構音障害者の少量学習データによる音声合成の検討*

☆南阪竜翔， 滝口哲也， 有木康雄 (神戸大)

1 はじめに

本研究では脳性麻痺などにより健常者と異なった発話となる構音障害者のコミュニケーション支援としてのテキスト音声合成 (Text-To-Speech) 及び声質変換を提案する．

構音障害者は健常者と比べて多くの発話データを収録することが困難である．しかし DNN によるテキスト音声合成では多くのデータ量が必要となる．そこでまず健常者で音声合成モデルを生成し，得られた合成音を声質変換することで構音障害者の音声合成システムを作成する．

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 [3] 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．

声質変換とは，入力音声に対して，発話内容は保持しながら，話者性や感情といった情報を変換する技術である．統計的声質変換の代表例として GMM (Gaussian Mixture Model)[4] がある．近年では深層学習に対する注目の高まりから DNN を用いたもの [5] もある．これらの声質変換には同一内容，同一発話長の発話が必要であるが，それを必要としない適応型制限ボルツマンマシン (Adaptive Restricted Boltzmann Machine) を用いた研究 [6] も行われている．

本稿では，Bidirectional LSTM[7] によるテキスト音声合成と GMM (Gaussian Mixture Model) による声質変換を用いた少量学習データによる音声合成について述べる．

2 提案手法

テキスト音声合成では Bidirectional LSTM を用いる．声質変換には GMM を用いる．Fig. 1 に本研究の概要を示す．

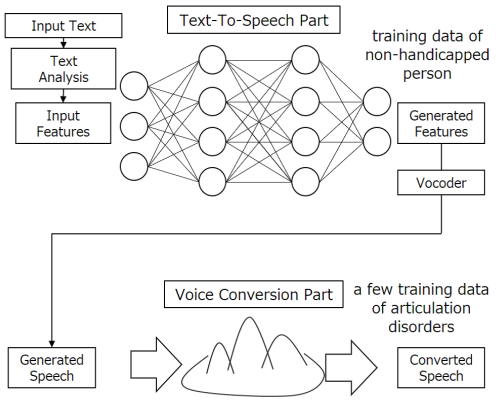

Fig. 1: A flow of proposed speech synthesis system

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する．その情報をバイナリや連続値で表現し，フレームで連結したものを言語特徴量として 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と動的特徴量をフレームで連結したものを用いる．

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．

2.1.1 Bidirectional LSTM 音声合成

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

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

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．

2.2 GMM 声質変換

GMM は，データの生成される確率を複数のガウス分布の重み付き和で表すモデルである．Fig. 3 に GMM 声質変換の概要を示す．入力特徴量系列を $x = [x_0, x_1, \cdots, x_t]^T$，出力特徴量系列を $y = [y_0, y_1, \cdots, y_t]^T$ とする．今回はメルケプストラムを特徴量に用いた．入力，出力特徴量を結合したベクトルを $Z = [x_t^T, y_t^T]^T$ とする．確率モデル関数は以下のように示せる．

$$p(Z|\lambda^{(Z)}) = \sum_{k=1}^{K} \pi_k N(Z; \mu_k^{(Z)}, \Sigma_k^{(Z)}) \quad (11)$$

添字 k は k 番目のガウス分布を表し，π_k はガウス分布の混合係数を表し総和は 1 である．μ_k はガウス分布の平均ベクトル，Σ_k はガウス分布の分散行列を表し，以下のように表せる．

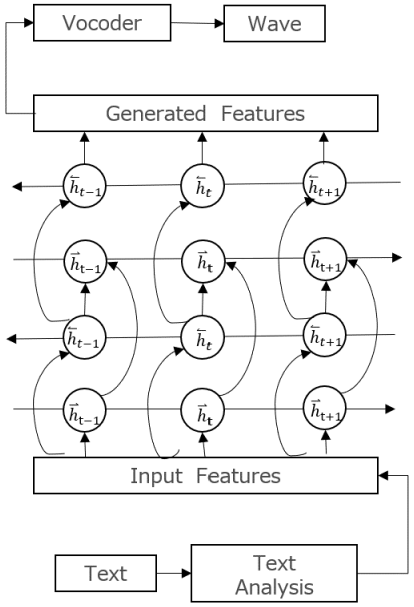

Fig. 2: Bidirectional LSTM speech synthesis system

$$\Sigma_k^{(Z)} = \begin{bmatrix} \Sigma_k^{(xx)} & \Sigma_k^{(xy)} \\ \Sigma_k^{(yx)} & \Sigma_k^{(yy)} \end{bmatrix} \mu_k^{(Z)} = \begin{bmatrix} \mu_k^{(x)} \\ \mu_k^{(y)} \end{bmatrix}$$

パラメータは最尤推定で学習する．声質変換された特徴量は以下の式で導出される．

$$\hat{y}_t = E[y_t|x_t] = \sum_{k=1}^{K} P(k|x_t, \lambda^{(Z)} E_{k,t}^{(y)} \quad (12)$$

ここで

$$P(k|x_t, \lambda^{(Z)}) = \frac{\pi_k N(x_t; \mu_k^{(Z)}, \Sigma_k^{(xx)})}{\sum_{n=1}^{K} \pi_k N(x_t; \mu_n^{(Z)}, \Sigma_n^{(xx)})} \quad (13)$$
$$E_{k,t}^{(y)} = \mu_k^{(y)} + \Sigma_k^{(yx)} \Sigma_k^{(xx)^{-1}} (x_t - \mu_k^{(x)}) \quad (14)$$

である．

2.2.1 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．

$$f0'_t = \frac{\sigma_{trg}}{\sigma_{src}}(f0_t - \mu_{src}) + \mu_{trg} \quad (15)$$

$f0'_t$ 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．

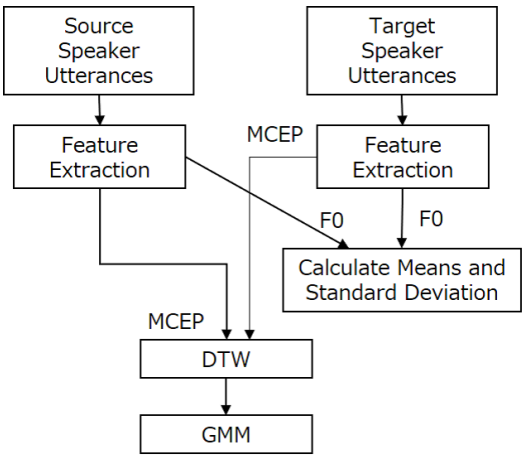

Fig. 3: GMM Voice Conversion system

3 評価実験

3.1 実験条件

実験データには構音障害者の男性 1 名，健常者の男性 1 名を使用した．健常者の選択にはヘッドホンを用いた聴取実験により構音障害者に一番声質が近い話者を選択し，ATR 音素バランス 450 文を学習，43 文を評価，10 文をテストに用いて実験を行った． 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[8] 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[9] 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．

声質変換においてはテキスト音声合成において得られた音声と構音障害者の ATR 音素バランス 50 文を学習に，10 文をテストに用いた．GMM 学習には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を用いた．基本周波数は式 (15) で変換を行った．帯域非周期性指標は入力話者のものを用いた．発話長は DP マッチングにより同一発話長とした．

評価基準として，メルケプストラム歪み (Mel-cepstrum Distortion:MelCD) を用いた．比較対象として，構音障害者の少量のデータのみを用いた合成音を用いた．学習データはそれぞれ 50 文または，100 文を用いて比較した．

3.2 実験結果・考察

声質変換によって得られた音声のスペクトログラムを Fig. 4 に示す．

構音障害者の音声は健常者に比べて高周波成分が弱い傾向がある．変換後のスペクトログラムにおいても 2000Hz 以上のパワーが弱くなっている．また低周波数域のフォルマントにおいても変換前に比べ構音障害者に近づいていることが分かる．

示した音声は「青い青い海は女性の美しさを持っている.」という文であるが，特に「持っている.」の部分において変換前のスペクトログラムから構音障害者のスペクトログラムへと近づきが見られた．今回入力話者の duration をそのまま用いたが，構音障害者の duration を用いることで更に話者性を考慮した音声合成が可能であると考える．

MelCD による評価結果を Fig. 5 に示す．評価結果はテスト 10 文の平均を取っている．構音障害者の 50 文，100 文で音声合成システムを構築した場合と比べ，提案手法の (健常者合成+) 声質変換を用いることにより，学習データ 50 文でもいい結果が得られた．

4 おわりに

本稿では，Bidirectional LSTM による TTS と GMM による声質変換を組み合わせた構音障害者の少量学習データによる音声合成についての提案を行った．

Fig. 4 より構音障害者のスペクトログラムにおける高周波成分特徴や低周波におけるフォルマントが上手く学習されていることを確認した．今後は少量の構音障害者データのみ用いたテキスト音声合成と提案手法の音声合成の主観評価による比較を行う．

また，パラレルデータを得るために DP マッチングを行ったが，アライメント誤りが音質に影響している可能性がある．アライメントを必要としない非パラレルデータ声質変換についても検討を行い，比較する．

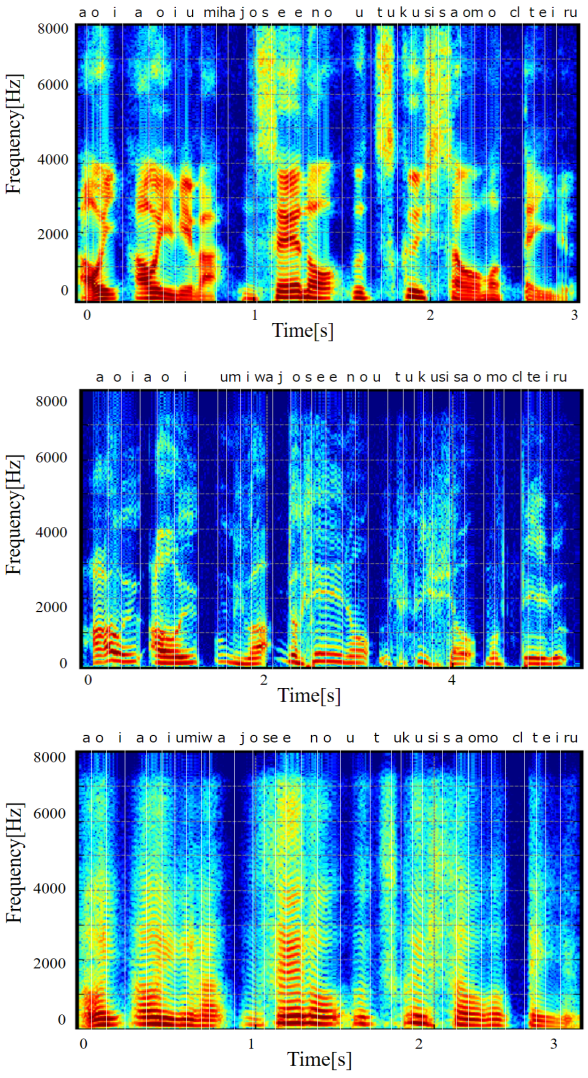

Fig. 4: Comparison of spectrogram
(Top: Source,Middle: Target,Bottom: Converted)

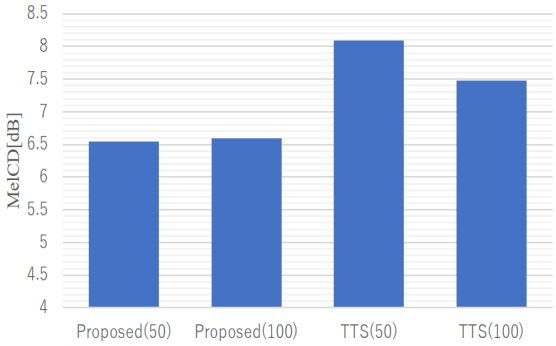

Fig. 5: Mel-cepstrum Distortion

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] 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.

[4] T. Toda *et al.*, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," in *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 8, 2007.

[5] S. Desai *et al.*, "Voice conversion using Artificial Neural Networks," in *ICASSP*, 2009.

[6] T. Nakashika *et al.*, "Non-Parallel Training in Voice Conversion Using an Adaptive Restricted Boltzmann Machine," in *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 24, No. 11, 2016

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

[8] Z, HeiGa *et al.*, "HMM 音声合成システム (HTS) の開発 ," in *IPSJ SIG Technical Reports*, 2007.

[9] M. Morise *et al.*, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications ," in *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877-1884, 2016.