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*

☆南阪竜翔，高島遼一， 滝口哲也， 有木康雄 (神戸大)

1 はじめに

脳性麻痺による構音障害者の発話は，筋肉の不随意運動により喉や口の筋肉を正常に動かすことが出来ないため，健常者と比べて不安定であり，コミュニケーションを難しくしている．また，多量の発話の収録は構音障害者に大きな負担を掛ける場合がある．

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．

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．音声合成は障害者支援にも用いられ，山岸ら [4] は様々な人々の音声を集めデータベースを構成し，ALS 患者の為の TTS システムを構築した．

TTS により直接構音障害者の音声を生成する手法もあるが，これには多くのデータが必要となる．

前研究 [5] では TTS で得られた健常者音声を，少量の構音障害者音声で学習した声質変換モデルで変換を行った．本研究では健常者音声で学習したモデルと構音障害者の音声で再学習したモデルの 2 つを用いることで音声を生成する．

2 章で提案手法，3 章で実験条件・結果，4 章でまとめ・今後の展開について述べる．

2 提案手法

Fig. 1 に本研究の概要を示す．まず健常者で音声合成モデルを生成し，そのモデルを構音障害者の少量データで再学習することによって，健常者と構音障害者の 2 つの音声合成モデルを作成する．また 2 つのモデルによるテキスト音声合成により，健常者と構音障害者，それぞれの正確なラベルが得られるため，これを用い音素の置き換え・基本周波数の変換を行い，明瞭性改善を行う．

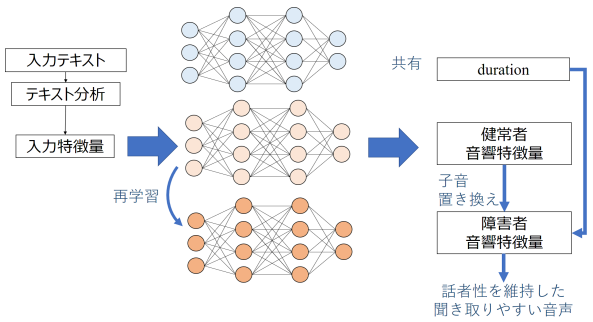


Fig. 1: 本音声合成システムの概要

2.1 DNN テキスト音声合成

まずテキストから当該音素，アクセント型，フレーム位置などの情報を抽出し言語特徴量として DNN の入力データとする．教師データとして，音声からボコーダを用いて抽出された音声パラメータを用いる．

音声合成時は任意のテキストから言語特徴量を抽出してモデルに入力する．そして得られた出力から最尤推定を用いて音響特徴量を推定し，ボコーダを用いて音声を合成する．

2.2 Bidirectional LSTM 音声合成

本研究ではテキスト音声合成に Bidirectional LSTM を用いる．Bidirectional LSTM は Recurrent Neural Networks に基いて構成される．隠れ層を $\boldsymbol{h} = (h_1, \cdots, h_T)$ ，出力を $\boldsymbol{y} = (y_1, \cdots, y_T)$， 入力を $x = (x_1, \cdots, x_T)$ ， 時系列を $t = 1 \cdots T$ とすると Recurrent Neaural Network (RNN) における，隠れ層の状態と出力の関係は以下のように示される．

$$h_t = H(\boldsymbol{W}_{xh} x_t + \boldsymbol{W}_{hh} h_{t-1} + b_h) \quad (1)$$

$$y_t = \boldsymbol{W}_{hy} h_t + b_y \quad (2)$$

W は重み行列を表す． ($\boldsymbol{W}_{xh}$ は入力と隠れ層間の重みを表す．) b はバイアスを表す． (b_h は

*Speech synthesis system using small amounts of data for articulation disorders，by Ryuka Nanzaka，Ryoichi Takashima, Tetsuya Takiguchi， Yasuo Ariki (Kobe univ.)

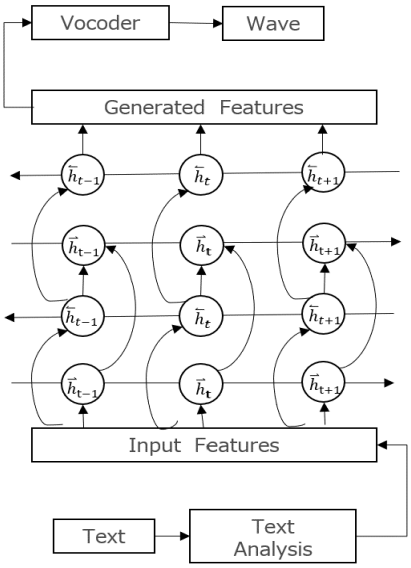


Fig. 2: Bidirectional LSTM テキスト音声合成

隠れ層のバイアスを示す．) H は非線形活性化関数を表す．LSTM では非活性化関数 H は以下のように示される．

$$i_t = \sigma(\boldsymbol{W}_{xi}x_t + \boldsymbol{W}_{hi}h_{t-1} + \boldsymbol{W}_{ci}c_{t-1} + b_i) \quad (3)$$

$$f_t = \sigma(\boldsymbol{W}_{xf}x_t + \boldsymbol{W}_{hf}h_{t-1} + \boldsymbol{W}_{cf}c_{t-1} + b_f) \quad (4)$$

$$c_t = f_t c_{t-1} + i_t tanh(\boldsymbol{W}_{xc}x_t + \boldsymbol{W}_{hc}h_{t-1} + b_c) \quad (5)$$

$$o_t = \sigma(\boldsymbol{W}_{xo}x_t + \boldsymbol{W}_{ho}h_{t-1} + \boldsymbol{W}_{co}c_t + b_o) \quad (6)$$

$$h_t = o_t tanh(c_t) \quad (7)$$

σ は sigmoid 関数，i, f, c は入力ゲート，忘却ゲート，出力ゲート，メモリーセルを表す．

Bidirectional RNN は隠れ層においてフォワード状態とバックワード状態の 2 つを持ち，以下で示される．

$$\overrightarrow{h_t} = H(\boldsymbol{W}_{x\overrightarrow{h}}x_t + \boldsymbol{W}_{\overrightarrow{h}\overrightarrow{h}}\overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}) \quad (8)$$

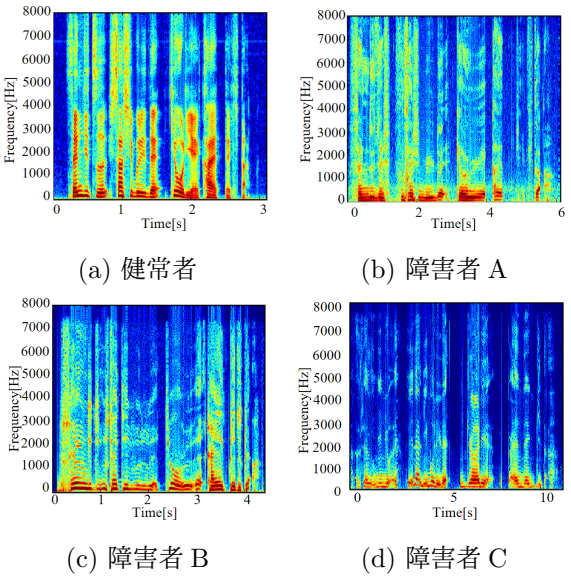
$$\overleftarrow{h_t} = H(\boldsymbol{W}_{x\overleftarrow{h}}x_t + \boldsymbol{W}_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}) \quad (9)$$

$$y_t = \boldsymbol{W}_{\overrightarrow{h}y}\overrightarrow{h}_t + \boldsymbol{W}_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y \quad (10)$$

Fig. 2 に Bidirectional LSTM によるテキスト音声合成の概要を示す．DNN 音声合成は隠れマルコフモデルを用いた音声合成に比べて高音質な音声合成が出来るが音声の時間的変動を考慮しているとは言えなかった．Bidirectional LSTM では各時刻の過去と未来の情報を考慮する．

2.3 健常者モデルの利用

本システムでは，テキスト情報から音素の継続長を推定する 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．



(a) 健常者　(b) 障害者 A

(c) 障害者 B　(d) 障害者 C

Fig. 3: 健常者と構音障害者のスペクトログラム

Fig. 3 に示される通り，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ため，聞き取りにくい原因となる．

そこで，acoustic モデルのみを構音障害者の音声で再学習し，duration モデルは健常者のものを使用することで明瞭性の改善を行う．しかし，話者性を考慮するために，母音の長さは構音障害者の平均継続長へと引き伸ばしを行う．

2.4 子音置き換え

一般的に，脳性麻痺による構音障害者は筋肉の不随意運動によって上手く発話が出来ず，子音の欠損が生じることが多く，発話の明瞭性に影響を与える．Fig. 3 により，話者によって差異はあるが，およそ 2000Hz から 4000Hz に掛けてパワーが弱いことが分かる．

これを解決するために，構音障害者の合成音に対し，2000Hz から 4000Hz において健常者モデルで得られた同一発話長の子音で置き換えを行うことで明瞭性改善を行う．本研究では特に欠損が見られた子音 s,sh,ts,m,k,h,y,r に対して置き換えを行った．

2.5 基本周波数 (F0) 変換

健常者から構音障害者への F0 変換は以下の式を用いて行う．

$$f0'_t = \frac{\sigma_{trg}}{\sigma_{src}}(f0_t - \mu_{src}) + \mu_{trg} \quad (11)$$

$f0'_t$ は変換後のフレーム t の F0, μ_{trg}, σ_{trg} は構音障害者の F0 の平均と分散，μ_{src}, σ_{src} は健常者の

F0 の平均と分散である．$f0_t$ は構音障害者のフレーム t の F0 である．

3　評価実験

3.1　実験条件

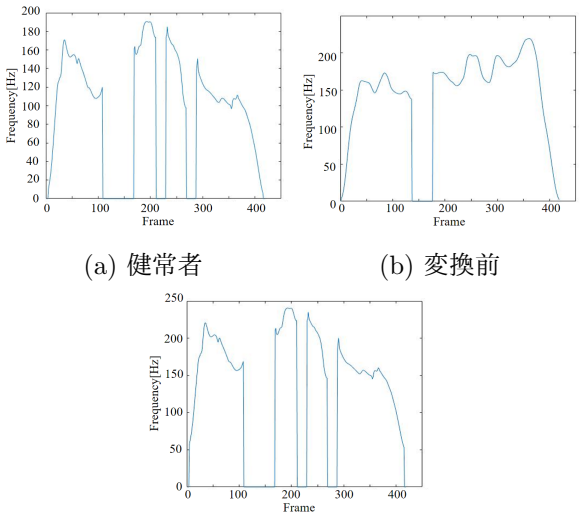
実験データには脳性麻痺による構音障害者の男性 3 名，健常者の男性 1 名を使用した．健常者モデルは ATR 音素バランス文 [7]450 文を用いて学習を行った．そしてそのモデルに対し，構音障害者の音声 50 文，100 文を用いて再学習を行った．サンプリング周波数は 16kHz，フレームシフトは 5ms とした．

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．音響特徴量には WORLD[8] 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．

3.2　実験結果・考察

3.2.1　基本周波数の変換



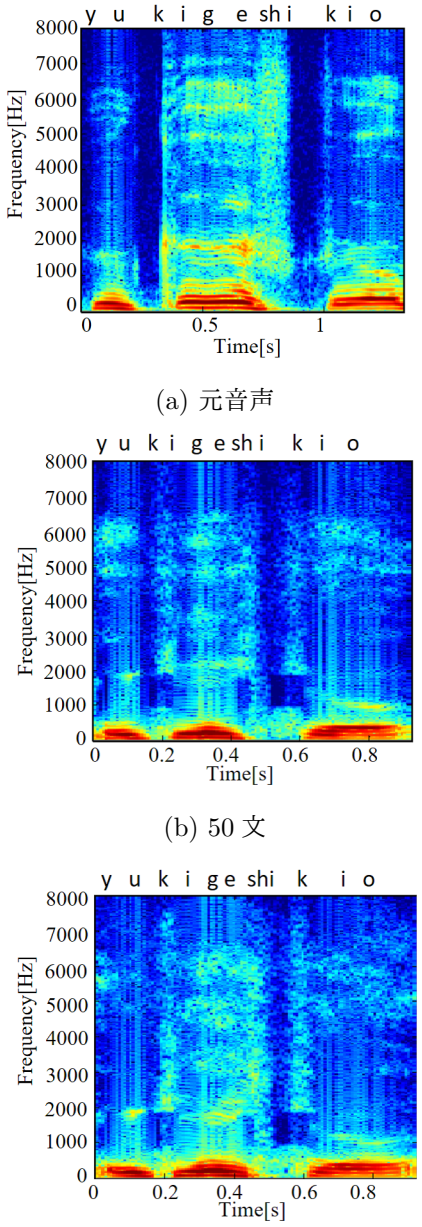
(a) 健常者　(b) 変換前

(c) 変換後

Fig. 4: 推定および変換された F0

基本周波数の変換結果を Fig. 4 に示す．構音障害者の平坦な基本周波数から平均は変えずに，健常者の抑揚のある基本周波数へと変換されていることが分かる．また，少ない学習データによりピッチの推定ミスが生じた場合にもこの変換によって音声の劣化を最小限にすることが可能である．

3.2.2　学習データ量による合成音の比較



(a) 元音声

(b) 50 文

(c) 100 文

Fig. 5: 元音声と合成音とのスペクトログラム比較

学習データ毎のスペクトログラム比較を Fig. 5 に示す．50 文，100 文での再学習において，いずれも構音障害者の発話の特徴を上手く学習している．また置換により sh,k の 2000Hz から 4000Hz にかけて，パワーが付加されていることが分かる．

しかし 50 文において, スペクトログラムの細

部が失われる傾向があった．

4　おわりに

本研究では，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を行った．3 人の脳性麻痺による構音障害者の発話の解析を行った結果，2000Hz から 4000Hz においてパワーが弱くなる傾向が見られた．そこで健常者の子音を用い，構音障害者の欠損した子音の置換を行うことで，明瞭性改善を行った．実験ではそのまま子音を置換を行ったが，滑らかな置換の検討が必要である．

構音障害者の学習データを 100 文から 50 文に減らした際に，スペクトログラムの細部が失われ音質劣化が見られたが，これは Bidirectional LSTM を用いた本モデルの表現力に起因する．

また基本周波数の変換を行うことにより，抑揚のある音声合成を実現した．しかし，構音障害者の基本周波数の推定ミスが大きい場合については検討が必要である．

今後，少量データを用いた話者適応モデルを用いることで精度改善を行う．また，話者ごとに置換する周波数帯域・子音の検討を行う．

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 synthesis,” 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] R. Ueda *et al.*, “Individuality-Preserving Voice Reconstruction for Articulation Disorders Using Text-to-Speech Synthesis,” in *ACM ICMI*, pp. 343-346, 2015.

[4] J. Yamagishi *et al.*, “Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction,” in *Acoustical Science and Technology*, vol. 33, no.1, pp. 1-5, 2012.

[5] R. Nanzaka *et al.*, “Hybrid Text-to-Speech for Articulation Disorders with a Small Amount of Non-Parallel Data,” in *APSIPA*, pp. 1761-1765, 2018.

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

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

[8] M. Morise *et al.*, “WORLD: A vocoder-based high-quality speech synthesis system for real-time applications,” in *The Institute of Electronics, Information and Communication on Engineers*, vol. 99, no. 7, pp. 1877-1884, 2016.