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*

☆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

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．複数の障害を併せ持つ者を考慮しなければ，国民のおよそ 7.6%が何らかの障害を有していることになる．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．

構音障害が起こる病気の一つに，脊髄性筋萎縮症（spinal muscular atrophy: SMA）がある．脊髄性筋萎縮症者の多くは身体を自由に動かすことができないため，その人にとって声は重要なコミュニケーション手段の一つとなる．しかしながら重度の脊髄性筋萎縮症者の場合は，発声に関しても健常者と比較するとその発話のスタイルが異なるため，その音は不明瞭となり聞き取りにくくなってしまう．

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．しかし現状の TTS アプリケーションによって作成される音声は，学習に用いられた人の声をもとに作成されるため，使用者とは大きく異なる声となってしまう．また，TTS のモデルを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が必要となり，障害者にとって長時間の収録は体への負担が非常に大きくなってしまう上，作成された音声は元の障害者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] 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，本人性を維持しつつ聞き取りやすい脊髄性筋萎縮症者の音声を合成するシステムを検討した．しかしこの手法では適応の強さによって明瞭性と話者性がトレードオフの関係になっており，また最適な音声を合成するための学習率，エポックなどの調整が困難であった．そこで本研究では，従来法の損失に加えて適応時に健常者の音声認識（automatic speech recognition: ASR）モデルによる損失を考慮することで，適応の際に明瞭性の低下を抑える手法を検討する．

2 脊髄性筋萎縮症について

脊髄性筋萎縮症（spinal muscular atrophy: 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 [4]．下位運動ニューロンは脊髄内の前角細胞から走行して筋肉を支配しているため，体幹，四肢の近位部優位に筋萎縮と進行性筋力低下を示す．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．ただし，出生前に発症した重症型を零型として呼ぶ場合もある [5]．Table 1 にその分類を示す．

本研究ではその中でも I 型の脊髄性筋萎縮症者を対象とする．I 型は脊髄性筋萎縮症の中で最も重症度が高く，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．このような理由から，I 型の脊髄性筋萎縮症者の音声は健常者とは異なる発声のスタイルとなり，その音声に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．ここで「勢い」（音素は/i k i o i/）という単語について実際の音声をスペクトログラムで表示して健常者の音声と比較したものを Fig. 1 に示す．脊髄性筋萎縮症者の音声の特徴としては次のようなものがある．

- 低周波成分と比べて高周波成分のパワーが弱い（これにより子音の判別がつきづらい）
- フォルマントの変化があまり見られない（母音の判別がつきづらい）
- 同じ音素の継続長にばらつきがある

3 話者適応に基づく高明瞭度音声合成

本研究では，聞き取りづらい脊髄性筋萎縮症者の音声を，その本人らしさ（話者性）を維持しつつ聞き取りやすい（明瞭性のある）音声を生成することを目的とする．大量の脊髄性筋萎縮症者のデータを用いて音声合成システムを作成することは，話者性においては優れているものの，明瞭性という面では改善が難しい．また，そもそも大量のデータの収録は本人

*Speech clarification in persons with spinal muscular atrophy based on speaker adaptation of acoustic model. by YOSHIMOTO, Takuma, TAKASHIMA, Ryoichi (Kobe Univ.), SASAKI, Chiho (Kumamoto Health Science Univ.), TAKIGUCHI, Tetsuya (Kobe Univ.)

Table 1　Clinical classification of SMA

	I 型	II 型	III 型	IV 型
発症年齢	0〜6 か月	7 か月〜1 歳半	1 歳半〜20 歳	20 歳以降
運動機能	坐位未獲得	立位未獲得	独歩可能	正常

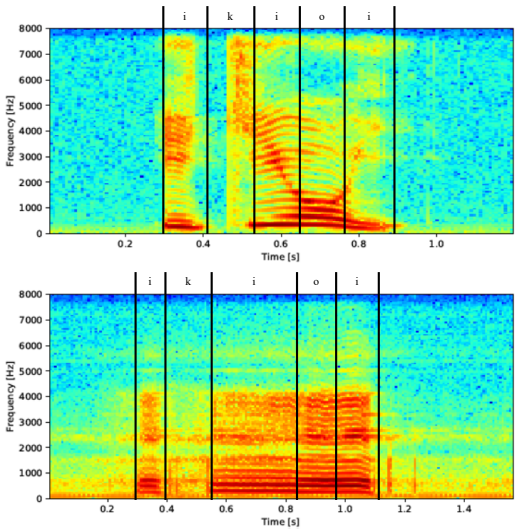

Fig. 1　Sample spectrograms (top:　a physically unimpaired person / bottom: a person with SMA)

の体への負担が大きいことから避けるべきことである．そこで，大量の健常者データではじめに音声合成システムを学習したのちに，少量の脊髄性筋萎縮症者データを用いて話者適応することを考える．先行研究 [3] では，適応の際に単純な脊髄性筋萎縮症者の音響特徴量との損失のみを考えていたが，これでは明瞭性と話者性がトレードオフの関係になっており，また最適な音声を合成するための学習率，エポックなどの調整が困難であった．そこで本研究では，従来法の損失に加えて適応時に健常者の音声認識モデルによる損失を考慮することで，適応の際に明瞭性の低下を抑えることを検討した．

Fig. 2 に今回提案するシステムの概要を示す．このシステムは，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 モデルの 3 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．

学習においては，はじめに健常者音声データとそのラベルを用いて TTS モデル，ASR モデルをそれぞれ学習する．ここで，モデルにはどちらも双方向 LSTM（Bidirectional LSTM）[6] 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．次に，学習された音響モデルに対して少量の脊髄性筋萎縮症者音声データとそのラベルを用いて音響モデルのみ再学習を行い，音響モデルの話者適応を行う．このとき，TTS モデルの出力を健常者データで学習しておいた 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も考慮する．すなわち，話者適応の際の全体の損失 L は次の式のように表せる．

$$L = L_{acoust} + \alpha \times L_{recog} \tag{1}$$

ここで，L_{acoust} は話者適応した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_{recog} は音響モデルで推定した音響特徴量を 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である．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，音響モデルを脊髄性筋萎縮症者音声に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．

合成においては，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には健常者データで学習したものをそのまま使用しているが，これは脊髄性筋萎縮症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反映されてしまい，聞き取りづらくなってしまう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本人らしさが反映されないとも言える．したがって，実際に用いる際には継続長モデルから出力される正規化された音素継続長について，次の処理を加える．

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

$d_{(norm)}$ は平均 0，分散 1 で正規化された音素継続長，s_{un} は健常者の音素継続長の標準偏差，$\bar{d}_{dys}$ は脊髄性筋萎縮症者の音素継続長の平均をそれぞれ表している．これで得られる $d_{(syn)}$ を用いてフレームレベルの言語特徴量を作成する．次に得られた特徴量を話者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，最終的にこの特徴量から音声を合成することになる．

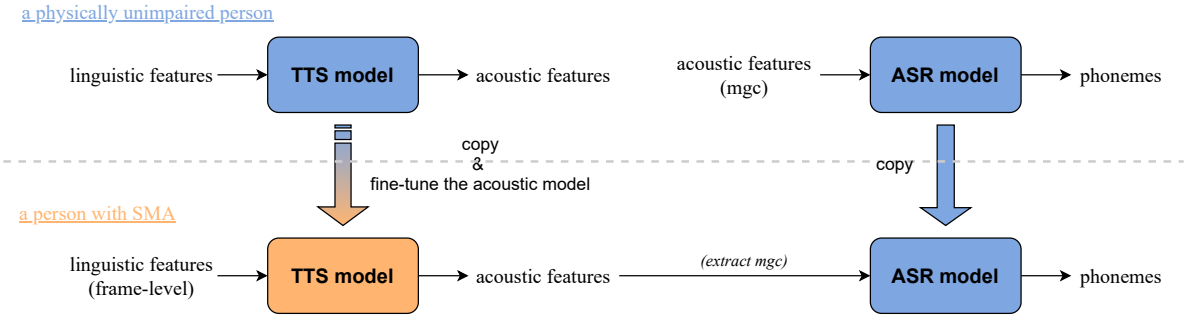

Fig. 2　Training procedure of the proposed TTS system

Table 2　Phoneme list

I	N	U	a	b	by	ch	cl	d	dy
e	f	g	gy	h	hy	i	j	k	ky
m	my	n	ny	o	p	py	r	ry	s
sh	t	ts	u	w	y	z	pau		

4　実験

4.1　実験条件

本実験では健常者，脊髄性筋萎縮症者ともに 1 名ずつのデータを使用する．脊髄性筋萎縮症者の音声には，女性の脊髄性筋萎縮症者 1 名が，ATR デジタル音声データベース [7] に含まれる音素バランス単語 216 語について 1 単語当たり 5 回発話した音声を収録したものを用いた．ただし一部収録の取りこぼしなどがあったため，実際は 5 回分発話された単語が 210 単語，4 回分だけ発話された単語が 5 単語，1 回も発話されていない単語が 1 単語となっている．また，健常者の音声は ATR デジタル音声データベースに含まれる音素バランス文 503 文を用いた．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．脊髄性筋萎縮症者の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）はすべて著者が手作業で行った．今回用いた音素体系は Table 2 に記載のものであり，pau は発話していない部分である．また，I と U は直前の無声子音によって無声化された母音，N は撥音である．健常者及び脊髄性筋萎縮症者のフルコンテキストラベルの作成には Open JTalk[8] のフロントエンド部を利用した．また音響特徴量から波形に変換するボコーダには WORLD[9, 10] を用いた．

本実験で用いる音響特徴量は，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る．また，有声/無声フラグ以外に関しては静的特徴量に加え 2 次までの動的特徴量を含んでおり，次元数は全部で 187 次元となる．音響特徴量は学習時，次元ごとに平均 0 分散 1 となるように正規化（標準化）を行った．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 [11]．話者適応時における学習率は 1e-5，エポック数は 20 とし，式 (1) における α の値について，従来法では $\alpha = 0.0$ であり，提案法では $\alpha = 2.0$ に設定した．

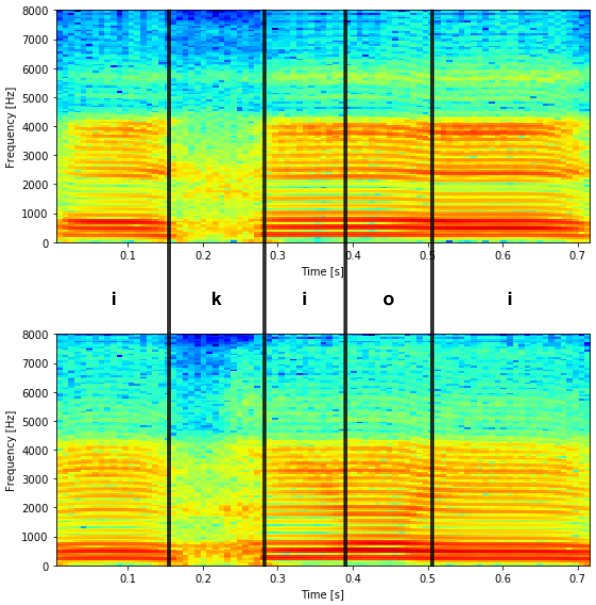

Fig. 3　Training of synthesized speech (top: a traditional method ($\alpha = 0.0$) / bottom: a proposal method ($\alpha = 2.0$))

4.2　実験結果

従来法および提案法により話者適応された音響モデルを用いて作成した合成音声「勢い」について，そのスペクトログラムを Fig. 3 に示す．この図において，従来法（上図）では適応が進みすぎたことで，Fig. 1 の下のような音声が作成されており，元の脊髄性筋萎縮症者の音声に近いため明瞭性が乏しい音声となっ

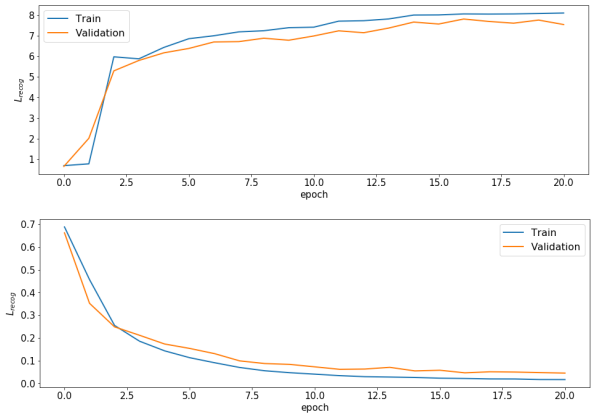

Fig. 4　Training curve of the L_{recog} (top: conventional method / bottom: proposal method)

ている．それに対して，提案法（下図）では脊髄性筋萎縮症者の音声の特徴に近づきつつも，Fig. 1 の上の音声に似た部分も残っている．例えば，勢いという単語では 3 つ目の音素から 5 つ目の音素までは母音が/i/，/o/，/i/と変化するが，健常者の場合は第 2 フォルマントが大きく変化する [12]．これは Fig. 1 の上の図で U 字状にパワーが強く出ていることで確認できる．それに対して Fig. 3 の下の図では，それよりは差が見えづらいものの U 字状にパワーが出ていることが確認できる．このことから，提案法では従来法と比べて母音の変化が現れていることが分かる．また，提案法の音素/k/の部分に着目すると，従来法よりも高周波成分が強い部分があることが確認できる．すなわち，従来法と比べて提案法による合成音声は子音が聞こえやすいことが分かる．

次に，今回新たに導入した ASR モデルに関する交差エントロピー損失の推移を Fig. 4 に示す．この 2 つを比較すると，従来法では損失が大きくなっているのに対して，提案法では損失が小さくなっている．損失が大きいとそれだけ TTS モデルの出力が音声認識しづらい音声になっているといえるため，従来法では話者適応を行うことで聞き取りづらい音声になってしまっていると考えられる．提案法では損失が大きくならずに小さくなっていることから，TTS モデルの出力は聞き取りやすい音声になっていることが分かる．

5　おわりに

本研究では，I 型脊髄性筋萎縮症者を対象として，発話スタイルが健常者と異なることによる不明瞭な音声を，健常者音響モデルの話者適応，およびその際に健常者 ASR モデルを損失に考慮することで，健常者音声の明瞭性を活かしつつ本人らしい音声を合成することを検討した．従来法と比較して，ASR モデルを組み込むことで適応が進んでも音声認識のしやすさを維持した音声を生成することが出来た．ただし，音声認識しやすいことが必ずしも明瞭性に直結しているかについては調査が不十分であるため，主観評価なども行い，話者性と明瞭性を定量的に評価する必要がある．また今後は，ASR モデルの学習方法などについてさらに調査を進め，より明瞭性と話者性を兼ねた音声が合成できるシステムの構築を目指す．

参考文献

[1] 内閣府, “令和 2 年版 障害者白書,” 2020.

[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] 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] E. Mercuri, E. Bertini, S. T. Iannaccone, “Childhood spinal muscular atrophy: controversies and challenges,” *The Lancet Neurology*, 11 (5), 443-452, 2012.

[6] M. Schuster, K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE transactions on Signal Processing*, 45 (11), 2673-2681, 1997.

[7] A. Kurematsu, *et al.*, “ATR Japanese speech database as a tool of speech recognition and synthesis,” *Speech Communication*, 9 (4), 357-363, 1990.

[8] “Open JTalk,” `http://open-jtalk.sourceforge.net/`

[9] M. Morise, F. Yokomori, K. Ozawa, “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.

[10] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” *Speech Communication*, 84, 57-65, 2016.

[11] 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.

[12] T. Hirahara, R. Akahane − Yamada, “Acoustic Characteristics of Japanese Vowels,” *Proc. ICA*, 3287 − 3290, 2004.