構音障害者を対象としたDNN音声合成*

☆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

本稿では，先天性の聴覚障害からなる構音障害者を対象として，音声合成を用いて彼らのコミュニケーションを支援するシステムを提案する．構音障害の中でも，聴覚障害と脳性麻痺からなる場合とでは発話の障害となる原因は異なる．私たちは，脳性麻痺からなる構音障害者を支援するシステム [1] を提案してきた．聴覚障害者の場合は，共鳴腔で作られた響きから舌や歯を用いて母音や子音を発音するための器官に不自由がない．しかし，耳が聞こえないため発音が曖昧になる場合，発話テンポなどの発話スタイルが健常者と異なる場合があり発話内容が伝わりにくく，健常者との発話を用いたコミュニケーションを難しくしている．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．

近年，テキスト音声合成 (Text-To-Speech) の枠組みとして，Deep Neural Networks (DNNs) を用いた音声合成 [2, 3] が広く研究されている．スマートフォンのアプリケーションなどで使用されており，高い自然性を持つ合成音が作成できる．さらに，多人数話者から平均声モデルを作成することで，少量の特定話者の音声データから合成音を作成する適応技術 [4] や，より高い自然性を持つ音声を作成できる WaveNet[5] が開発されている．本稿では，テキスト音声合成を応用することにより聴覚障害者の発話支援システムを提案する．

聴覚障害者の発話は，イントネーションが不安定である場合や，一部の母音や子音を曖昧，もしくは極端に発話する場合があるため，聞き取りが難しい．そのため，聴覚障害者の収録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．本研究では，健常者の音声パラメータを用いて構音障害者のモデルを修正することで，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 DNN音声合成

DNN 音声合成の学習時は，入力となるテキストを解析して得られる言語特徴量と，教師となる音声を分析して得られる音響特徴量の関係を DNN を用いて学習する．Fig. 1 に DNN 音声合成の合成時の概要を示す．

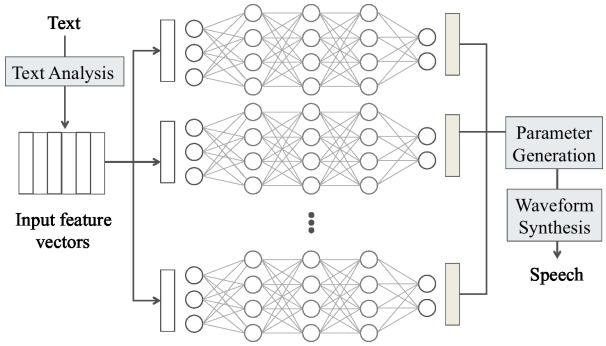


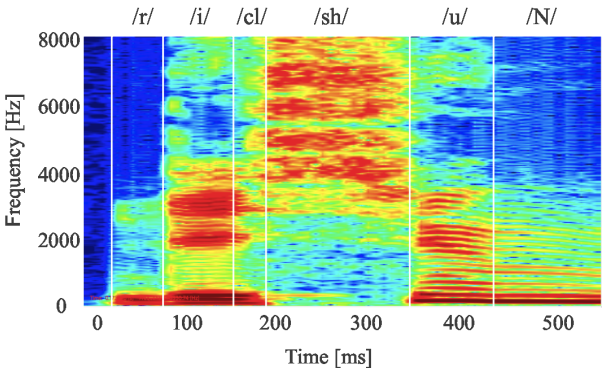
Fig. 1 A flow of speech synthesis using deep neural networks.

言語特徴量を DNN に入力して得られた音響特徴量には，静的特徴量と二次までのデルタ特徴量が含まれており，MLPG アルゴリズム [6] を用いてトラジェクトリを求め，ボコーダを用いて音声波形を生成する．本研究では，学習基準に Trajectory Training [3] を用いている．これは，学習時にデルタ特徴を含む DNN の出力である音響特徴量と教師との二乗誤差を DNN に逆伝播させるのではなく，DNN の出力に対して MLPG アルゴリズムを適用しトラジェクトリを導出し，教師との二乗誤差を求め誤差を伝播することで DNN を学習する基準である．

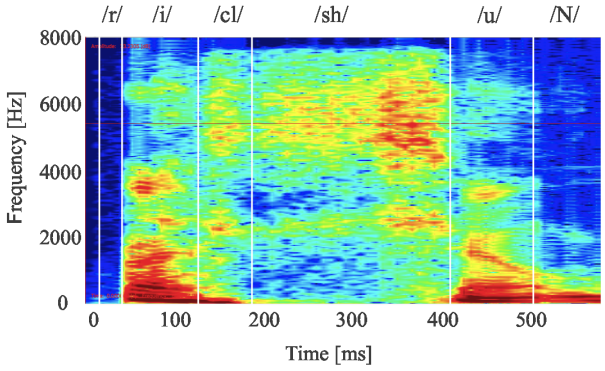
3 Two Type DNNsを用いた音声合成

本研究では，聴覚障害者と健常者の音声を用いる．Fig. 2 に健常者と聴覚障害者の「立春」(/r/ /i/ /cl/ /sh/ /u/ /N/) という発話のスペクトログラムを示す．Fig. 2 から，健常者と比較して聴覚障害者は高域のエネルギーが弱くこもった発話となっている．また，Fig. 2 では健常者と聴覚障害者の発話時間は概ね等しいが，音素継続長が間延びする音素 (ex: /cl/, /sh/) や欠落する音素 (ex: /r/) が存在する．これらが音声の明瞭度を下げている一因であり，話者性を維持しつつ明瞭度の高い音声を作るため，聴覚障害者と

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



(a) a physically unimpaired person.



(b) a person with hearing loss.

Fig. 2　Sample spectrograms of /r i cl sh u N/.

健常者の音声から得られる特徴量の両方を使用する．

3.1　音響周波数修正

Fig. 3 に基本周波数 F_0 の修正方法の概要を示す．健常者と比較して構音障害者の発話は抑揚が少ない．Table 1 に聴覚障害者と健常者 3 名の F_0 の平均と分散を示す．なお，F_0 の導出には音声分析システム WORLD[7] を用いており，平均と分散の算出時には無音区間は省いた．Table 1 から，対象とした聴覚障害者の分散が小さいことからイントネーションに変化が少なく，淡々と同じ高さで話していることから聞き取りが難しい原因となっている．

聴覚障害者の合成音のイントネーションを修正するため，合成時に健常者の F_0 の概形を入力として用いるネットワークを用いる．まず，学習時には健常者と聴覚障害者の 2 つの DNN を独立に学習する．健常者の DNN は入力に言語特徴量，教師に健常者の音響特徴量を使用して学習を行う．聴覚障害者の DNN は

Table 1　F_0 の平均と分散

	聴覚障害者	1	2	3
平均	107	137	118	113
分散	644	1058	929	734

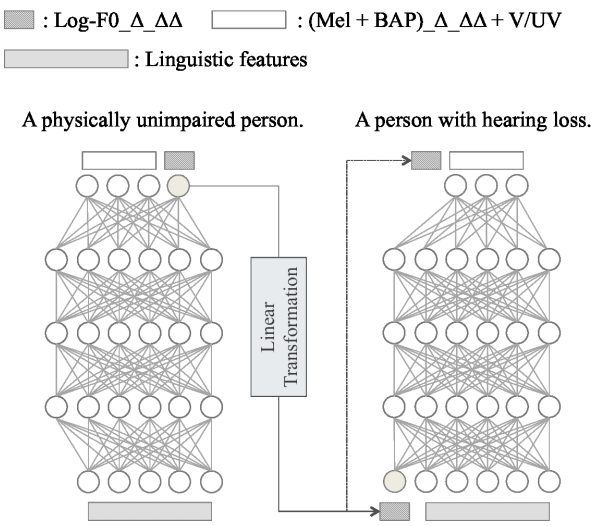


Fig. 3　A flow of the F_0 and spectrum modification method.

入力に言語特徴量と聴覚障害者の F_0 特徴量，教師としてスペクトル特徴量と非周期性指標を用いて学習を行う．これにより，合成時に入力の F_0 に対応したフォルマント構造を持つスペクトルを推定する聴覚障害者 DNN を構築する．

Fig. 3 に 2 つの DNN を用いた F_0 修正及びスペクトル推定について示す．合成時は，健常者 DNN に対して言語特徴量を入力して得られる F_0 特徴量に，式 (1) を用いて線形変換を行うことで F_0 系列の平均値を聴覚障害者の値に修正する．

$$\hat{w}_t = \frac{\sigma_x}{\sigma_w}(w_t - \mu_w^{(F_0)}) + \mu_x^{(F_0)} \tag{1}$$

式 (1) において，w_t は健常者のフレーム t の対数 F_0，$\mu_x^{(F_0)}$ と σ_x はそれぞれ w 系列の平均と分散，$\mu_w^{(F_0)}$ と σ_x はそれぞれ聴覚障害者の対数数 F_0 系列の平均と分散を表す．この F_0 特徴量と言語特徴量を聴覚障害者 DNN に入力として用いることで，入力 F_0 に対応したスペクトルを得ることが可能であり，推定されたスペクトル特徴，非周期性指標及び入力として用いた F_0 から音声を生成する．なお，式 (1) を用いて F_0 系列の平均値を聴覚障害者の値に変換した理由は，声の高さに話者性が多く含まれていることに加えて，聴覚障害者の学習データ中に含まれている F_0 系列から外れた高さを持つ F_0 を入力すると，未知データの推論となるためである．

合成時の言語特徴量は，健常者と聴覚障害者について同じ Duration 情報を用いる必要があるが，次節の音素継続長修正を用いて推定した Duration を用いる．

3.2　音素継続長修正

Fig. 2 で示した通り，聴覚障害者の音素継続長が不安定であるため，発話のテンポが健常者と異なる

場合がある．そこで，発話のテンポ (各音素の長さの比率) に健常者の値を用いて，平均音素継続長 (各音素の長さの平均値) には話者性が多く含まれているとして聴覚障害者の値にするため，以下の式を用いる．

$$y_i = t_i - \mu_w^{(Dur)} + \mu_x^{(Dur)} \tag{2}$$

$$\mu_w^{(Dur)} = \frac{\sum_{i=1}^{I} \mu_{ti}}{I} \tag{3}$$

$$\mu_x^{(Dur)} = \frac{\sum_{i=1}^{I} \mu_{xi}}{I} \tag{4}$$

式 (2) で，t_i は健常者の音素継続長モデルの i 番目のノードの値である．式 (3)，式 (4) において，I はモデル中のノード数，u_{ti} は健常者モデルの i 番目のノードの平均値であり，u_{xi} は聴覚障害者モデルの i 番目のノードの平均値である．

4　評価実験

明瞭度 (DMOS) と話者性 (一対評価) に関する主観評価実験を行なった．日本語話者 9 人に対してオープンなテキスト 10 文について，録音音声に加えて 4 つの方法で合成音を作成し，ヘッドホンを用いて聴収実験を行った．

- **Original**: 録音音声
- **Conventional**: 従来の DNN 音声合成
- **Prop_Dur**: **Conventional** + 音素継続長修正
- **Prop_F$_0$**: **Conventional** + 基本周波数修正
- **Proposed**: **Prop_Dur** + **Prop_F$_0$**

Conventional では，聴覚障害者の高域のエネルギーが弱い子音や，欠落している音素のうち話者性が現れにくい音素について，スペクトル特徴量のフレームを健常者のフレームと置換して学習を行なっている．今回，置換した音素は「/s, sh, k, t, ts, z, ch/」である．

なお，DMOS は 5 段階評価で「5：劣化が全く認められない，4：劣化が認められるが気にならない，3：劣化がわずかに気になる，2：劣化が気になる，1：劣化が非常に気になる」である．

4.1　実験条件

実験データとして，聴覚障害者 1 名と健常者 1 名の音声を用いた．音声は健常者と聴覚障害者共に 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 の入出力に関して言語特徴量は 395 次元，音響特徴量は WORLD を用いて抽出した 229 次元（メルケプストラム 50 次元，帯域非周期性指標 25 次元，基本周波数 1 次元に二次までの動的特徴量に加えた 228 次元に有声無声パラメータ 1 次元を加えた）を用いた．言語特徴量は [0-1] 正規化，音響特徴量は平均 0 分散 1 に正規化をしている．

4.2　実験結果と考察

Fig. 4 に明瞭度の一対評価の実験結果を示す．図中のエラーバーは 95%信頼区間を示している．Fig. 4 から，**Conventional** が **Original** より低い明瞭度を示していることに対し，**Proposed** が **Original** を上回っている．また，**Proposed** は，**Prop_Dur** や **Prop_F$_0$** より高い明瞭度を示している．これにより，基本周波数修正と音素継続長修正の両方を使用することで，より高い明瞭度を持つ音声が作成できる．

Fig. 5 に話者性の DMOS テストの結果を示す．図中のエラーバーは 95%信頼区間を示している．Fig. 5 から **Conventional** が最も話者性が高く **Original** に近い．また，健常者の音声特徴量を用いると話者性は失われるが，**Proposed** の DMOS スコアは 3.19 であり話者性がわずかに気になるが保たれている結果となった．

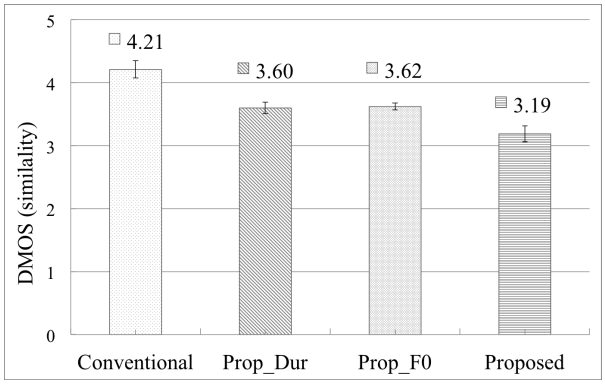
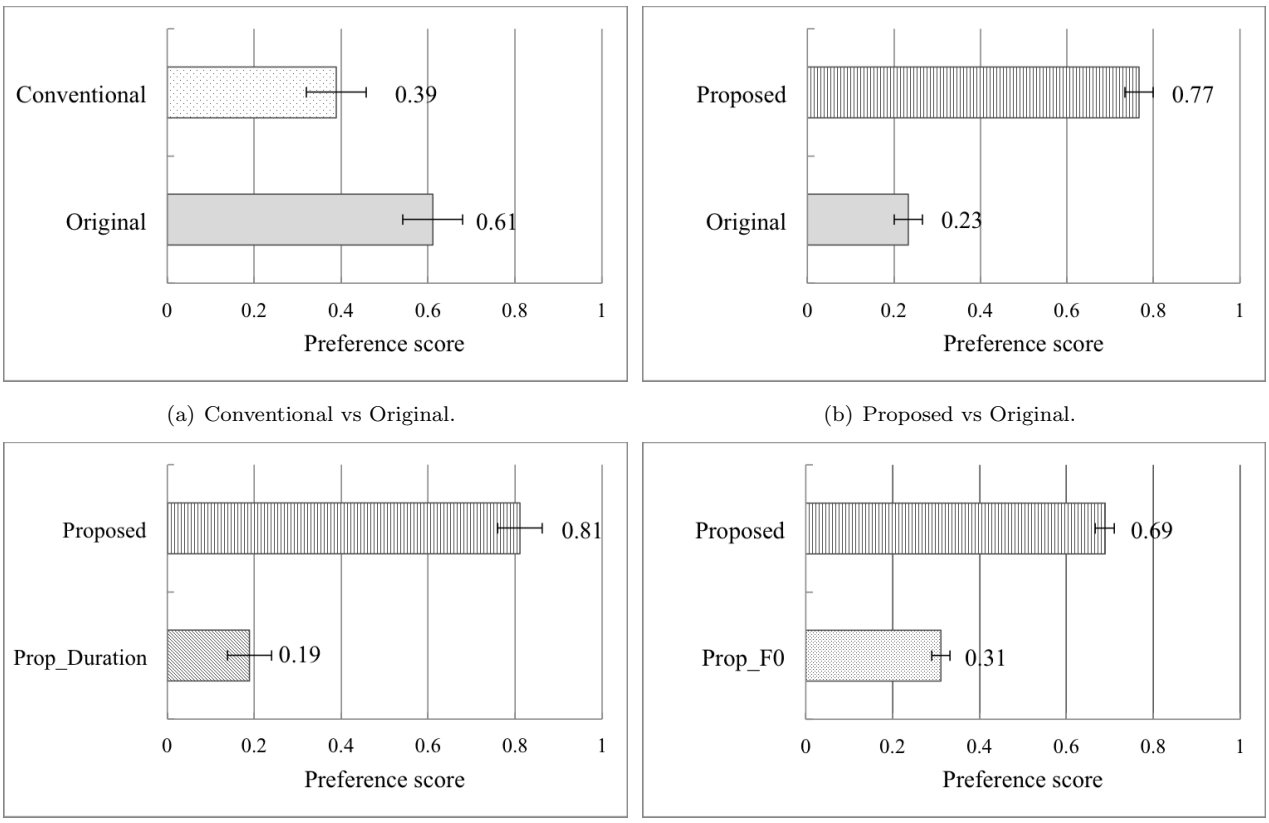


Fig. 5　Speaker similarity to the hearing loss person based on subjective evaluations.

よって，Fig. 4 と Fig. 5 より，提案手法により聴覚障害者の話者性を維持しつつ聞き取りやすい音声が作成されたことがわかる．

5　おわりに

本研究では，Deep Neural Networks を用いた構音障害者の発話を支援する音声合成の手法を提案した．話者性を維持しつつより明瞭度の高い合成音を作成するため DNN のモデルを修正した．主観評価実験により話者性と明瞭度の試験を行った結果，従来手法では録音音声より低い明瞭度を示した一方で，提案手法では録音音声や従来手法よりも高い明瞭度を示した．今後は，話者性を多く含み明瞭度が低い音素をより聞き取りやすくするため，話者性と音韻性を分離し再合成できるような特徴量を検討する．



(a) Conventional vs Original.

(b) Proposed vs Original.

(c) Proposed vs Prop_Dur.

(d) Proposed vs Prop_F_0

Fig. 4　Preference scores for the listening intelligibility based on subjective evaluations.

謝辞　本研究の一部は，JST さきがけ JPMJPR15D2，JSPS 科研費 JP17H01995 の支援を受けたものである．

参考文献

[1] R. Ueda, T. Takiguchi, and Y. Ariki, “Individuality-preserving voice reconstruction for articulation disorders using text-to-speech synthesis,” in *ACM ICMI*, 2015, pp. 343–346.

[2] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in *IEEE ICASSP*, 2013, pp. 7962–7966.

[3] Z. Wu and S. King, “Improving trajectory modelling for DNN-based speech synthesis by using stacked bottleneck features and minimum generation error training,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1255–1265, 2016.

[4] S. Takaki, S. Kim, and J. Yamagishi, “Speaker adaptation of various components in deep neural network based speech synthesis,” in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 153–159.

[5] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: a generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.

[6] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *IEEE ICASSP*, vol. 3, 2000, pp. 1315–1318.

[7] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.