Kobe University Repository : Kernel

PDF issue: 2025-05-08

健常者音声モデルの話者適応に基づく脊髄性筋萎縮症者の高明瞭度音声合成

吉本，拓真
高島，遼一
佐々木，千穂
滝口，哲也

(Citation)
神戸大学都市安全研究センター研究報告,25:191-196

(Issue Date)
2021-03

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/81013299

(URL)
https://hdl.handle.net/20.500.14094/81013299

KOBE UNIVERSITY

神戸大学都市安全研究センター
研究報告，第25号，令和３年３月

健常者音声モデルの話者適応に基づく脊髄性筋萎縮症者の高明瞭度音声合成

High-quality speech synthesis for persons with spinal muscular atrophy using model adaptation of a physically unimpaired person

吉本拓真 [1)]
Takuma Yoshimoto
高島遼一 [2)]
Ryoichi Takashima
佐々木千穂 [3)]
Chiho Sasaki
滝口哲也 [4)]
Tetsuya Takiguchi

概要：脊髄性筋萎縮症(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、重度の場合は呼吸不全が見られる。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。本論文ではこのSMAを抱えた方の発話を分析したうえで、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。

近年は目覚ましい発展を遂げる深層学習を用いた音声合成手法は、高品質で自然な音声を生成することができる。近年ではこれらの技術を利用したコミュニケーション支援のための音声合成アプリケーションも存在するが、そこで生成される音声は学習に用いた人の音声をもとに作成されるため、使用者本人の声とは別人のものとなってしまう。

そこで本研究では音声合成システムの作成に健常者音声にて作成したモデルを脊髄性筋萎縮症者へ話者適応するアプローチを提案する。これにより、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。
キーワード：構音障害、介護支援、音声合成、深層学習

１．はじめに

わが国では、2013年に「障害者の日常生活及び社会生活を総合的に支援するための法律（障害者総合支援法)」が施行された。その中では、全ての国民が障害の有無によって分け隔てられることなく、相互に人格と個性を尊重し合いながら共生する社会を実現し、全ての障害者及び障害児が可能な限りその身近な場所において必要な日常生活又は社会生活を営むための支援を受けられることが目指されている。

本研究では、脊髄性筋萎縮症者を対象として、その音声の分析を行い、またコミュニケーション支援としての Text-to-Speech （TTS)を提案する。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TTSアプリケー

ションが開発され用いら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異なる声となってしまう。また、TTS のモデルを本人の声で学習するためには多くのデータが必要となるが、障害者にとって長時間の録音は非常に大きな負担となってしまう上、作成された音声は元の音声と同様に不明瞭なものとなる。そこで、これまで脊髄性筋萎縮症者の音声に対して、本人性を維持しつつ聞き取りやすい音声を健常者音声の話者適応により作成することを目指す。

２．脊髄性筋萎縮症者の音声分析

まず脊髄性筋萎縮症者音声と健常者音声の違いを分析するため、特定話者の孤立単語認識実験を行った。脊髄性筋萎縮症者の音声は、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[1]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。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 種類、4 回分発話された単語が 5 種類、1 回も発話されていない単語が 1 種類となっている。健常者の音声は、コーパス内の女性 1 名(ラベル:FTK)の音素バランス単語 216 語音声を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)の音声が作成されるように用意した。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。

表 1. 孤立単語音声認識実験

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

実験結果を表 1 に示す。健常者(FTK)の音声は単語モデル(word)、音素モデル(phoneme)ともに認識率 100%であった。一方、脊髄性筋萎縮症者(DYS)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、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。

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。健常者の音声と比較すると、脊髄性筋萎縮症者の音声は 1) 低周波成分と比較して高周波成分のパワーが弱い 2) 音素ごとの継続長が一定でない 3) 母音の変化が明確でないなどの特徴がみられる。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。

３．提案する音声合成手法

(1) 音声生成システムの概要

モデルの概要を図 2 に示す。学習部では、はじめに従来のテキスト音声合成[2]同様に、健常者音声のデータとそれに対応するラベル(テキスト)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。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、複製した音響モデルをもとに、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(テキスト)を用いて再学習を行うことにより、話者適応を行う。

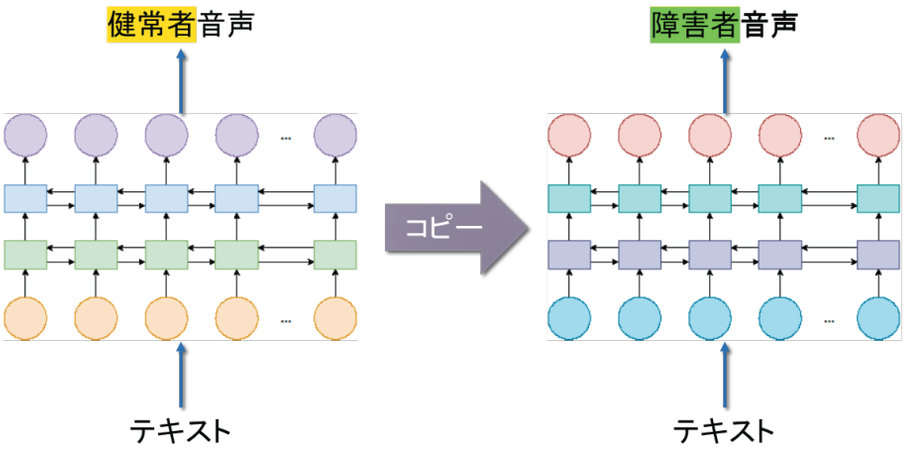

図2.　提案手法の音声合成システムの概要

合成部では、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、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。

学習及び合成を行うにあたって 2 点ほど注意すべき箇所がある。1 つは話者適応の学習について、もう 1 つは合成時の継続長の取り扱いである。話者適応が進むにつれて、モデルの出力は健常者の特徴量から脊髄性筋萎縮症者の特徴量へと変化していく。この適応が進みすぎると、健常者の特徴量で得た明瞭性を活かすことが出来ない。そのため、学習率に注意しなければならない

また、合成時には各音素の継続長は健常者データで学習した継続長モデルを使用すると述べた。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ことで、継続長の線形変換を行う。

４．評価実験

(1) 実験条件

本実験では脊髄性筋萎縮症者の音声は 2 章で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 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。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[3]のフロントエンド部を利用した。

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。

実験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。

(2) 実験結果 -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脊髄性筋萎縮症者の生音声のスペクトログラムを比較したものを図 3 に示す。

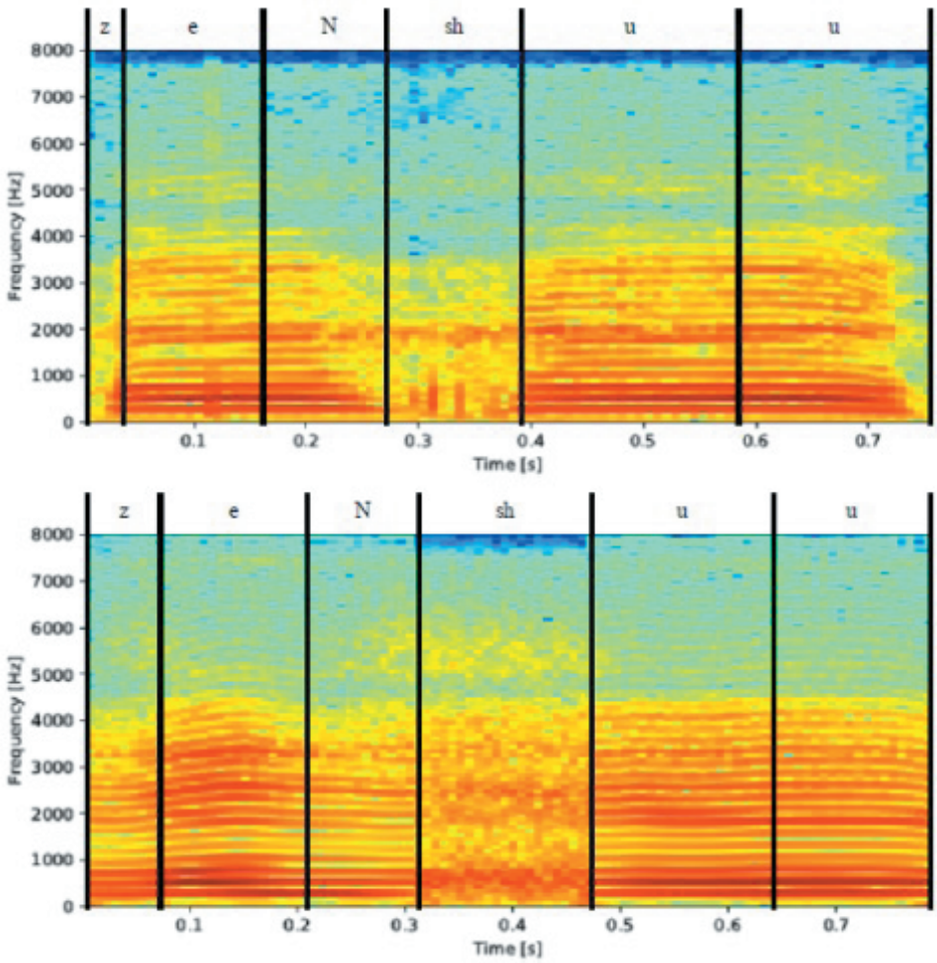

図 3. 元音声（上）と合成音下）のスペクトログラム

音素/sh/に着目すると、/sh/は摩擦音であるため健常者の音声では高周波成分が大きいはずである。しかし脊髄性筋萎縮症者の生音声では/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取れる。合成音声では特に 5,000～6,000 Hz 程度の成分が生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。

(3) 実験結果 -主観評価実験の結果-

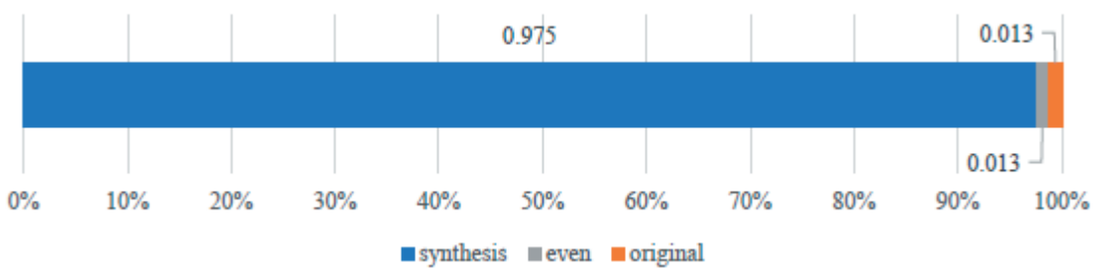

図 4. 明瞭性実験における AB 評価結果

明瞭性の評価結果を図 4 に示す。図中の 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」、「生音声(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生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。

話者性の評価結果を図 5 に示す。図中の DYS、FTK はそれぞれ「脊髄性筋萎縮症者の生音声に近い」、「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ある程度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。

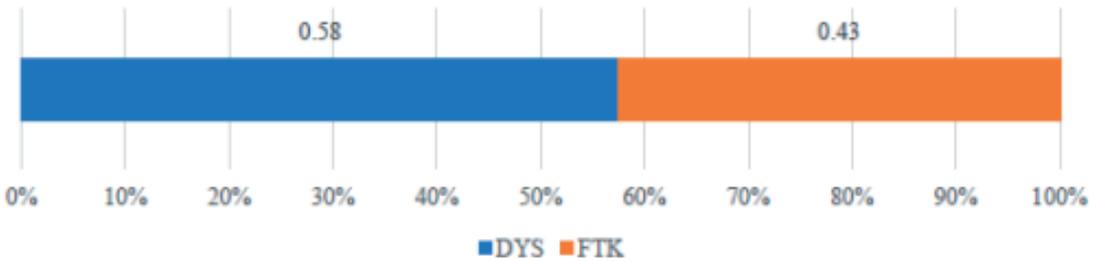

図 5. 話者性実験における ABX 評価

4．まとめ

本論文では、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。

本研究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た。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。今後は損失関数の改善を行い、さらに明瞭性と話者性を兼ね備えた音声を合成することを目指す。

参考文献

[1]: A. Kurematsu et al., “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol. 9, pp. 357-363, 1990.

[2]: Zen, Heiga, Keiichi Tokuda, and Alan W. Black. “Statistical parametric speech synthesis,” Speech Communication 51.11 (2009): 1039-1064.

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

筆者： 1）吉本拓真、情報知能工学科、学生；2）高島遼一、都市安全研究センター、准教授；3）佐々木千穂、熊本保健科学大学、准教授；3）滝口哲也、都市安全研究センター、教授；

High-quality speech synthesis for persons with spinal muscular atrophy using model adaptation of a physically unimpaired person

Takuma Yoshimoto
Ryoichi Takashima
Chiho Sasaki
Tetsuya Takiguchi

Abstract

Since a law for supporting handicapped people was established in 2013, supports for handicapped people joining society have become more strong in Japan. Recently, the information technology has been often utilized for supporting handicapped people.

In this research, we focus on people suffering from spinal muscular atrophy (SMA). SMA is an atrophy caused by a disease of motor cells in a spinal and often causes a speaking disorder. Because it is very difficult to hear the speech of who having this type of disorder, they cannot have fluent communication. In this paper, we analyze the characteristic of the SMA speech and propose a text-to-speech (TTS) synthesis system for them.

There have been various methods of the speech synthesis. Among them, methods based on the deep neural network (DNN) have been actively studied and can synthesize high quality and natural speech. TTS systems for the communication assistance using DNN have also been studied; however, the TTS system can generally synthesize only speech of the speaker whose speech is used as a training data. Therefore, if a SMA person use a conventional TTS system, it will synthesize a speech of a different speaker from the SMA person. In order to synthesize speech of a target SMA person, the system requires a lot of training data; however, it is a large burden for him/her.

In this paper, we propose a TTS system for a SMA person based on model adaptation technique. In our proposed system, we train a TTS model by using a lot of training data uttered by a physically unimpaired person, and then, we fine-tune the pre-trained model by using few training data uttered by the target SMA person.

©2021 Research Center for Urban Safety and Security, Kobe University, All rights reserved.