話速補正に基づく話者性を維持した構音障害者のための音声合成システム *

☆上田怜奈 (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支援するための音声合成法を提案する. 構音障害者にとって健常者との会話は困難を伴うものである. 障害者支援のための研究において，Veaux *et al.* [1] は筋萎縮性側索硬化症（ALS）患者のための話者性を維持した音声再構築を試みた．また，山岸ら [2] は様々な人々の音声を集めデータベースとし，それを用いた ALS 患者のための TTS システムを構築した．構音障害者のコミュニケーションの障害となりうる要因としてはピッチ，スペクトルなどの問題が挙げられる．このような問題に対して，これまでの我々の研究では，彼らのピッチ，スペクトルに対し健常者の音声を用いて変換，修正を行う事でより聞き取りやすい音声を実現した [3][4]．これまでの研究ではピッチやスペクトルの修正は行ってきたものの，話速に関する修正は行ってこなかった．本研究では，健常者のラベリング情報を用いて構音障害者の間延びした話速を修正する．実験では提案法が構音障害者の話者性を維持しつつより聞き取りやすい合成音を実現していることを示す.

2　構音障害者のための HMM 音声合成

構音障害者の音声は収録した段階で不安定な音声となっているため，構音障害者の音声から得られた音声特徴でパラメータ学習をすると得られる合成音は聞き取りずらいものになってしまう．そこで本研究では，話者性の近い健常者と構音障害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．Fig. 1 は提案手法の概要である．提案手法において，構音障害者と健常者の両方を学習データとして使用する．初めに，STRAIGHT[5] を用いて二人の話者から 3 つの音声パラメータ (F0 概形，スペクトラム包絡，非周期成分 (AP)) を抽出する．特徴量を抽出したのち，健常者の F0 系列を修正する（2.1 節）．音素継続長モデルについては構音障害者，健常者それぞれのコンテキスト依存ラベルからそれぞれのモデルを作成したのち修正を行い，修正後音素継続長モデルを得る (2.2 節)．その後，修正後音素継続長モデルから生成したコンテキスト依存ラベル系列と学習した HMM に基づいて，スペクトラム，F0，AP パラメータが生成される．F0 パラメータは修正した F0 モデルから生成，スペクトル，AP パラメータは構音障害者のモデルから生成する．最後に，パラメータ系列を合成フィルタにかけることによって合成音が生成される．2.1 節，2.2 節では F0 と音素継続長に対する処理の詳細を記述する.

2.1　F0 系列の修正

構音障害者の F0 系列はしばしば不安定なものであるので，本研究の F0 の修正法では，健常者の F0 系列を基本として F0 モデルを学習する．F0 系列に構音障害者の話者性を付与するため，F0 系列を構音障害者の特徴へと変換する．F0 モデルはこの変換後の F0 系列を用いて学習するので，構音障害者の話者性が含まれていることになる，F0 系列の変換には Eq. (1) のような線形変換を利用する.

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

Eq. (1) において，w_t は健常者の t フレーム目の対数 F0，μ_w，σ_w は健常者の F0 系列の平均・分散，μ_x，σ_x は構音障害者の対数 F0 系列の平均・分散をそれぞ

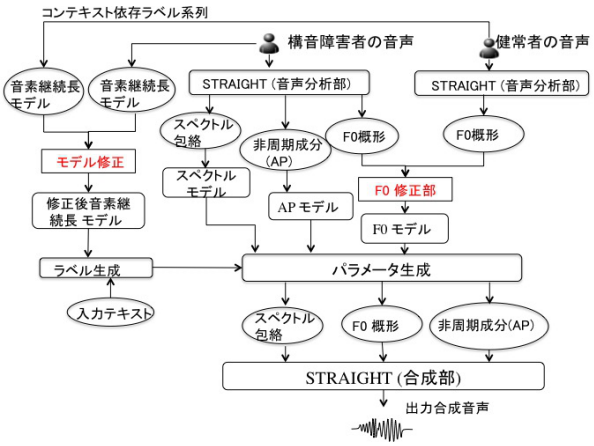
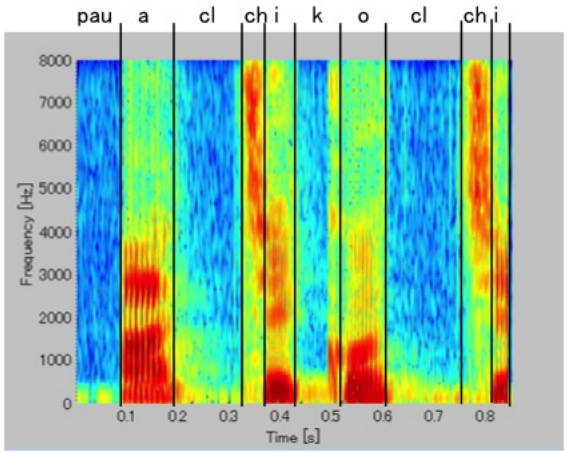
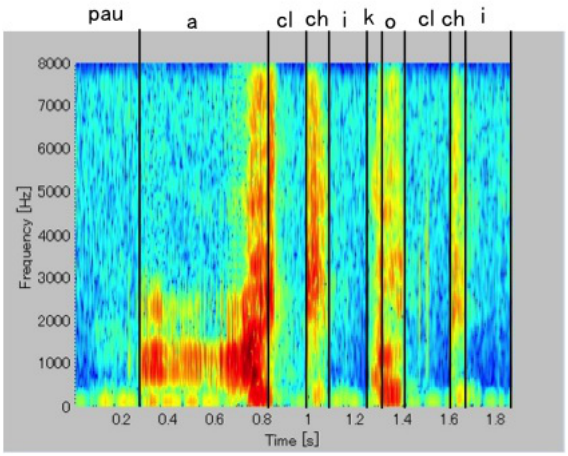


Fig. 1: 構音障害者のための HMM 音声合成手法の概要

*Individuality-Preserving Sound Synthesis System for Articulation Disorders besed on Duration Modification, by Reina Ueda（Kobe University），Tetsuya Takiguchi（Kobe University/JST PRESTO），Yasuo Ariki（Kobe University）



(a) 健常者



(b) 構音障害者

Fig. 2: 元音声スペクトルの一例// pau a cl ch i k o cl ch i

れ表している．

2.2 音素継続長の修正

Fig. 2 は健常者と構音障害者の元音声の“ あっちこっち ”と発声しているスペクトログラムである．なお，図のアライメントは手動で行っている．Fig. 2 において，構音障害者の発話時間は健常者と比較して２倍以上の時間を要している．また，２つめの音素である“ a ”を見比べると構音障害者のスペクトルは間延びしていることが分かる．これらの現象は障害によって筋肉の緊張が起こり意図した発話ができないことによって引き起される．これにより発話の間延びや音素長のばらつきが発生し学習音声のアライメントにもずれが起こってしまう．

Table 1 は構音障害者，健常者の学習データからそれぞれ音素継続長モデルを作成し，母音・子音・無音区間ごとに分布の平均，分散を算出したものである．各話

Type	母音	子音	無音区間
平均	4.54	4.91	6.25
分散	48.73	60.76	137.67

(a) 構音障害者

Type	母音	子音	無音区間
平均	3.29	3.65	4.89
分散	11.26	14.75	58.26

(b) 健常者

Table 1: 学習後音素継続長モデルの平均，分散

者の数値を見比べると，平均，分散ともにどの場合でも構音障害者の数値が高くなっていることが分かる．特に，子音の平均値や母音，子音，無音区間の分散値が高くなっていることが合成音の聞き取りにくさに影響していると考えられる．そこで，本研究では音素継続長モデルを修正し，話者性は維持しつつより聞き取りやすくなる音声を作成する．提案法では，健常者の音素継続長モデルをベースとして修正を行う (Fig. 3)．そして，健常者のモデル中の母音の平均値に対して修正を行う．修正はノードごとに以下のように行う．

$$\hat{y}_i = y_i - \mu_y + \mu_z \tag{2}$$

$$\mu_y = \frac{\sum_{i=1}^{I} \mu_{yi}}{I} \tag{3}$$

$$\mu_z = \frac{\sum_{i=1}^{I} \mu_{zi}}{I} \tag{4}$$

Eq. (2) において，y_i は健常者音素継続長モデル中の i 番目のノードの平均値，μ_y，μ_z は Eq. (3)，Eq. (4) のようにして求められる．Eq. (3)，Eq. (4) において，I はモデル内の母音の全ノード数，μ_{yi} は健常者モデルの i 番目の母音ノードの平均値，μ_{zi} は構音障害者モデルの i 番目の母音ノードの平均値をそれぞれ表している．

3 評価実験

3.1 実験条件

学習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．健常者音声には ATR データベースの音声を用いた．本実験では Table 2 のような，３つの条件のもとで３種類の合成音をそれぞれ ATR データベース中の１０文を基に作成した．それぞれの合成音の作成にあたり，学習データとして健常者音声は ATR データベース５０３文，障害者音声は収録し

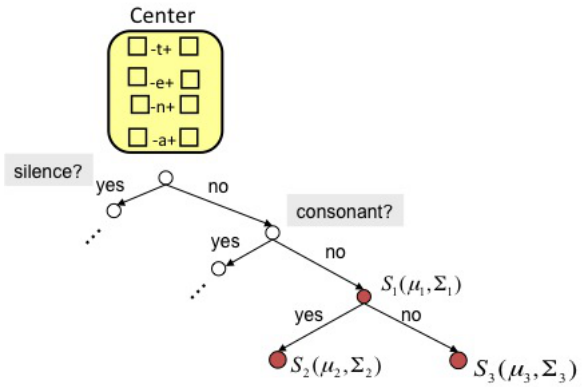


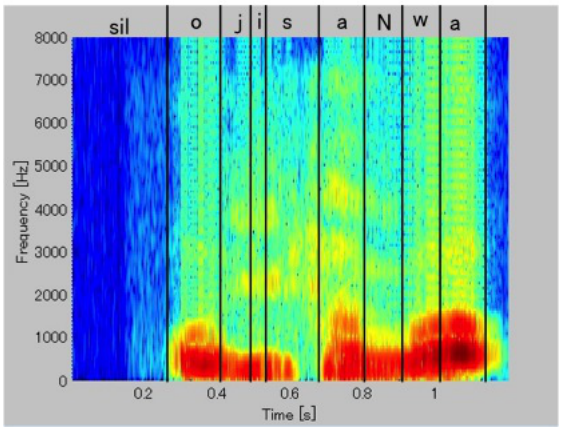
Fig. 3: 健常者音素継続長モデルの修正

Table 2: 実験で比較した合成音の生成条件

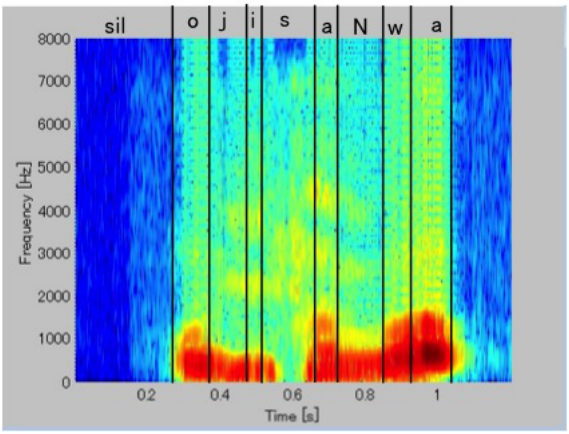
Type	Duration Model	F0 Model	AP Model	Spectral Model
Prop	Modification	Modification	AD	AD
Ref1	PU	Modification	AD	AD
Ref2	AD	Modification	AD	AD

（**Prop**: 提案手法，**AD**: 構音障害者，**PU**: 健常者）

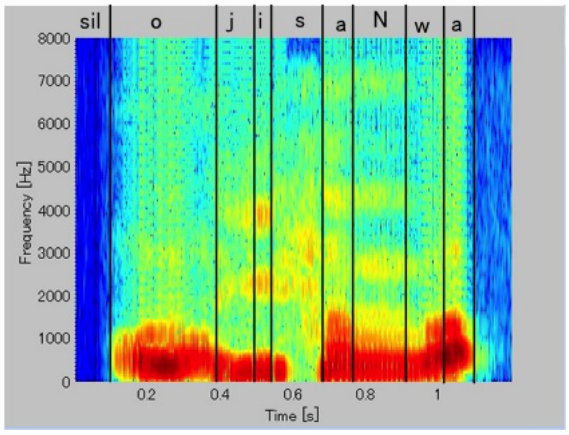
た同じデータベースの４２９文を使用した．特徴量についてはサンプリング周波数は１６ kHz，フレームシフトは５ ms で音声特徴量は STRAIGHT を用いて抽出し，スペクトルパラメータ系列には，２５次元のメルケプストラムとその Δ，ΔΔ，励起パラメータには対数 F0，５周波数帯域の非周期性指標とその Δ，ΔΔ を使用，学習，合成には５状態のコンテキスト依存 HMM を使用した [6]．提案法の有用性を示すため，本研究では話者性と聞き取りやすさの２つの観点からの実験を試みた．９人の日本人に対して聴取実験を行った．話者性に関する実験には本研究では MOS(Mean Opinion Score) テストを実施した．このテストではそれぞれの音に対して５段階で評価をした．(5:非常に似ている, 4: とても似ている, 3:まあまあ似ている, 2: 似ていない, 1: 全く似ていない）聞き取りやすさの評価には一対比較法を行い２つの音声のうちより聞き取りやすい方を選んで評価した．



(a) Prop



(b) Ref1



(c) Ref2

Fig. 4: 合成スペクトルの一例// sil o j i s a N w a

3.2 実験結果

Fig. 4 は３種類の条件を基に作成した合成音のスペクトルである．テキストは全て同じ発話である．Fig. 4c は構音障害者の音素継続長モデルで作成したスペクトルである．発話は全体として長くなっており，特に発話中の “o”が間延びしていることが分かる．Fig. 4b は健常者の音素継続長モデルで作成したスペクトルである．Ref1 と Ref2 を比較すると Ref1 の方が発話時間が短く音素の間延びも見られないことが分かる．Fig. 4a は提案法の音素継続長モデルで作成したスペクトルである．音素長のバランスが整えられ，発話中の “o”も間延びしていないことが分かる．これは，音素継続長モデル作成の際に分散をすべて健常者のものを使用したことが原因と考えられる．

また F0 修正について，Fig. 5 は Ref3 と Prop の合成音のピッチを表示したものである．Ref3 作成の際

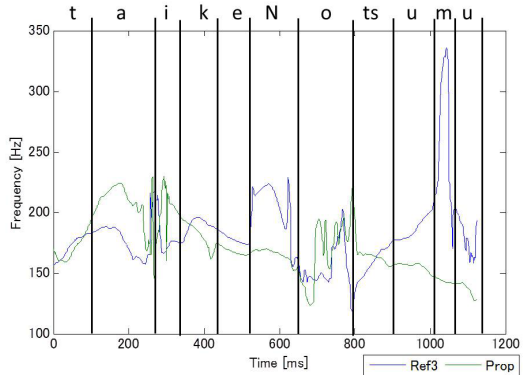
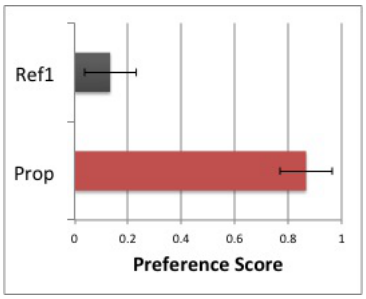
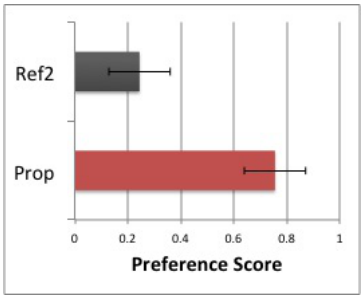


Fig. 5: F0 修正前と修正後の比較 //t a i k e N o ts u m u



(a) 話者性に関する比較



(b) 聞き取りやすさに関する比較

Fig. 6: 実験結果

は F0 修正を行わずに合成した．どちらも提案法の音素継続長モデルを使用し，フレームは一致している．Fig. 5 では日本語で “体験を積む”と発話している．Ref3 において “m”の部分のピッチが不自然に上がっているのに対して，Prop では正しいイントネーションに修正されていることがわかる．

Fig. 6a は話者性に関する実験結果である．Ref1 と Prop を比較したところ Prop がより良い値となり，提案法が健常者の音素継続長モデルを使用した Ref1 よりもより話者性を維持出来ているということが確認できる．Fig. 6b は聞き取りやすさに関する実験結果である．Ref2 と Prop を比較したところ，こちらでも Prop が良い値となった．これは提案法が構音障害者の音素継続長モデルをそのまま使用した Ref2 よりもより聞き取りやすい合成音を実現出来ていると言える．Fig. 6 の結果より，提案法によって障害者の話者性を維持しつつ聞き取りやすさは向上した合成音声を作成出来ていることが分かった．

4 おわりに

本研究では話速補正に基づく構音障害者のための音声合成手法を提案した．実験を通して提案法が補正前の合成音と比較して話者性を維持し聞き取りやすい音声を実現出来ることが示された．今後はスペクトルモデルに対しても音素ごとに修正を行うことを検討したい．

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] C. Veaux *et al.*, “Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders,” in *Proc. of Interspeech*, 2012.

[2] J. Yamagishi *et al.*, “Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction,” Acoustical Science and Technology, vol. 33, no. 1, pp. 1–5, 2012.

[3] R. Ueda *et al.*, “Individuality-preserving spectrum modification for articulation disorders using phone selective synthesis,” in *6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2015, pp. 118–123.

[4] R. Ueda *et al.*, “Individuality-preserving voice reconstruction for articulation disorders using text-to-speech synthesis,” in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 343–346.

[5] H. Kawahara *et al.*, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” Speech communication, vol. 27, pp. 187–207, 1999.

[6] T. Yoshimura *et al.*, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” in *Proc. of Eurospeech*, 1999, pp. 2347–2350.