音響特徴量補正による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)

1 はじめに

テキスト音声合成とは，入力されたテキストの言語情報から合成音声を生成する技術である．本研究では，アテトーゼ型脳性麻痺から起こる構音障害を持つ人々を対象として，テキスト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が音声を用いた発話を行う際，発話スタイルが健常者と異なるため聞き取りには困難を伴う．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) を用いた音声合成が広く研究されている．音声認識と組み合わせたスマートフォンの対話アプリケーションなどで使用されており，高い自然性を持つ音声が作成できる．近年では，高い音質と自然性を持つ音声を合成できるend-to-end の音声合成システムである Tacotron 2[2] をはじめとし，複数話者の音声を合成できる適応手法 [3] などが開発されている．

本研究では，DNNs を用いた音声合成技術を用いて構音障害者の合成音声を作成する．脳性麻痺者の収録音声は，読み上げ時の筋肉の不随意運動により明瞭度が低く不安定なものとなっている．よって，これらの音声を教師データとして音響モデルを学習し合成音を生成すると，学習データと同様に合成音の明瞭度も低くなる．本稿では，脳性麻痺者と健常者の音響特徴量とを比較し，脳性麻痺者の音響特徴量を健常者のパラメータを用いて修正を行うことにより，明瞭かつ話者性を保持した合成音を作成する．

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，入力としてテキストから言語情報を抽出して得られる言語特徴量と，教師として音声波形を分析して得られる音響特徴量と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．

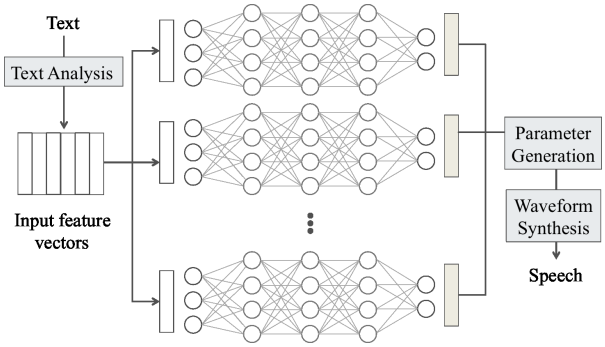


Fig. 1 A flow of speech synthesis using deep neural networks.

本研究では，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，発話中に起きる筋肉の不随意運動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．また，舌や口腔などの筋肉を動かすために健常者と比較して多くの力が必要であり，特定の音素が発音出来ない場合や他の音素に置換される場合がある．

Fig. 2 に脳性麻痺者の発話例として,「日本のエスペラントとして，〜」と発話した音声信号を示す．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,「エスペラント・として」の「として」の出だしの振幅が大きくなっており，健常者と比較して振幅が不安定であることがわ

*Speech Synthesis System Using Deep Neural Networks for Articulation Disorders based on modification of acoustic features. by Tsuyoshi Kitamura (Kobe University), Tetsuya Takiguchi (Kobe University/ JST PRESTO)

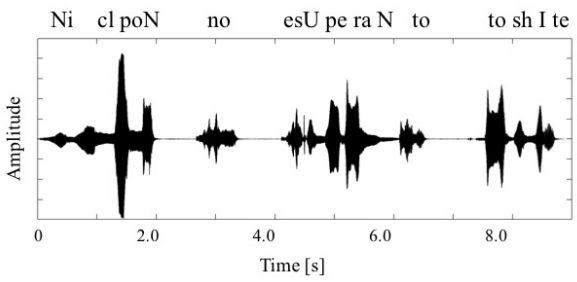
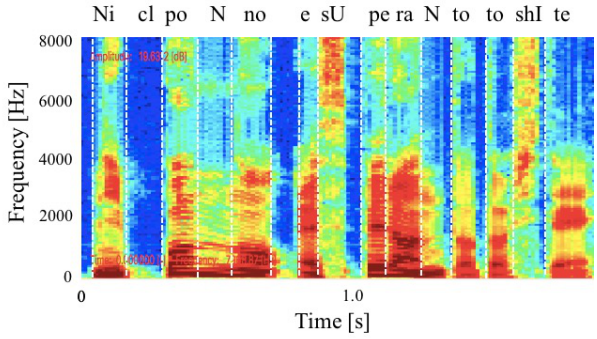
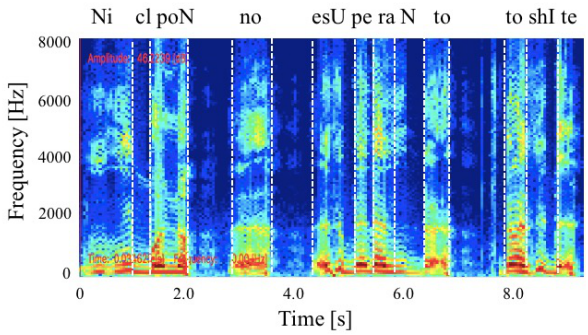


Fig. 2 Sample speech signal of a person with an articulation disorder.



(a) a physically unimpaired person.



(b) a person with an articulation disorder.

Fig. 3 Sample spectrograms.

かる．

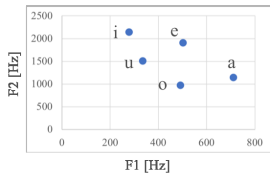
続いて，Fig. 3 に健常者と脳性麻痺者の発話した「日本のエスペラントとして，〜」のスペクトログラムを示す．なお，図中のアライメントは手動で行っている．Fig. 3 から，健常者と比較して脳性麻痺者は，全帯域においてエネルギーが弱く明瞭度が低くこもった発話となっている．母音は第 3〜第 4 フォルマントを多く含む 2000Hz から 4000Hz の帯域が特に弱く，聞こえ度が低くなっている．また，図中の「/s/ /sh/」のような摩擦音など，舌先や唇などの細かい筋肉の制御を必要とする子音のエネルギーの欠落が大きい．また，「エスペラント」の「/t/」音素の発音時，「/d/」のような音となっており音素の置換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/」や「エスペラント」の「ン (/N/)」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 3 に男性の脳性麻痺者 3 名と男性の健常者 3 名の平均と分散を示す．表中

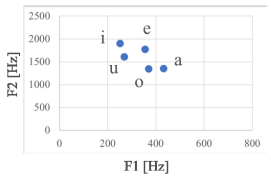
Table 1 F_0 の平均と分散（D：障害者，P：健常者）

	D1	D2	D3	P1	P2	P3
平均	198	195	180	137	118	113
分散	2823	3324	3409	1058	929	734

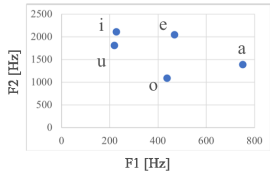
の P1，P2，P3 は健常者を，D1，D2，D3 は脳性麻痺者を表している．なお，F_0 の導出には音声分析システム WORLD[4] を用いており，平均と分散の算出には各話者について 50 文の同一の文章から，無音区間を省いて計算した．Table 3 から，脳性麻痺者の分散がどの話者についても大きいが，筋肉の緊張により声が裏返るなどの突発的な変動があり，聞き取りが難しい原因となっている．



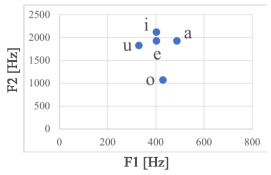
(a) average of physically unimpaired persons.



(b) D1.



(c) D2.



(d) D3.

Fig. 4 F1 and F2 of vowels.

続いて，健常者と脳性麻痺者について，Fig. 4 に母音の第 1 フォルマント F1 と第 2 フォルマント F2 の平均値を示す．母音の音韻は F1 と F2 によっておおよそ識別されるが，F1 は舌の上下位置や開口度，F2 は舌の前後位置によって決定する．Fig. 4 より，脳性麻痺者は健常者と比較して母音の発音時の開口度合いや舌位置の変化が小さく，各母音間が近い F1 と F2 を持っている．そのため，例として D1 や D3 の場合は「/e/」と「/i/」母音の発音が混ざる場合などが多い．また，D2 の各母音の F1 と F2 の点位置は，健常者の空間を狭めたものとなっているため，脳性

麻痺者の母音の発音を識別する訓練をすることで聞き取りができる可能性があるが，例として健常者の「/o/」の空間に脳性麻痺者の「/a/」が位置しているため，単音節明瞭度は低いと推測される．

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，不安定な振幅と継続長，基本周波数の分散の高さ，F1-F2 の空間が狭いことによる母音の曖昧化及び子音の欠落や置換が挙げられる．構音障害者を対象とした音声合成手法として，テキスト音声合成と声質変換技術を組み合わせた手法 [5] が研究されている．本研究ではテキスト音声合成の枠組みで，話者性の近い健常者と脳性麻痺者の両方の音響特徴量を使用することで，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．

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 を用いて推定を行う．Fig. 2 と Fig. 3 から，音素の間延び等により継続長が不安定であり，各音素の長さの比が健常者と大きく異なる場合がある．そこで，健常者の継続長モデルの推定値をベースとして，各音素の長さの比率には健常者の値を用いる．各音素の長さの平均値は話者性を多く含んでいるため障害者の値となるよう線形変換を行うことで継続長を推定した．この時，破裂を含む破裂音と破擦音に対しては線形変換を行っていない．

4.2 音響特徴量修正

音響特徴量の修正は，基本周波数 F_0，パワー及びスペクトルに対して行う．学習時は，独立に脳性麻痺者と脳性麻痺者と声質が近しい健常者について二つ，２節で示した音響モデルを学習する．音声合成時は，テキストから 4.1 節に基づいて継続長を推定，言語特徴量を作成し二つの音響モデルに入力することで，健常者と構音障害者のパラレルな音響特徴量を得る．

F_0 の修正について，脳性麻痺者の F_0 は分散が高く不安定であるため，式 (1) を用いて脳性麻痺者の F_0 系列を推定する．

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

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．式 (1) により，脳性麻痺者の平均声高を保ちつつ健常者のピッチ概形を用いる．

続いて，舌先，歯や唇の筋肉などを使用する発音が難しい子音の明瞭度を向上するため，健常者合成音のスペクトルを用いて障害者スペクトルの修正を行う．修正は式 (2) を用いて行う．

$$\hat{S}^{(ij)} = f_{PU}^{(j)} S_{PU}^{(ij)} + f_{AD}^{(j)} S_{AD}^{(ij)} \tag{2}$$

この時，S_{PU}，S_{AD}，$\hat{S}$，i，j はそれぞれ，健常者スペクトル，障害者スペクトル，修正後スペクトル，フレームのインデックス，周波数次元のインデックスを示している．重み関数 f_{PU}，f_{AD} は以下のように定義される．

$$f_{PU}^{(j)} = \frac{1}{1 + e^{(-j+c)}} \tag{3}$$

$$f_{AD}^{(j)} = \frac{1}{1 + e^{(j-c)}} \tag{4}$$

この時，f_{PU} が健常者スペクトルに対する重み関数，f_{AD} は障害者に対する重み関数，c は制御変数をそれぞれ表している．式 (2) により，c により決定される閾値以上の周波数を持つ帯域は健常者，閾値以下の周波数帯域は障害者のスペクトルが使用される．周波数の閾値を制御する変数 c は式 (5) により決定する．

$$c = \frac{th}{fs} \times D \tag{5}$$

式 (5) において，th は周波数の閾値，fs はサンプリング周波数，D はスペクトルの次元数を表している．th の周波数の閾値については，無声破裂音，無声摩擦音，無声破擦音については話者性が低い子音であるので閾値を 0Hz とすることで，障害者の濁音化した音声に含まれる低域のボイスバーを除去しつつ高域を付与する．有声破裂音，有声摩擦音，有声破擦音は 3500Hz を閾値とすることで，障害者の低域のボイスバーを保ちつつ高域のエネルギーを付与することで話者性を保ちつつ子音の明瞭度を向上させる．

3 節から，脳性麻痺者は F1-F2 空間が狭く母音が曖昧になる場合がある．母音の明瞭度を向上させるために，話者ごとに最適化が必要であるが，健常者平均の F1 もしくは F2 の周波数と大きく乖離を持つ母音について，健常者の平均 F1-F2 の周波数に位置する障害者のエネルギーに対して対数スペクトル上でフィルタを用いることで，スペクトル中のピークを強調する．本研究では強調フィルタとして，式 (6) を用いた．

$$f_k = a \times \exp(-\frac{(k-c)^2}{2\sigma^2}) + 1 \tag{6}$$

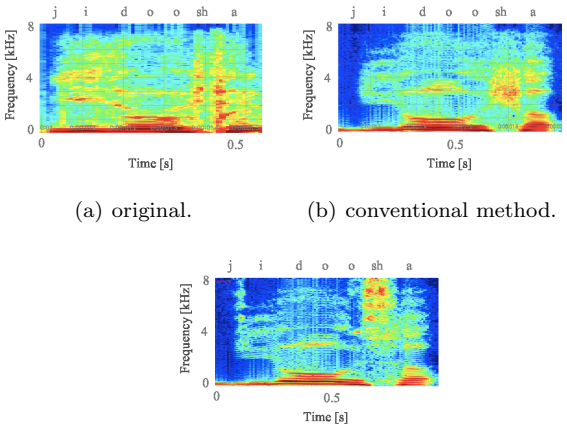
f_k は障害者対数スペクトルの次元 k に対するフィルタ重み，c と σ^2 は強調を行う F1-F2 のいずれかにおける健常者フォルマント周波数に対応するインデックスの平均と分散を示している．a は対数スペクトルに対する強調の度合いを制御する変数であり，本研究では最適な数値を探索により求めた．また，本研究では障害者の不安定な振幅を修正するために，健常者の音響モデルによる推定値を用いた．

5　評価実験

5.1　実験条件

実験データとして，脳性麻痺者 1 名 (3 節の D3) 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．脳性麻痺者の音声のアライメントは HMM を用いた強制アライメント後に，エラー箇所を修正することで求めた．音響モデルの入出力に関して，言語特徴量はアクセント句を最大の単位とした特徴ベクトル 390 次元を，音響特徴量は WORLD を用いて抽出した 199 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 2 次までの動的特徴量に有声無声パラメータ 1 次元) を使用した．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．D3 の母音は，第 1 フォルマントの空間が狭いので,「/i/, /a/」母音の第 1 フォルマントが健常者と大きく乖離しており明瞭度が低い．そのため,「/i/, /a/」のフレームについてそれぞれ，280Hz と 710Hz の周波数を中心に強調フィルタを適用した．

5.2　実験結果と考察



(a) original.　(b) conventional method.

(c) proposed method.

Fig. 5　Sample spectrograms.

Fig. 5 中の conventional は，音素継続長のみを修正した合成音である．Fig. 5 より，録音音声及び従来の音声合成手法では出だしの「/j/」について，ボイスバーは存在するが高域の破裂や摩擦の成分が失われているのに対し，提案手法では高域が付与されている．また，障害者の録音音声について,「/sh/」などの子音は健常者と同様の発音ができず別の発音手法を用いているため，エネルギーが中高域に現れ低域にボイスバーも出ているが，提案手法では高域のみが存在している．しかし，各音素に対してフレーム毎に音響特徴量を修正したため，フォルマント遷移などが失われており自然性の低下が考えられる．

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．健常者と脳性麻痺者の音響特徴量を比較することで，F1-F2 の空間が狭く母音の明瞭性が低いことを示した．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，基本周波数，音素継続長やパワーなどの音響特徴量を健常者の値を用いることで修正した．また，母音や子音の明瞭性を向上するためフィルタを用いて強調を行った．今後は，より自然性の高い音声を生成するため，フォルマント遷移を加味した子音や母音の強調を行う．

謝辞　本研究の一部は，JSPS 科研費 JP17J04380, JST さきがけ JPMJPR15D2 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] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” in *IEEE ICASSP*, 2018, pp. 4779–4783.

[3] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno *et al.*, “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” *arXiv preprint arXiv:1806.04558*, 2018.

[4] 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.

[5] R. Nanzaka and T. Takiguchi, “Hybrid text-to-speech for articulation disorders with a small amount of non-parallel data.” in *APSIPA*, 2018.