Vol.2021-MUS-131 No.33
Vol.2021-SLP-137 No.33
2021/6/18

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

吉本 拓真[1,a)]　高島 遼一[1,b)]　佐々木 千穂[2]　滝口 哲也[1]

概要：近年，音声信号処理技術は障害者支援に用いられており，その需要はますます増加している．本研究では，脊髄性筋萎縮症者を対象とする．脊髄性筋萎縮症（SMA）は神経筋疾患の一つであり，気管切開などによる人工呼吸器の装着，口を動かす筋肉の萎縮などが原因で，彼らの発話は健常者のものと比較して不明瞭なものとなる．そのためその言葉を聞き取ることが容易でなく，コミュニケーションを円滑にとれないという問題がある．そこで本論文では，脊髄性筋萎縮症者の発話を分析し，コミュニケーション支援のためのテキスト音声合成システムを提案する．本システムでは健常者音声にて作成したモデルを脊髄性筋萎縮症者へモデル適応するアプローチを行うことで，健常者音声に由来する明瞭性と脊髄性筋萎縮症者音声に由来する本人性を兼ね備えた音声を合成することを目指す．

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をとる際に大きな障壁となりやすく，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．

本研究では，脊髄性筋萎縮症（spinal muscular atrophy: SMA）による構音障害を対象とする．脊髄性筋萎縮症は，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 [3]．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．しかしながら脊髄性筋萎縮症者を含む構音障害者の音声は，健常者と比較するとその発話のスタイルが異なるため，発話が不明瞭となり聞き取りにくい音声となる．そこで，構音障害者のコミュニケーションを支援するために，近年では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，そのアプリケーションで使用するモデルを事前に学習する際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．使用者には「自分らしい声で会話したい」というニーズがあり，そのためには 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．しかしそれを実現するためには多くの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．

脊髄性筋萎縮症の音声に対する研究はほとんど行われていないため，本研究ではまず障害者本人の音声をどれだけ認識できるかを調査した．次に，明瞭性のある健常者音声モデルを障害者本人のものへ話者適応することにより，本人性を維持しつつ聞き取りやすい音声を作成する音声合成システムを検討した．

2. 脊髄性筋萎縮症者の音声

本研究では，音声合成システムを検討する前に，脊髄性筋萎縮症者の音声がどれだけ聞き取りづらいかを音声認識実験により調査した．また，スペクトログラムを健常者と比較し，脊髄性筋萎縮症者の音声の特徴を分析した．

[1] 神戸大学
Kobe University
[2] 熊本保健科学大学
Kumamoto Health Science University
[a)] yoshimoto_t@stu.kobe-u.ac.jp
[b)] rtakashima@port.kobe-u.ac.jp

表 1　音素リスト
Table 1　Phoneme list.

I	N	U	a	b	by
ch	cl	d	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

2.1　音声認識実験

2.1.1　実験条件

本実験では脊髄性筋萎縮症者 1 名と健常者 1 名について，特定話者の孤立単語認識実験を行った．モデルの学習および評価は，Hidden Markov Model Toolkit （HTK）[4] を用いて行った．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（ラベル:DYS）が，ATR デジタル音声データベース [5] に含まれる音素バランス単語 216 語を 1 単語当たり 5 回繰り返し発話した音声を収録したものを用いた．ただし一部の音声には収録の取りこぼしがあったため，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．健常者の音声には，コーパス内の女性 1 名（ラベル:FTK）が話す音素バランス単語 216 語音声をもとに話速とピッチを操作して 1 単語あたり 5 種類（オリジナルを含む）作成して用いた．5 回分ある発話データのうち，1 回分の発話を評価データ，残り 4 回分の発話を訓練データとして使用した．例えば 1 回目の発話の認識を行う場合は，2〜5 回目の発話を訓練データとして用いた．これを 5 回分に対してそれぞれ行い，認識結果を認識率の平均値により求めた．

実験に用いる音響特徴量には, 12 次元のメル周波数ケプストラム係数（mel frequency cepstrum coefficients: MFCC）とその 1 次微分を用いた．音声のサンプリング周波数は 16 kHz，ハミング窓長は 25 msec，フレーム周期は 10 msec である．

音声認識モデルには GMM-HMM の単語モデルと音素モデルを用意した．HMM の状態数は 3（始端終端を除く），GMM の混合分布数は 4 とした．また，発話辞書には音素バランス単語 216 語のみが含まれるようにした．音素モデルについて，日本語の音素体系には複数の定義が存在するが，今回は**表 1** に示す 36 種類を使用した．[*1]

2.1.2　音声認識結果

実験結果を**表 2** に示す．健常者の音声は単語モデル，音素モデルともにどのセットの発話を評価に用いても認識率 100%を記録した．一方，脊髄性筋萎縮症者の音声では単語モデルと音素モデルとで認識率に大きな差があった．単語モデルの結果から，健常者ほど明確な差があるわけではないが，おおよその単語の区別はつけることができるといえる．しかし，音素モデルの結果が極端に低いことから，音素モデルはうまく学習できていないことが分かる．この理由としては，脊髄性筋萎縮症者の音素体系が健常者のものに一致していないことが考えられる．

[*1] このほかに，発話でない部分を表す pau も用意した．

表 2　音声認識の結果
Table 2　ASR results.

model	DYS	FTK
word	73.88%	100%
phoneme	15.41%	100%

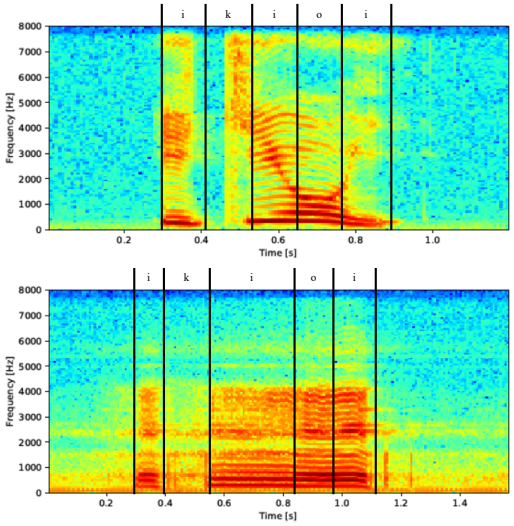

図 1　スペクトログラムの例 （上:健常者 下:脊髄性筋萎縮症者）
Fig. 1　Sample spectrograms of a physically unimpaired person (top) and a person with SMA (bottom)

2.2　スペクトログラムの比較

実際，構音障害者の発話は健常者の発話に比べて高周波成分の欠落や音素の間延びなどが起こりやすい．ここで例として，健常者と脊髄性筋萎縮症者の「勢い /i k i o i/」という発話のスペクトログラムとその音素アライメントを**図 1** に示す．健常者の音声と比較すると，脊髄性筋萎縮症者の音声は

- 低周波成分と比較して高周波成分のパワーが弱い．
- 音素ごとの継続長が一定でない（図中では 2 回目の音素/i/が間延びしている）．
- 母音の変化が明確でない（図中では音素/o/から音素/i/への変化などがスペクトログラムだけで判断できない）．

などの特徴がみられる．健常者のスペクトログラムに見られるように，子音には高周波成分が多く含まれるため，その高周波成分の弱い脊髄性筋萎縮症者の音声は，特に子音が聞き取りづらいことがわかる．

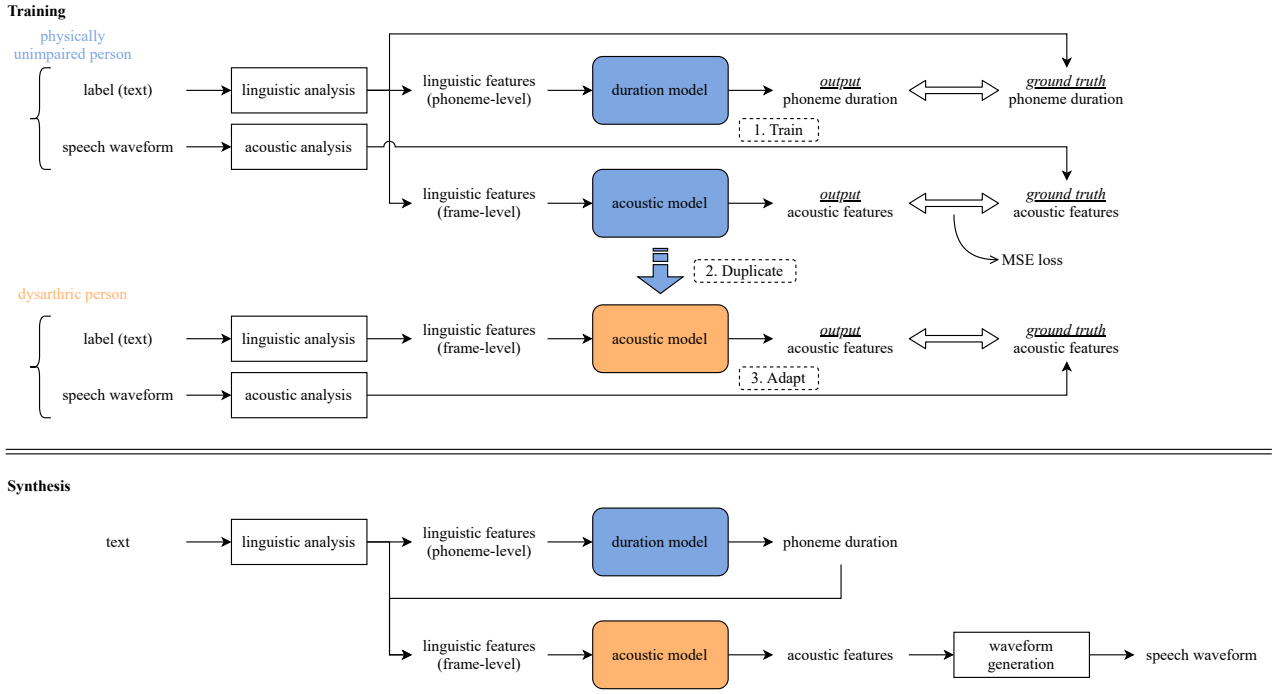

図 2 提案する音声合成システムの概要
Fig. 2 Overview of the proposed TTS system.

```
3900000 4850000 xx^sil-a+r=a/A:-2+1+4/B:xx-xx_xx/C:07_xx+xx/D:02+xx_xx/E:xx_xx!xx_xx-xx/F:4_3#0_xx@1_2|1_9/G:5_5%0_xx_1/H:xx_xx/I:2-9@1+2&1-5|1+26/J:3_17/K:2+5-26
4850000 5100000 sil^a-r+a=y/A:-1+2+3/B:xx-xx_xx/C:07_xx+xx/D:02+xx_xx/E:xx_xx!xx_xx-xx/F:4_3#0_xx@1_2|1_9/G:5_5%0_xx_1/H:xx_xx/I:2-9@1+2&1-5|1+26/J:3_17/K:2+5-26
```

図 3 ラベルの一例
Fig. 3 An example of labels.

3. 話者適応による音声合成システム

本研究で提案するシステムの概要を**図 2** に示す．ここで，図中にある両方向の矢印は学習における損失関数を表し，本研究ではいずれも平均二乗誤差（mean squared error: MSE）を用いる．また，ここでのラベルとは，**図 3** のようなフルコンテキストラベルを指す．

学習時は，はじめに従来のテキスト音声合成システムと同様に，多量の健常者音声のデータとそれに対応するラベルを用いて，音素継続長を推定する継続長モデルと音響特徴量を推定する音響モデルの 2 つを学習する．ここで，2 つのモデルにはどちらも双方向 LSTM[6] を用いている．次に，学習した 2 つのモデルのうち，音響モデルのみ複製する．最後に，少量の脊髄性筋萎縮症者音声のデータとそれに対応するラベルを用いて，複製した音響モデルに対して再学習を行うことにより，話者適応を行う．

合成時は，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含まれる各音素の継続長を推定する．次に，脊髄性筋萎縮症者データで話者適応した音響モデルを用いて，先ほど推定した音素継続長に基づいて作成したフレームレベルの言語特徴量から音響特徴量を推定する．最後に，推定した音響特徴量をもとに合成音声を作成する．

ここで，合成時，音響特徴量の推定には話者適応した音響モデルを使用するのに対し，各音素の継続長の推定には健常者データで学習した継続長モデルをそのまま使用している．これは，脊髄性筋萎縮症者の音声の特徴の一つである音素ごとの継続長が一定でない問題を解決するためである．しかし，そのままでは脊髄性筋萎縮症者の話速などの話者性が失われてしまう．したがって，本手法で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を denormalize する際に脊髄性筋萎縮症者の音素継続長の平均を活用する．フレームレベルの言語特徴量を作成する際に用いる音素継続長 $d^{(syn)}$ は次のように表せる．

$$d^{(syn)} = d^{(norm)} \times s_{un} + \bar{d}_{dys} \quad (1)$$

ここで，$d^{(norm)}$ は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表している．

4. 音声合成実験

4.1 実験条件

本実験では脊髄性筋萎縮症者の音声は，音声認識実験（2.1 節）の際に用いた収録音声と同じものを使用した．また，健常者の音声は ATR デジタル音声データベースに含

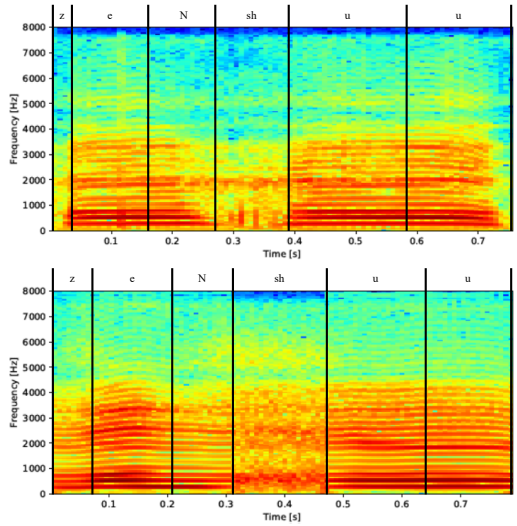

図 4 スペクトログラムの例 （上:収録音声 下:合成音声）
Fig. 4 Sample spectrograms of recorded speech (top) and synthesized speech (bottom).

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）はすべて手作業で行った．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[7] のフロントエンド部を利用した．またボコーダには WORLD[8], [9] 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．

また，実験で得られる合成音声の評価には主観評価実験を用いており，明瞭性と話者性に関する評価を行った．明瞭性は脊髄性筋萎縮症者の生音声と合成音声を比較してどちらが聞き取りやすいかを AB 評価にて，話者性は合成音声が脊髄性筋萎縮症者の生音声と健常者の音声とのどちらに近いかを ABX 評価にて行った．

4.2 実験結果

4.2.1 スペクトログラムの変化

例として「全集」という単語について，得られた合成音声と元の脊髄性筋萎縮症者の収録音声とでスペクトログラムを比較したものを**図 4** に示す．図中の音素/sh/の部分に着目する．/sh/は摩擦音であるため，健常者の音声では高

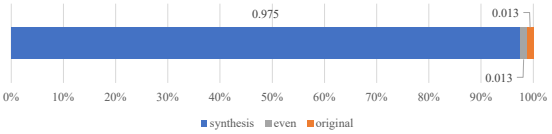

図 5 明瞭性の主観評価実験の結果
Fig. 5 Subjective evaluation of intelligibility.

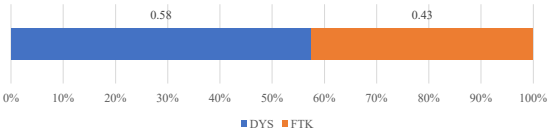

図 6 話者性の主観評価実験の結果
Fig. 6 Subjective evaluation of individuality.

周波成分が大きくなる部分である．しかし，脊髄性筋萎縮症者の収録音声では/sh/に相当する部分の高周波成分がかなり小さくなっていることが見てとれる．それに対して，合成音声では特に 5,000〜6,000 Hz 程度の成分が収録音声より大きくなっており，このことから明瞭性の改善がうかがえる．

4.2.2 主観評価実験の結果

はじめに明瞭性の評価結果を**図 5** に示す．図中の凡例 synthesis，even，original はそれぞれ「合成音声の方が聞き取りやすい」，「どちらともいえない」，「元の収録音声の方が聞き取りやすい」という選択である．

ほとんどの評価において元の収録音声よりも合成音声の方が明瞭性があるという結果が得られた．評価者からはサ行の音の明瞭性に差が大きく感じられたという意見もあった．前項でも触れたように，高周波成分が補強されたことで主に子音がよりクリアになり，また，各音素の継続長のばらつきが抑えられたことで明瞭性が改善されたと考えられる．

次に話者性の評価結果を**図 6** に示す．図中の凡例 DYS，FTK はそれぞれ「合成音声は脊髄性筋萎縮症者の収録音声に近い」，「合成音声は健常者の音声に近い」という選択である．

合成音声は脊髄性筋萎縮症者の音声に近いという結果が 60%程度で優勢となった．上手くモデル適応を行うことで多くの音声を脊髄性筋萎縮症者の音声に近づけることが出来たものの，明瞭性が損なわれないよう適応を抑える必要があったため，その影響が話者性に及んでいると考えられる．

5. おわりに

本研究では，脊髄性筋萎縮症者を対象に，音声認識を用いた発話分析を行い，その分析から得た不明瞭性の原因を改善するための音声合成について検討した．健常者音声モデルの話者適応に基づいた音声合成を行うことで，話者性

を維持しつつ明瞭性を向上させた合成音声が生成できることを示した．

現段階では話者適応を行う際に，脊髄性筋萎縮症者の特徴量との損失のみを考えている．適応が進みすぎることによる明瞭性の低下を防ぐためには，話者適応の際にも健常者の特徴量との損失も考慮するなどの検討が必要だろう．今後はモデルの構造や損失関数の改善を行い，さらに明瞭性と話者性を両立させた音声を合成することを目指す．

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] SMA 診療マニュアル編集委員会（編）：脊髄性筋萎縮症診療マニュアル，金芳堂 (2014).
[4] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D. et al.: The HTK book, *Cambridge university engineering department*, Vol. 3, No. 175 (2002).
[5] Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S.,Kuwabara, H. and Shikano, K.: ATR Japanese speech database as a tool of speech recognition and synthesis, *Speech communication*, Vol. 9, No. 4, pp. 357–363 (1990).
[6] Schuster, M. and Paliwal, K. K.: Bidirectional recurrent neural networks, *IEEE transactions on Signal Processing*, Vol. 45, No. 11, pp. 2673–2681 (1997).
[7] Open JTalk, `http://open-jtalk.sourceforge.net/` .
[8] Morise, M., Yokomori, F. and Ozawa, K.: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, *IEICE TRANSACTIONS on Information and Systems*, Vol. 99, No. 7, pp. 1877–1884 (2016).
[9] Morise, M.: D4C, a band-aperiodicity estimator for high-quality speech synthesis, *Speech Communication*, Vol. 84, pp. 57–65 (2016).