構音障害者を対象としたDNN音声合成に関する言語特徴量の検討*

☆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，脳性麻痺から起こる構音障害を持つ人々を対象として，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．構音障害を持つ人々の発話スタイルは健常者と異なるため，音声を用いた会話には困難を伴う．また，構音障害の種類や程度によっても発話の障害となる原因は異なる．聴覚障害者の場合は耳が聞こえないため，発話の基本周波数やスペクトルが不安定であることが聞き取りを難しくしており，これらを修正する音声合成システム [1] を作成した．一方で，脳性麻痺者の場合は筋肉の不随意運動により，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までの器官に不自由がある．そのため，健常者と脳性麻痺者のコニュニケーションは文字盤や，体の一部を動かすことにより意思を伝える場合がある．そこで，音声合成の使用を試みることも考えられるが，一般的な音声合成システムでは，脳性麻痺者の話者性を維持しつつ聞き取りが容易な音声を作成することが出来ない．

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] 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 [3]，より高い自然性を持つ音声を作成できる WaveNet[4] が開発されている．

本稿では，私たちが以前提案した，構音障害者のスペクトルや基本周波数を修正し，話者性を維持しつつ明瞭度の高い音声を作成するシステム [1] をベースラインとして，脳性麻痺者の音声を作成する．脳性麻痺者の収録音声には，読み上げ時の筋肉の不随意運動により，同一のブレスグループやアクセント句の中に多くの息詰まりやブレークが存在する．よって，テキストを形態素解析して得られる言語特徴量と学習に用いる音声との間にミスマッチが存在する．本稿では学習及び推定に用いる言語特徴量に使用する特徴を検討することで，より明瞭性の高い音声を作成するシステムの実現を目指す．

2 深層学習を用いた音声合成

深層学習を用いた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ニューラルネットワークを用いて学習する．Fig. 1 に深層学習を用いた音声合成の合成時の概要を示す．

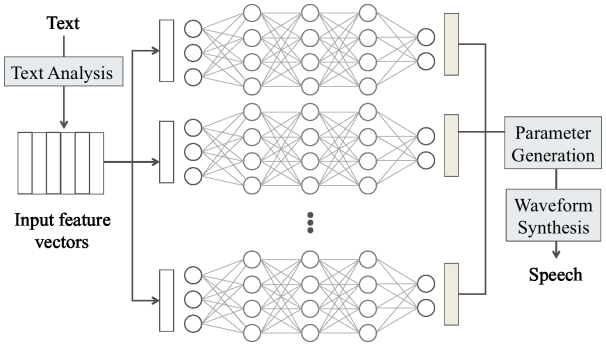


Fig. 1 A flow of speech synthesis using deep neural networks.

言語特徴量には，テキストを形態素解析して得られる { 音素，モーラ，単語，アクセント句，ブレスグループ，文 } に関する特徴が含まれている．深層学習を用いた音声合成では，音響特徴量と言語特徴量が同フレーム数である必要があるため，言語特徴量を引き延ばす必要があるが，この時にフレーム位置に関する特徴を付与することで，時間的変動を考慮することが可能となる．言語特徴量をニューラルネットワーク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，音声合成時には最尤推定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，多層双方向 LSTM を中間層に用いた音響モデルを学習及び合成に用いた．時間的な変動をモデルにより保持することで，より高い自然性と音質を実現することが可能となっている．

3 脳性麻痺者の発話

脳性麻痺者は，筋肉の不随意運動などによって意図した発話が出来ず，健常者と比較して発話が不安定となる場合がある．Fig. 2 に脳性麻痺者の発話として，「日本のエスペラントとして，〜」と発話した音声信

*Linguistic Features of 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)

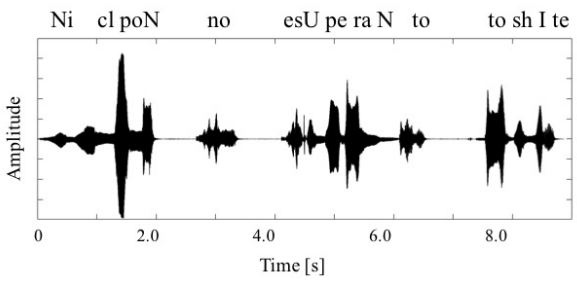
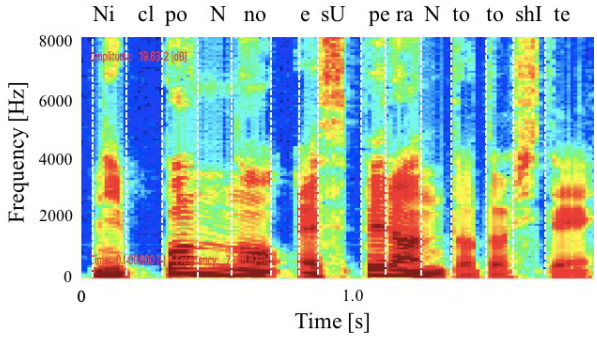


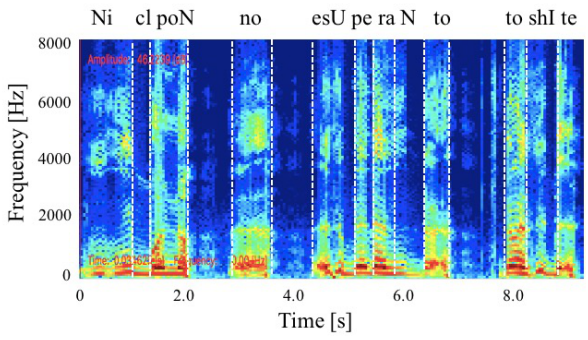
Fig. 2 Sample speech signal of a person with an articulation disorder.

号を示す．Fig. 2 から，「日本・の・エスペラント・として」のように途切れ途切れの発話となっている．これは，発話の区切り位置をあらかじめ決めて発話を試みるが，健常者のように意図した発話が出来ず，区切り位置が不安定となっているためである．また，「日本 (/n/ /i/ /cl/ /p/ /o/ /N/)」の「/p/」子音は発音時に口先の筋肉の制御を必要とする音素であり，Fig. 2 から，他の音素と比較して振幅が大きい．細かい筋肉の制御を必要とする音素の発音では，このように振幅が不安定となる場合がある．

続いて，Fig. 3 に健常者と脳性麻痺者の発話した「日本のエスペラントとして，〜」のスペクトログラムを示す．なお，図中のアライメントは手動で行っている．



(a) a physically unimpaired person.



(b) a person with an articulation disorder.

Fig. 3 Sample spectrograms.

Fig. 3 から，健常者と比較して脳性麻痺者は全帯域のエネルギーが弱く，こもった発話となっている．特に，図中の「/s/ /sh/」のような高域のエネルギーの欠落が大きく，不明瞭な発音となっている．また,「エスペラント」の発音が「エスペランド」となるように，音素の置換や濁音化が起きる場合がある．発話時間を比較すると，健常者の 1.5 秒に対して脳性麻痺者は 9 秒近く必要としており，音素継続長が長くなっている．また，発話の出だしの「日本 (/n/ /i/ /cl/ /p/ /o/ /N/)」の「/n/ /i/」が筋肉の緊張により間延びしていることなど，各音素の音素長比率が不安定であることが聞き取りを難しくしている．

基本周波数について，Table 1 に脳性麻痺者と健常者 3 名の平均と分散を示す．なお，F_0 の導出には音

Table 1 F_0 の平均と分散

	脳性麻痺者	1	2	3
平均	180	137	118	113
分散	3409	1058	929	734

声分析システム WORLD[5] 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，対象とした脳性麻痺者の分散が大きいが，筋肉の緊張により声が裏返るなどの突発的な変動があり，聞き取りが難しい原因となっている．

4 構音障害者を対象とした音声合成

3 節から，脳性麻痺者の音声の聞き取りが難しい原因として，音素の欠落や置換，不安定な振幅と継続長，及び基本周波数の分散の高さが挙げられる．よって，脳性麻痺者の収録された音声のみを使用して音声合成を行うと，合成音も同様に聞き取りが難しいものとなってしまう．そこで本研究では，話者性の近い健常者と脳性麻痺者の両方の音声を学習データとして，話者性は維持しつつより聞き取りやすい合成音を作成した．

4.1 音素継続長修正

音素継続長は言語特徴量を入力，継続長を教師とする DNN を用いて推定を行った．Fig. 2 と Fig. 3 から，音素の間延びなどにより継続長が不安定であるため，発話のリズムとテンポが健常者と異なる場合がある．そこで，健常者の継続長モデルをベースとして，発話のリズム (各音素の長さの比率) に健常者の値を用いて，発話のテンポ (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脳性麻痺者の値にするため線形変換を行うことで音素継続長情報を推定した．

4.2 音響特徴量修正

脳性麻痺者の合成音の基本周波数 F_0 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ニューラルネットワークを用いる．まず，学習時には健常者と脳性麻痺者の 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．脳性麻痺者の DNN は，入力に言語特徴量と聴覚障害者の F_0，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 DNN を構築する．

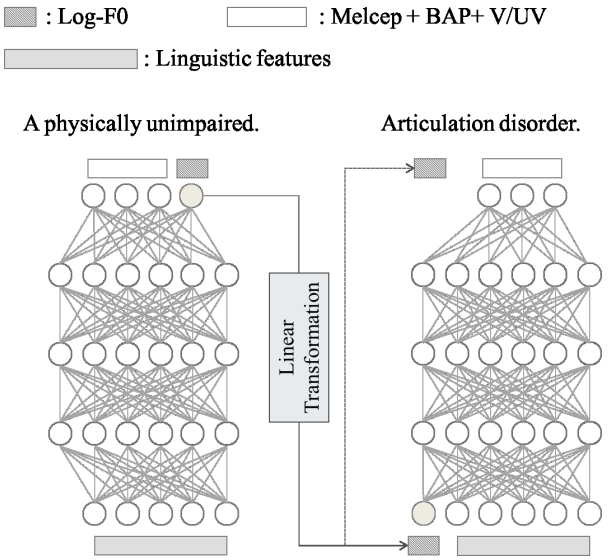


Fig. 4　A flow of the F_0 and spectrum modification method.

Fig. 4 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 系列の平均値を脳性麻痺者の値に修正する．

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

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 はそれぞれ脳性麻痺者の対数数 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．合成時の言語特徴量は，健常者と脳性麻痺者について 4.1 節により推定した継続長情報を用い，健常者と脳性麻痺者の二つの音響特徴量を推定する．

また，明瞭性を保証するために，一部の子音を健常者のものと入れ替えを行う．これは，脳性麻痺者が発音できない，もしくは明瞭性の低い音素が存在する場合があるが，その中で高域のみにエネルギーを有する摩擦音や破裂音のような話者性が現れにくい音素については，健常者の音素のフレームと入れ替えて学習を行うことにより，スペクトル情報を補完する．なお，今回対象とした話者で置換した音素は「/s, sh, k, t, ts/」である．

5　評価実験

Fig. 2 から，脳性麻痺者は「日本・の・エスペラント・として」のように，単語やモーラ毎に区切られた発話を行う場合がある．この場合，一般的な TTS で用いられるアクセント句やショートポーズで区切られたブレスグループが存在しない．TTS の合成時にテキストを形態素解析して得られる単位の中には例としてブレスグループの単位が存在し言語特徴量に入れられるが，ブレスグループの単位で読み上げられた文章が学習データ中に多数存在しない場合には，有効ではないと考えられる．アクセント句や単語単位についても同様であり，言語特徴量がどの単位の特徴まで合成音の明瞭性に有効であるかを評価する．

評価のために録音音声に加えて以下の三つの方法で合成音を作成し，明瞭度 (一対評価)，話者性 (DMOS) と自然性 (MOS) に関する主観評価実験を行った

- **Original**: 録音音声
- **Breath**: ブレスグループまで全ての特徴を含む
- **Accent**: アクセント句までの特徴を含む
- **Word**: 単語までの特徴を含む

各評価にはオープンなテキスト 10 文ずつを用意し，日本人 10 人に対して聴取実験を行った．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．

5.1 実験条件

実験データとして，脳性麻痺者 1 名と健常者 1 名の音声を用いた．音声は健常者と脳性麻痺者共に ATR 音素バランス 380 文学習データ，40 文を開発データとして用いた．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 **Breath** が 402 次元，**Accent** が 390 次元，**Word** が 336 次元である．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 60 次元，帯域非周期性指標 5 次元，基本周波数 1 次元に二次までの動的特徴量を

加えて使用した．言語特徴量は [0-1] 正規化，音響特徴量は平均 0 分散 1 となるように正規化をしている．

5.2 実験結果と考察

Fig. 5 に明瞭度，Fig. 6 に話者性，Fig. 7 に自然性についての実験結果を示す．各図中のエラーバーは 95%信頼区間を示す．

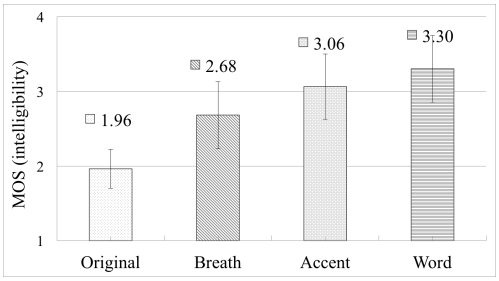


Fig. 5 MOS scores for the listening intelligibility.

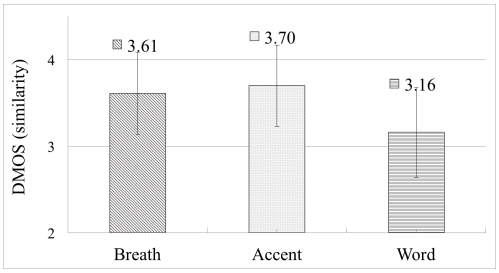


Fig. 6 Speaker similarity to the person with an articulation disorder.

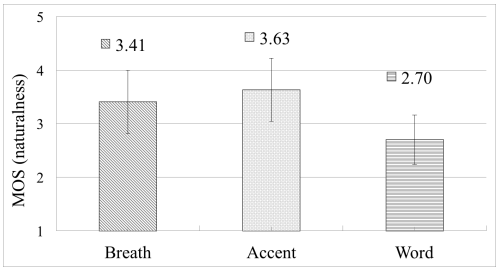


Fig. 7 MOS scores for the naturalness.

Fig. 5 から，**Original** と比較して 3 種類の合成音の明瞭度が上回っており，音響特徴量修正 DNN モデルが有効であることが示された．また，**Breath**, **Accent** と **Word** を比較すると，長い言語単位の特徴が合成音の明瞭度に寄与していない．

Fig. 6 と Fig. 7から話者性と自然性において **Word** が低い値を示した．発話の助詞などが前後の繋がりがなく独立に発音をされる場合があり，明瞭な発音とはなるが自然性の低下の原因と考えられる．よって，**Accent** のようにアクセントフレーズまでの特徴を用いることで，録音音声と比較して明瞭度が高く，話者性と自然性を維持した合成音声が作成できる．

6 おわりに

本研究では，深層学習を用いた脳性麻痺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 音声合成のモデルについてスペクトルや基本周波数を修正するモデルを使用するとともに言語特徴量の検討を行った．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，提案したモデルで作成した音声は話者性を維持しており，録音音声と比較して高い明瞭度を示した．また，学習時にはブレスグループに関する特徴を使わず，アクセント句の特徴を最大の単位として，音声の合成を行うことで話者性を維持しつつ，より聞き取りが容易な音声が作成されることを示した．今後は，話者を増やすとともに，より少ないデータによる構音障害者の音声合成の実現を検討する．

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] T. Kitamura, T. Takiguchi, Y. Ariki, and K. Omori, “Individuality-preserving speech synthesis system for hearing loss using deep neural networks,” *CHAT-17*, pp. 95–99, 2017.

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

[3] 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.

[4] 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.

[5] 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.