少量データを用いた構音障害者音声合成の健常者モデルによる明瞭性改善*

☆南阪竜翔，高島遼一， 滝口哲也 (神戸大)

1 はじめに

テキスト音声合成とは，任意に与えられたテキストから対応する音声を合成する技術である．これまでテキスト音声合成を実現するための多くの手法が提案されており， 従来手法として隠れマルコフモデル (Hidden Markov Model: HMM) を用いたもの [1] が最も代表的であった．

近年では DNN (Deep Neural Network) を用いた音声合成 [2] が，従来の隠れマルコフモデルを用いた場合と比べ高音質の合成音が作成できるため DNN を用いた音声合成が主となっている．また，高い音質と自然性を持つ音声を合成できる音声合成システムである Tacotron 2[3] を始めとした end-to-end 音声合成の研究も盛んである．

構音障害とは，口唇，舌，軟口蓋などの形態的・機能的障害により構音が正しくできなくなる言語障害であり, 種類や程度によって構音の障害となる原因は異なる．

我が国において 2016 年に「障害者差別解消法」が施行され，ハードウェア・ソフトウェアの両面から各分野で障害者差別の解消に向けた環境の整備が推進されている. このような背景から，介護・福祉分野における情報技術支援の重要性は非常に高まっている．

本研究では脳性麻痺による構音障害者を対象に実験を行う．脳性麻痺者は筋肉の不随意運動や麻痺により喉や口の筋肉を正常に動かすことができないため，健常者と比べて不安定であり，コミュニケーションが難しい．また，一般的に TTS には多くのデータが必要であるが，脳性麻痺者にとって長時間の収録は大きな負担となる．

そこで本研究では，脳性麻痺による構音障害者を対象としたコミュニケーション支援として構音障害者の少量データを用いた，テキスト音声合成 (TTS) システムを提案する．本システムは構音障害者のモデル学習時に健常者モデルを併用することで合成音の明瞭性改善を行う．

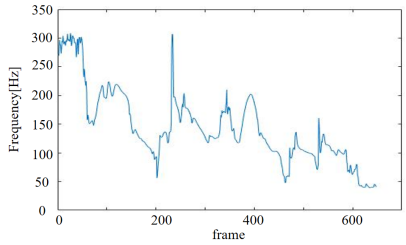
本論文では 2 章で構音障害者の発話分析，3 章で提案手法，4 章で実験条件・結果，5 章でまとめ・今後の展開について述べる．

2 構音障害者の発話分析

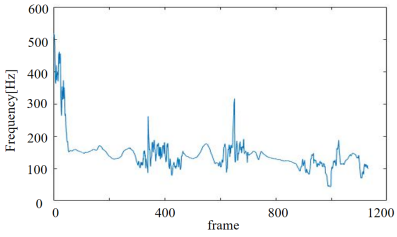
脳性麻痺者は，筋肉の不随意運動によって意図した発話ができず，健常者と比較して発話が不安定となる場合がある．また，麻痺により喉や口の筋肉を動かす事が難しく，音素が欠損したり別の音素に置き換わる場合がある．

2.1 構音障害者の基本周波数

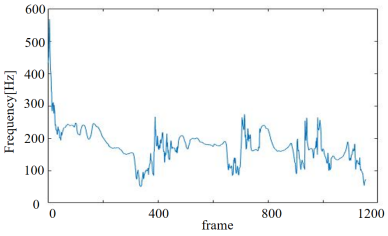
健常者と構音障害者の基本周波数を Fig. 1 に示す．健常者の基本周波数は抑揚があるが，脳性麻痺による構音障害者は筋肉を上手く動かせないため平坦な発話となっている．また発話時の緊張により筋肉が強張り，大きく乱れる場合もある．



(a) Physically unimpairedA



(b) DysarthricA



(c) DysarthricB

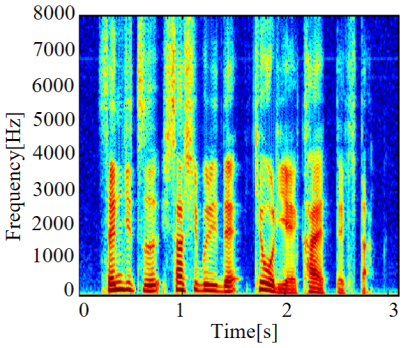
Fig. 1: Comparison of fundamental frequency

*Speech synthesis for dysarthric people using a small amount of data and improvement of clarity using a healthy person model，by Ryuka Nanzaka，Ryoichi Takashima, Tetsuya Takiguchi(Kobe univ.)

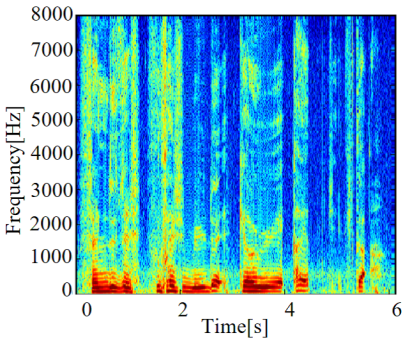
2.2 構音障害者のスペクトル特徴

健常者と構音障害者のスペクトログラムを Fig. 2 に示す．構音障害者の発話は健常者の発話に比べ音素が間延びし，発話が断続的である．

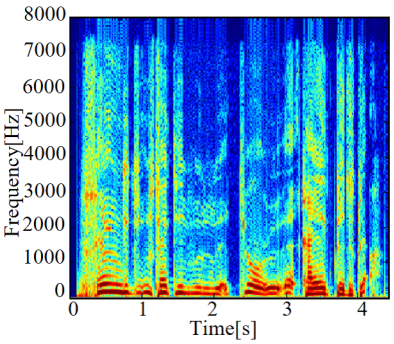
また，2000 Hz 以上の中高域の周波数帯において健常者に比べて，パワーが弱い傾向がある．これらの要因が, 発話の明瞭性の低下に影響していると考える．



(a) Physically unimpairedA



(b) DysarthricA



(c) DysarthricB

Fig. 2: Spectrograms of a healthy person and two dysarthric persons

3 提案手法

Fig. 3 に本システムの概要を示す．まず多量に用意可能な健常者のデータを用いて, テキスト情報から音素継続長を推定する duration モデルと音響特徴量を推定する acoustic モデルの 2 つを学習する．その後，acoustic モデルのみを構音障害者の少量の音声を用いて再学習を行う．

音素継続長は言語特徴量を入力，継続長を教師とする DNN を用いて推定を行う．構音障害者の発話は音素の間延び等により継続長が不安定であり，各音素の長さの比が健常者と大きく異なる場合がある．そこで，本研究では構音障害者の音声合成には健常者の duration モデルを用いるが，話者性を失わないために音素の長さの比はそのままに，音素継続長は構音障害者の平均話速へと引き伸ばす．

duration モデル，acoustic モデルともに, 双方向 LSTM[4] を用いた．本モデルは双方向の長期的な依存関係を学習する．

本章では，acoustic モデルの再学習時に健常者スペクトルを用いた学習と，健常者の音素認識モデルを用いた学習の 2 つの手法について述べる．

3.1 健常者スペクトルを用いた学習

構音障害者の acoustic モデルを学習時に，元の健常者の acoustic モデルにも同様の言語特徴量を入力し, 同じフレーム数の音響特徴量を得る．得られた特徴量のうち，スペクトル特徴量の 2000 Hz 以上に対応する次元において平均二乗誤差を取り，これを λ とする．

また構音障害者の教師の音響特徴量との Loss において，スペクトル部分を L_{spec}，それ以外の音響特徴量を L_{other} とすると，損失関数は以下で示される．

$$L_{total} = L_{other} + \lambda L_{spec} \tag{1}$$

3.2 音素認識を用いた学習

構音障害者の acoustic モデルを学習時に，推定された音響特徴量のうちスペクトル特徴量を，事前に複数話者の健常者音声で学習した音素認識モデルに入力する．得られた CrossEntropyLoss を L_{recog} とする．構音障害者の教師特徴量との Loss を L_{tts} とすると損失関数は以下で示される．

$$L_{total} = L_{tts} + L_{recog} \tag{2}$$

音素認識モデルには，TTS と同様に双方向 LSTM を用いた．

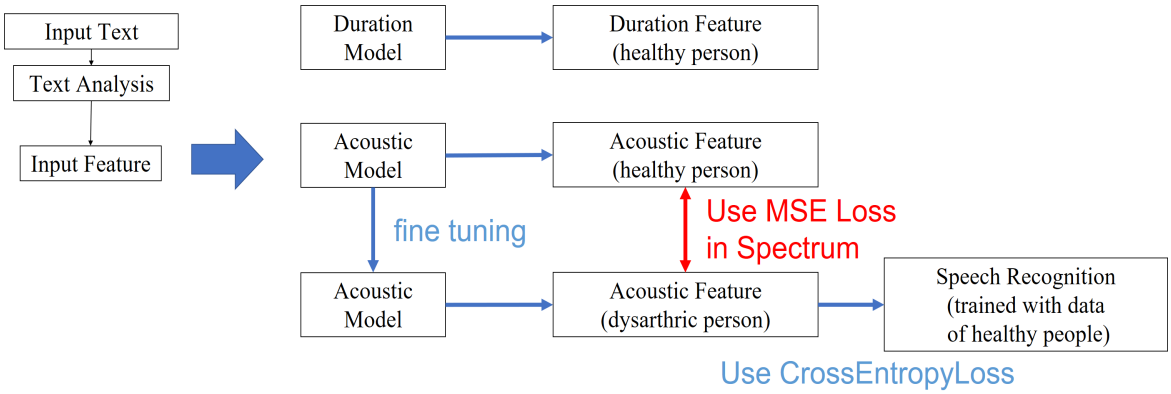


Fig. 3: Flow of Proposed Method

4 評価実験

4.1 実験条件

実験には ATR 音素バランス文 [5] を用いて学習を行った．TTS においては健常者の男性 1 名 450 文，再学習時において脳性麻痺による構音障害者の男性 1 名 100 文を 2 ペア用いた．

音素認識においては，健常者 5 名 503 文，1 名 450 文を用いた．サンプリング周波数は 16 kHz，フレームシフトは 5 ms とした．

言語特徴量にはコンテキストラベルに対して HTS 形式の Question を適用して抽出した 979 次元を使用する．脳性麻痺者の音声のアライメントは HMM を用いた強制アライメント後に，エラー箇所を修正することで求めた．

音響特徴量には WORLD[6] を用いて抽出したスペクトル包絡にメルフィルタバンクを使用し，低次 60 次元メルケプストラム係数，対数基本周波数 F0，帯域非周期性指標 1 次元とそれらの 2 次までの動的特徴量，1 次元の有声無声パラメータを用いた．

言語特徴量は最小値 0， 最大値 1 となるように，次元ごとに正規化を行った．音響特徴量は平均 0 分散 1 となるように，正規化を行った．

4.2 実験結果・考察

再学習のみ行った場合と各手法で得られたスペクトログラムとの比較を Fig. 4 に示す．spec，recog はそれぞれ 3.1，3.2 節の手法で得られたスペクトログラムである．

健常者スペクトルを用いた手法，音素認識を用いた手法ともに，中高域の周波数帯域においてパワーが確認できる．話者 A では ts，sh，s において 2000 Hz から 8000 Hz にかけて，話者 B では h において 2000 Hz から 4000 Hz にかけてパワーが残っている．

健常者スペクトルを用いた手法においては，構音障害者と健常者スペクトルとの差を取った重み λ が，2000 Hz 以上の学習を抑制していると考える．

また音素認識を用いた手法においては，健常者で学習された音素認識 Loss を加えることで，健常者らしい発話を学習し同様に高周波成分の学習の抑制効果があったと考える．

5 おわりに

本研究では，少量データを用いた構音障害者の音声合成および明瞭性改善を行った．

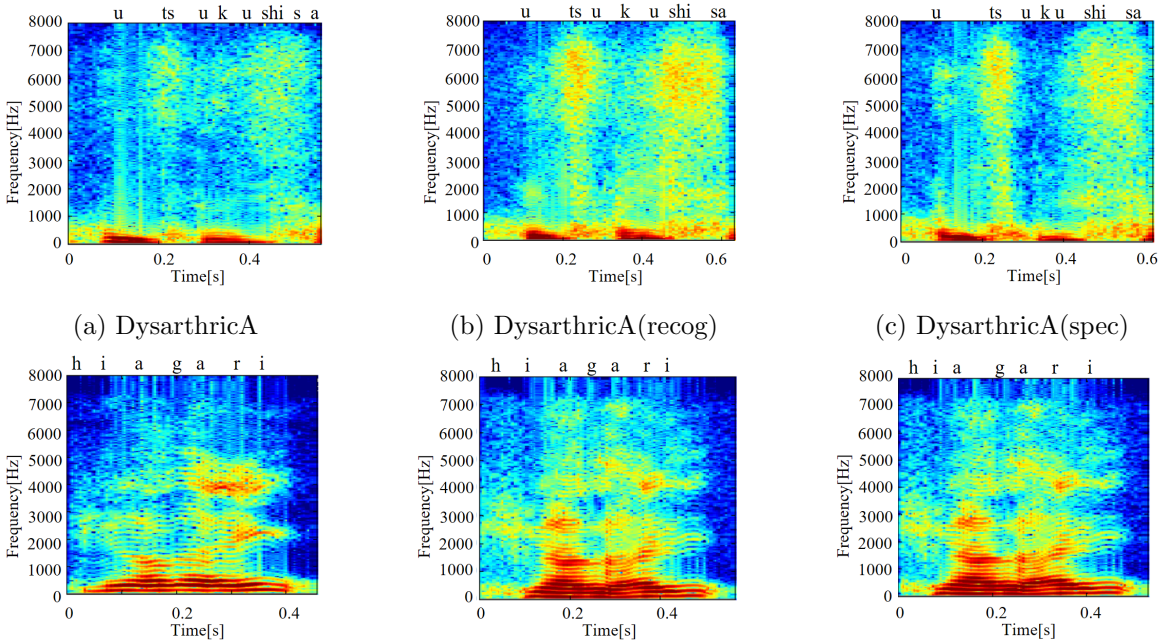
健常者スペクトルを用いた手法，音素認識を用いた手法ともに高周波成分のパワーの欠損に対して有効性を示した．しかし，健常者スペクトルを用いた手法においてスペクトル全体に重み付けを行ったが，各話者の音素毎に重み付けする必要があると考える．また，音素認識を用いた手法において得られたの合成音の音素認識率を調べる必要がある．

今回の実験では，合成された音響特徴量に対して修正は行わなかったが，前研究 [7] の健常者特徴量による基本周波数の修正を行うことでピッチのブレを抑えることができ，更に明瞭性の高い音声を生成することが可能だと考える．

謝辞 本研究の一部は，JSPS 科研費 JP17H01995 の支援を受けたものである．

参考文献

[1] K. Tokuda *et al.*, “Speech parameter generation algorithms for HMM-based speech syn-



(a) DysarthricA (b) DysarthricA(recog) (c) DysarthricA(spec)

(d) DysarthricB (e) DysarthricB(recog) (f) DysarthricB(spec)

Fig. 4: Spectrogram comparison

thesis," in *ICASSP*, 2000, pp. 1315-1318.

[2] H. Ze *et al.*, "Statistical parametric speech synthesis using deep neural networks," in *ICASSP*, 2013, pp. 7962-7966.

[3] Shen *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP*, 2018, pp. 4779-4783.

[4] Y. Fan *et al.*, "TTS Synthesis with Bidirectional LSTM based Recurrent Neural Networks ," in *INTERSPEECH 2014*.

[5] Y. Sagisaka *et al.*, "A large-scale Japanese speech database," in *ICSLP*, pp. 1089-1092, 1990.

[6] M. Morise *et al.*, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," in *IEICE*, vol. 99, no. 7, pp. 1877-1884, 2016.

[7] 南坂ら., "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," in 日本音響学会 *2019* 年秋季研究発表会, pp. 1011-1014, 2019.