JTalk [14] 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する。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 [15, 7]。音響特徴量抽出およびボコーダには WORLD [16, 17] を使用する。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み計 187 次元となる。また，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。

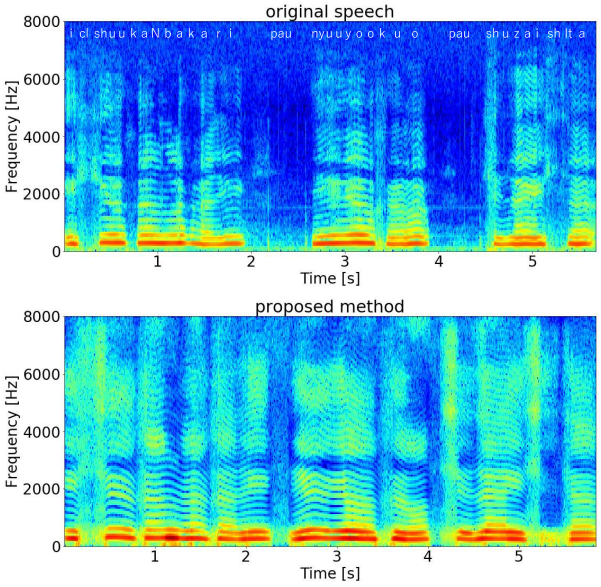

Fig. 3　Spectrograms of recorded speech (upper) and synthesized speech (lower).

TTS モデルおよび ASR モデルにおける双方向 LSTM は 3 層とし，SR モデルにおける TDNN は 4 層，プーリング層ではフレーム（時間）方向の平均を計算している。また，話者埋め込みは SR モデルの出力層の一つ手前の隠れ層の出力としている。モデル適応時の最適化には Adam を使用しており，学習率は 1e-4 とした。また，式 (1) における α および β の値はそれぞれ 0.5，0.01 に設定した。

3.2　実験結果

合成音声の例として，「一週間ばかり，ニューヨークを，取材した」という文章について，SMA 者による収録音声と提案手法による合成音声のスペクトログラムを Fig. 3 に示す。ここでは明瞭性について着目する。第 1 章でも触れたが，SMA 者の音声の特徴として，摩擦音をはじめとする子音が発声しづらい

Table 1　Mean f0 of recorded/synthesized speeches.

Method	original	previous [7]	proposal
F0 [Hz]	243.5	235.4	240.0

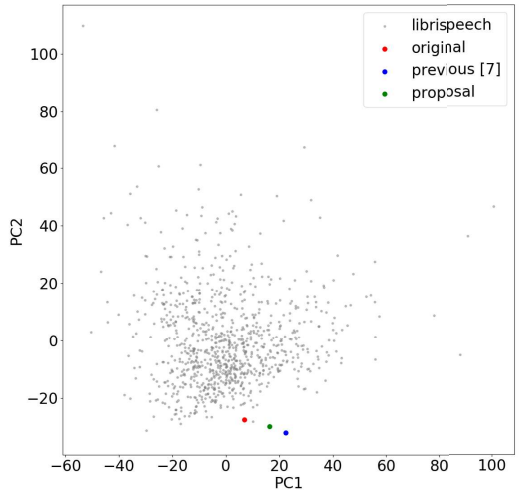

Fig. 4　Plots of speaker embeddings.

というものがある。健常者の摩擦音では通常高周波成分に強いパワーが現れるが，SMA 者収録音声では音素/sh/などの摩擦音部分での高周波成分はほかの音素と比較して強くなっているとはいえない。これに対して提案手法による合成音声では，文章中に現れる 3 回の/sh/について，いずれも高周波成分に強いパワーが確認できる。音声全体として調波構造がより明確にもなっており，収録音声と比較して合成音声は明瞭性に優れているといえる。

また，この音声に対する基本周波数（fundamental frequency; f0）の平均を，従来法 [7] の結果（すなわち式 (1) における $\beta = 0$ の場合）とともに示したものを Table 1 に示す。今回の提案法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，本人らしい声になっているといえる。

さらに，512 次元からなる x-vector を主成分分析（principal component analysis; PCA）によって 2 次元に次元圧縮を行いプロットしたものを Fig. 4 に示す。従来法による合成音声から抽出された話者埋め込みと比較して，提案法の音声から抽出された話者埋め込みは SMA 者の収録音声に近い部分に埋め込まれていることが分かる。適応時に話者埋め込みが SMA 者のものに近づくような損失を加えた効果が表れており，このことからも提案法の合成音声の方が話者性に優れているといえる。

4　おわりに

本研究では，健常者 TTS モデルを SMA 者のデータを用いてモデル適応させる際に，従来法であった健常者音声認識の損失を考慮することに加え，合成音声の話者埋め込みが SMA 者の収録音声の話者埋め込みに近づくような損失を考慮して，明瞭性および話者性が両立した SMA 者の音声を生成することを目指した。その結果，提案法は収録音声と比較して明瞭で，従来法より話者性に優れた音声を生成することができた。しかし，話者性に対する今回の提案法の効果は大きいものとはいえず，依然として収録音声との差がみられる。今後は，主観評価実験も行いながら，さらに話者性が優れた明瞭な SMA 者音声合成手法について検討する。

謝辞　本研究の一部は，JSPS 科研費 JP21H00906, JP22K12168 の支援を受けたものである。

参考文献

[1] 内閣府, “令和 4 年版 障害者白書,” 2022.
[2] 厚生労働省, “平成 30 年版 厚生労働白書,” 2019.
[3] 菊谷武 他, “歯科医師のための構音障害ガイドブック,” 医歯薬出版, 2019.
[4] 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.
[5] 小牧宏文, “脊髄性筋萎縮症の診断と治療,” 臨床整形外科, 55(1), 2020.
[6] T. Yoshimoto *et al.*, “Highly Intelligible Speech Synthesis for Spinal Muscular Atrophy Patients Based on Model Adaptation,” *Proc. 1st Workshop on Speech for Social Good (S4SG)*, 36-40, 2022.
[7] 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.
[8] N. Dehak *et al.*, “Front-End Factor Analysis for Speaker Verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, 19 (4), 788-798, 2011.
[9] E. Variani *et al.*, “Deep neural networks for small footprint text-dependent speaker verification,” *ICASSP*, 4052-4056, 2014.
[10] D. Snyder *et al.*, “X-Vectors: Robust DNN Embeddings for Speaker Recognition,” *ICASSP*, 5329-5333, 2018.
[11] M. Schuster, K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE transactions on Signal Processing*, 45 (11), 2673-2681, 1997.
[12] A. Kurematsu *et al.*, “ATR Japanese speech database as a tool of speech recognition and synthesis,” *Speech Communication*, 9 (4), 357-363, 1990.
[13] V. Panayotov *et al.*, “Librispeech: an asr corpus based on public domain audio books,” *ICASSP*, 5206-5210, 2015.
[14] “Open JTalk,” `http://open-jtalk.sourceforge.net/`
[15] 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.
[16] 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.
[17] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” *Speech Communication*, 84, 57-65, 2016.