DNNを用いた聴覚障害者の音声合成の検討*

☆北村毅（神戸大），滝口哲也（神戸大/JST さきがけ），有木康雄（神戸大）

1　はじめに

本研究では，先天性の聴覚障害からなる構音障害者を対象としている．構音障害の中でも，聴覚障害と脳性麻痺からなる場合［1］とでは，発話の障害となる要因は異なる．聴覚障害者の場合は，筋肉を自由に動かすことができるため，共鳴腔で作られた響きから舌や歯を使い子音や母音の発音を容易に行える．しかし，聴覚障害者は発音を訓練により習得するが，耳が聞こえないため，発音があいまいになる場合や発音の方法が健常者と異なる場合があり，発話内容が伝わりにくい．そのため，健常者と聴覚障害者のコミュニケーションは手話通訳者を介する，もしくは筆談が主に用いられる．例えば手話通訳者の代わりに音声合成の使用を試みることも考えられるが，現状の音声合成システムでは，聴覚障害者の話者性を十分に反映させることが出来ない．

音声合成技術は近年，深層学習の進歩に伴い音質を向上させている．スマートフォンの音声合成にも Deep Neural Network (DNN)が用いられており, 従来の Hidden Markov Model (HMM)を用いた場合と比べ，高い音質の合成音を作成できることが報告されている[2]．さらに，平均声モデルを作成することで少量の学習データのみで合成音を作成できる適応技術［3］や，高い自然性を持つ合成音の作成が可能な WaveNet［4］が登場している．本稿では，DNN を用いて，聴覚障害を持つ人々を支援するためのテキスト音声合成法を提案する．

聴覚障害者の発話は，言語習得時に母音などの発音をあいまいに，または，ある子音だけ極端に発話するケースがある．よって，聴覚障害者音声を用いて学習したモデルから得られる合成音も聞き取りが難しくなる．そこで，本研究では，健常者音声を用いて聴覚障害者の音声パラメータを修正し，それらを用いて DNN を学習することで，話者性を維持しつつ，より聞き取りやすい合成音を作成するシステムの実現を目指す．

2　従来の音声合成法

DNN を用いた音声合成の手法は文献[2, 5]で提案されている．Fig. 1 に従来の音声合成の概要を示す．まずテキストから当該音素，周辺音素，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値もしくは整数値で表現し，それらを連結し，言語特徴量として DNN の入力データとして使用する．教師データには，音声にボコーダから抽出された音声パラメータと，動的特徴量を計算し，それらを連結した教師データを用いて学習を行う．音声合成時は，テキストから言語特徴量を抽出し，モデルに入力して得られる出力から音声パラメータを合成し，ボコーダを用いて音声を合成する．

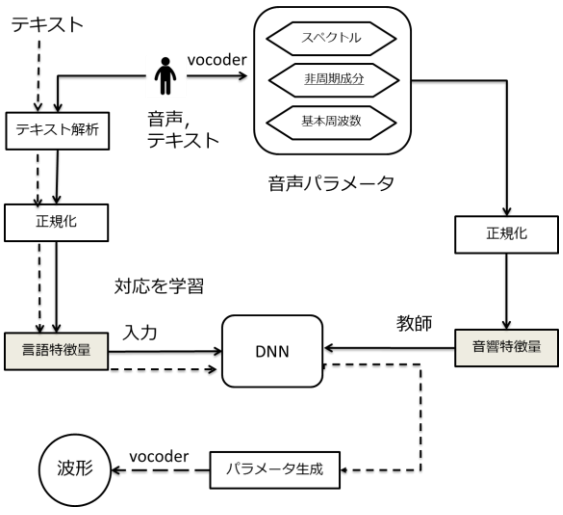

Fig. 1　Text-to-Speech system; solid arrows indicate the training stage; dash arrows indicate the synthesis state

3　提案手法

Fig. 2 は提案手法における DNN の学習手法の概要である．健常者と聴覚障害者の両方の特徴量を，変換及び選択し DNN を学習して

* Speech Synthesis System Using Deep Neural Networks for Hearing Disorders. by Tsuyoshi Kitamura (Kobe University), Tetsuya Takiguchi (Kobe University/ JST PRESTO), Yasuo Ariki (Kobe University)

いる．変換した音声パラメータは，F0 とスペクトルである．

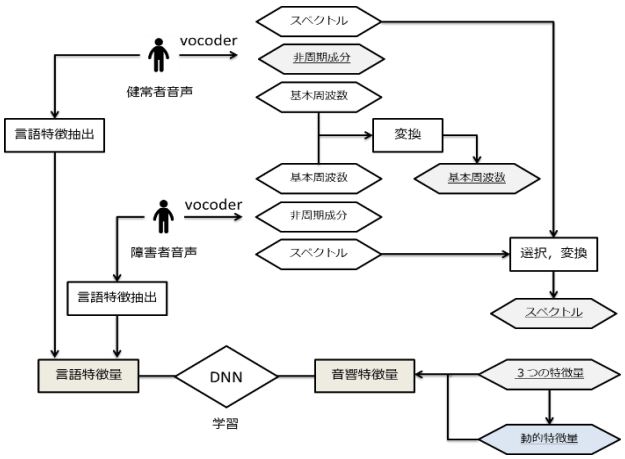

Fig. 2　Flow of the proposed method

3.1 F0 の修正

聴覚障害者の発話は，抑揚がなく淡々と話す場合や単語中のアクセント位置が健常者と異なる場合があり，聞き取りを難しくする原因となっている．本研究では，健常者の F0 系列に変換したものを用いて，DNN を学習した．聴覚障害者の話者性を維持するため，健常者の F0 系列を以下の式(1) を用いて変換する．

$$\hat{\omega}_t = \frac{\sigma_x}{\sigma_w}(\omega_t - \mu_w) + \mu_x \qquad (1)$$

ただし，$\hat{\omega}_t$ は変換後のフレーム t の対数 F0，ω_t は健常者のフレーム t の対数 F0 である．μ_x，σ_x は聴覚障害者の対数 F0 系列の平均と分散，μ_w，σ_w は健常者の F0 系列の平均と分散である．

3.2 スペクトルの修正

発音練習をした聴覚障害者は，正しく発音できない音素でも，同じ音素は同じ発音方法で発音するため，正しく発音できない音素と正しく発音できる音素とで分けることができる．Fig. 3 に健常者と聴覚障害者の「熱風が」の発話のスペクトルを示す．健常者のスペクトルと比べて全体的に高周波の成分が欠落しているため，こもった響かない声となる．また，発話中の「Q(促音)」の長さが約 0.3 秒であるが，発話全体の話速が速いため，発話が急に途切れたように感じる．また，「p」の子音が相対的に強く発話されており，自然な発話となっていない．以上から，話速（音素継続長）及びスペクトルを修正する必要があると考えられる．

スペクトルの修正部では，音声中の正しく発音できない音素を，健常者の音素に置換，もしくは低域や高域の周波数成分のみを置換し，学習を行うことで聞き取りやすい合成音の生成が期待できる．DNN では，言語特徴量と音響特徴量がフレーム単位で結びついているので，音素を健常者と聴覚障害者で入れ替えた場合，対応する言語特徴量等も入れ替えて修正を行った．

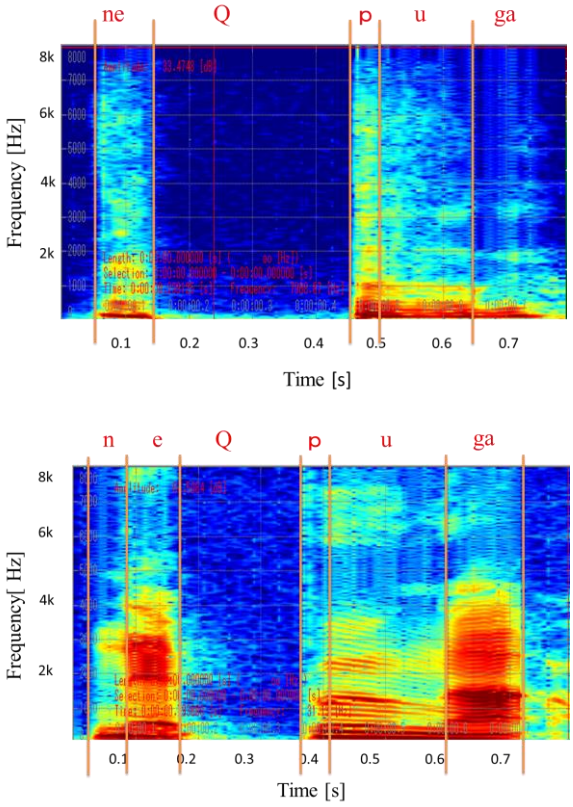

Fig. 3　Example of spectrograms for an utterance (/neppuga/ in Japanese) spoken by a hearing disorder (top) and a physically unimpaired person (bottom)

3.3 提案手法の流れ

まず，WORLD［6］を用いて健常者と聴覚障害者の音声パラメータ（基本周波数（F0），スペクトラム包絡，非周期成分（AP））を抽出する．特徴量を抽出した後，健常者の F0 と障害者のスペクトルに対して修正処理を行う．その後，それらの特徴量を用いて DNN を学習する．AP，F0 及び有声無声パラメー

タ用の DNN と，スペクトルを学習するための２つの DNN を学習した．これは，スペクトルに対して修正処理を行っており，フレーム数が増減しているため，対応する AP 等の成分が存在しないためである．AP 等の学習時には，修正処理を行っていないスペクトルを教師データの一部として使用している．

作成したモデルに，言語特徴量を入力して得られる DNN の出力から音響パラメータを生成し，WORLD を用いて音声を合成する．合成時の音素継続長は，あらかじめ健常者と聴覚障害者それぞれに対し，コンテキスト依存ラベルから音素継続長モデルを作成し，求めている．聴覚障害者の音素継続長が健常者と比較して極端に異なる音素は，健常者のモデルから音素継続長を作成した．

4　評価実験

4.1 実験条件

実験データには聴覚障害者の男性 1 名，健常者の男性 6 名を使用した．音声は健常者と聴覚障害者ともに ATR 音素バランス 503 文を用いた．サンプリング周波数は 16kHz，フレームシフトは 5ms とした．

DNN の入力には，414 次元の言語特徴量を用いた．この言語特徴量は，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．学習時の聴覚障害者音声の音素継続長は強制アライメントにより求めた．

本研究では，2 つの DNN を用いた．1 つ目は AP，F0，有声無声パラメータの推定用であり，2 つ目はスペクトルの推定用である．前者の教師信号として，WORLD を用いて抽出した音声パラメータから求めた 60 次元メルケプストラム係数, 対数 F0, 25 帯域 AP と，これらの Δ，ΔΔに有声無声パラメータを加えた 259 次元のデータを用いた．F0 は式(1)修正を行った. 後者の教師データは 60 次元メルケプストラム係数とその Δ，ΔΔでありスペクトルは健常者のデータを用いて修正している．どちらの DNN も入力データは平均 0 分散 1 に，教師データは 0~1 の値に次元ごとに正規化した．DNN の隠れ層の数は 4 層で，各層 1024 個のユニットを持つ. 隠れ層は各層 Batch Normalization を使用しており，活性化関数にはシグモイド関数を用いた．最適化アルゴリズムは Adam を用いた．

健常者 6 名の中から 1 人，メルケプストラムの距離が聴覚障害者に近い話者を Mel-CD (Mel-cepstrum distortion)により測定し，この話者に F0 とスペクトルの修正を行った．残りの 5 名のデータで DNN の学習を行い，このモデルパラメータを初期値として転移学習を行った. DNN の出力から音声パラメータを生成する際は，MLPG アルゴリズム[7]を用いることで動的特徴量を考慮した．

提案法の有効性を示すため，12 人の日本人話者に対してヘッドホンを用いた主観評価実験を行った．話者性の評価基準として，MOS (Mean Opinion Score)により 5 段階(5:非常に似ている，4:とても似ている，3:まぁまぁ似ている，2:似ていない，1:全く似ていない)を用いた．そして聞き取れた単語の個数の割合を評価した．単語は読み上げる単語の一覧を与えておらず，スクリプトなしで聞き取れるかを評価している．

4.2 実験結果と考察

Fig. 5 に主観評価実験の結果を示す (Prop:提案手法，Conv:従来手法，Orig:原音声). Fig. 5 は単語の理解した個数の割合を示している．発話の明瞭性は，提案手法が原音声を上回っていることが確認できる．Fig. 6 に，提案手法と従来手法により作成した文章中の「認識し」のスペクトルを示す．従来法で音素の境界が不明瞭な部分でも，提案手法では明瞭である．また，子音「s」の高周波の摩擦成分が存在する．子音「N」を無音区間で表している従来手法と比べて，提案手法では「N」の成分が確認できる．Fig. 5 と Fig. 6 より，健常者の特徴量を学習に利用することで，聞き取りやすい音声を生成できた．

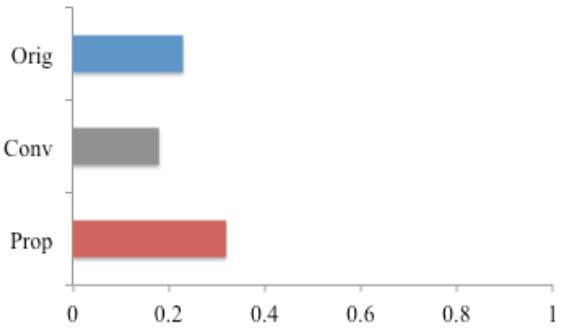

Fig. 5　Correct answer rate

Fig. 7 では，話者性が Ref2 より低い結果となった．この理由として，特徴的なアクセントや音素継続長を健常者のモデルを用いて修正したことが考えられる．つまり，イントネーションなどが話者性として評価されたためと考えられる．

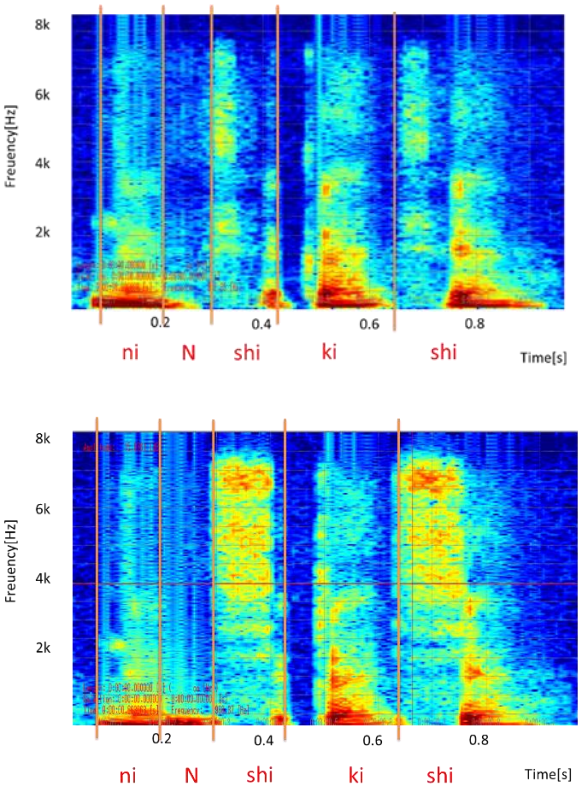

Fig. 6 Spectrogram of a conventional synthesized speech (top) and a proposed synthesized speech (bottom)

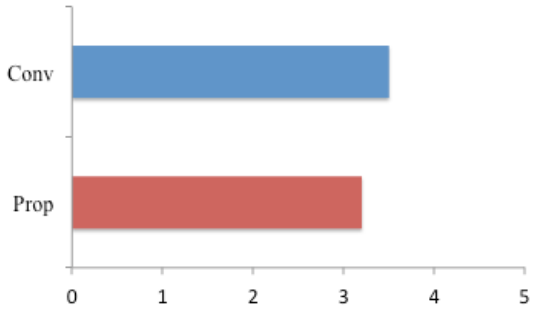

Fig. 7 MOS test on similarity

5 おわりに

本研究では Deep Neural Network を用いた聴覚障害者のための音声合成の方法を提案した．主観評価実験により，提案手法により生成した合成音が，パラメータ修正を行わない従来手法の合成音と比較してより明瞭な音声であることを示した．

今後は，母音等の低周波域にパワーを持つ音素の修正を検討する．また，聴覚障害者特有のこもった声を明るい声に変換するため，フォルマント成分を推定するネットワークの構築を検討する．

謝辞

本研究の一部は，JST さきがけの支援を受けたものである．

参考文献

[1] Reina Ueda, Tetsuya Takiguchi, Yasuo Ariki, “Individually-Preserving Voice Reconstruction for Articulation Disorders Using Text-to-Speech Synthesis, ” ICMI2015, pp.343-346, 2015

[2] H.Zen, A.Senior, and M.Schuster, “Statistical parametric speech synthesis using deep neural networks,” in Proc, IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2013

[3] S.Takaki, S.Kim, J.Yamagisi, “Speaker Adaptation of Various Components in Deep Neural Network based Speech Synthesis,” Speech Synthesis Workshop pp.167-173, 2016

[4] WaveNet: A Generative Model for Raw Audio https://deepmind.com/blog/wavenet-generative-model-raw-audio/

[5] Z.Wu and S.King, “Minimum Trajectory error training for deep neural networks, combined with stacked bottleneck features,” Proceedings of Interspeech 2015, pp.309-313, 2015

[6] M.Morise, “An attempt to develop a singing synthesizer by collaborative createon,” in Proc. Stockholm Music Acoustic Conf., 2013, pp.287-292.

[7] K.Tokuda, T.Yoshimaura, T.Kobayashi, T,Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2000