音素選択型スペクトル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, 滝口哲也, 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] は様々は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，本研究では彼らのコミュニケーションの手助けとなるような隠れマルコフモデル (HMM) を用いた音声合成システムを提案する．提案法においては，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する．これまでの我々の研究では他手法との比較は行っていなかった [3]．本研究の実験では話者適応を用いた合成音との比較も行い，提案手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[4] 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）．学習部・合成部両方において，それぞれのパラメータに対して別々の処理を行う．音素継続長モデルについては構音障害者のコンテキスト依存ラベル系列のみを用いて学習する．合成の際はこのように生成した音素継続長モデルと入力テキストに基づいて，コンテキスト依存ラベル系列が生成される．その後，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，AP パラメータは構音障害者の AP モデルから生成する．スペクトルパラメータに関しては構音障害者，健常者のそれぞれのスペクトルモデルからそれぞれ生成する．スペクトルパラメータを生成した後，障害者のスペクトルパラメータを健常者のスペクトルパラメータを用いて修正する（2.2 節）．パラメータ系列 (スペクトル，F0，AP 成分) はすべて STRAIGHT で扱うことのできる形式に変換される．最後に，STRAIGHT によって最終的な合成音が生成される．2.1 節，2.2 節では F0 とスペクトルパラメータ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

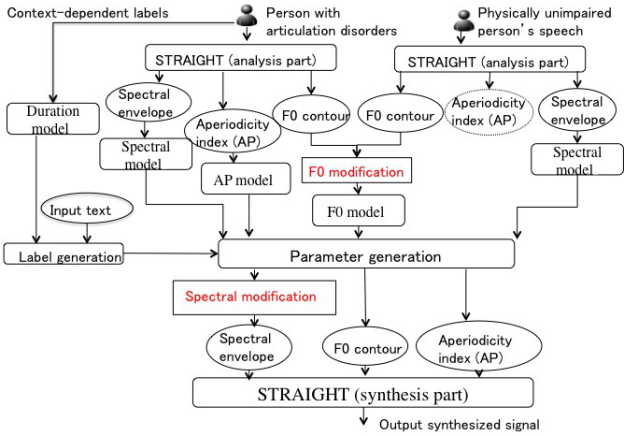
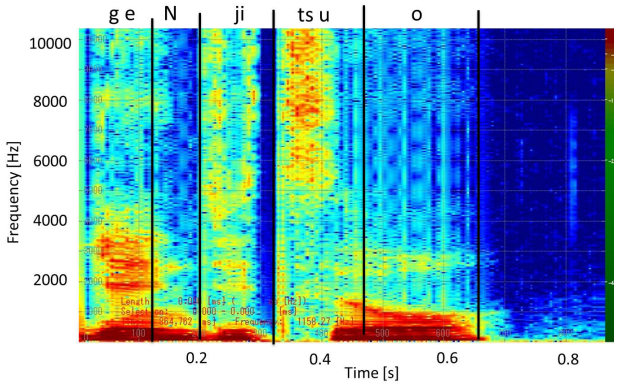
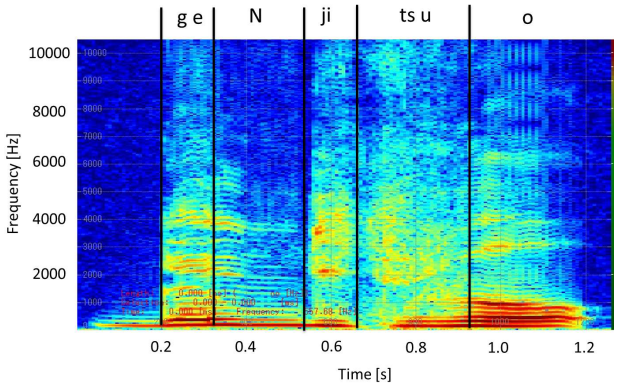


Fig. 1: 構音障害者のための HMM 音声合成手法の概要

*Individuality-Preserving Spectrum Modification for Articulation Disorders Using Phone Selective Synthesis



(a) 健常者



(b) 構音障害者

Fig. 2: 元音声スペクトルの一例// g e N j i ts u o

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．

$$\hat{x}_t = \frac{\sigma_y}{\sigma_x}(x_t - \mu_x) + \mu_y \tag{1}$$

Eq. (1) において，x_t は健常者の t フレーム目の対数 F0，μ_x，σ_x は健常者の F0 系列の平均・分散，μ_y，σ_y は構音障害者の対数 F0 系列の平均・分散をそれぞれ表している．

2.2 スペクトラム系列の修正

Fig. 2 は健常者と構音障害者の元音声の“現実を”と発声しているスペクトログラムである．Fig. 2 にあるように，構音障害者のスペクトルの高周波成分は健常者のものと比べて弱くなっている．これは構音障害者の発声の子音成分が弱くなっておりそのことが聞き取りにくさの原因となっていることを示している．そこで Fig. 1 のようにテキストが入力された後，それぞれの話者のスペクトルモデルからスペクトルパラメータを生成する．そして高周波成分を健常者のスペクトルパラメータで補完し，低周波域は構音障害者のスペクトルパラメータを使用し話者性を維持しつつより聞き取り易くなるように修正を行う．このような修正はすべての音素に対して行うのではなく，摩擦音，破擦音等高周波成分にパワーを持つ音素 (sh/s/z/ch/ts/j) に対してのみ行い，その他の音素に対しては修正を行わず構音障害者のスペクトルパラメータを使用する．この修正は以下の式で実現される．

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

このとき，S_{PU}，S_{AD}，$\hat{S}$，i，j はそれぞれ健常者スペクトル（Physically Unimpaired），構音障害者スペクトル（Articulation Disorder），修正後スペクトル，フレームのインデックス，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．

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

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

このとき，f_{PU} は健常者スペクトルに対する重み関数，f_{AD} は構音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．Eq. (2) を用いることにより，高周波領域では健常者のスペクトル成分によって補完され，より子音部分が明瞭に聞こえるようにし，低周波領域では構音障害者のスペクトル成分を保持することにより話者性を保つということを実現する．周波数の閾値を制御する変数 c は Eq. (5) によって閾値が 4000Hz になるように設定する．

$$c = \frac{4000}{f_s} \times D \tag{5}$$

Eq. (5) において，f_s 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．

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．健常者のデータは出来る限り構音障害者の話者性に近い人を選ぶことが望ましい．そこで本研究では，ATR データベースセット B の話者 10 名と構音障害者間の話者性の類似度を求めるためにスペクトル，メルケプストラム，ケプストラムの話者間の距離をそれぞれ算出した．Table 1 はその結果である．Table 1 より，使用する構音障害者には MTK が最も類似していることが分かり，以下の実験では健常

Table 1: 健常者と構音障害者の類似度

	SpD	MelCD	CepD
FKN	3.72.E-03	1.42	16.72
FKS	2.87.E-03	1.14	13.01
FTK	2.97.E-03	1.01	12.13
FYM	3.60.E-03	1.22	14.39
MMY	3.02.E-03	1.11	13.06
MTK	**2.79.E-03**	**0.97**	**11.64**
MHO	3.62.E-03	1.17	13.85
MHT	3.79.E-03	1.32	15.81
MSH	3.46.E-03	1.14	13.70
MYI	3.65.E-03	1.22	14.52

Table 2: 実験で比較した合成音の生成条件

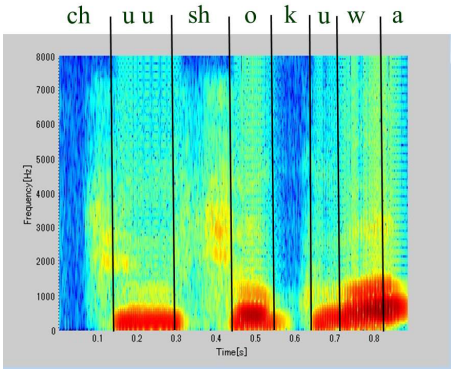
Type	Duration Model	F0 Model	AP Model	Spectral Model
ADM	AD	AD	AD	AD
Ref	AD	Modification	AD	AD
Prop	AD	Modification	AD	Modification
PUM	PU	PU	PU	PU
ADPT	Adapt PUM to the target characteristic			

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

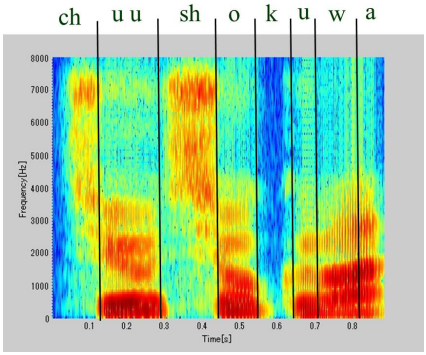
者音声には MTK の音声データを採用した．本実験では Table 2 のような，5 つの条件のもとで 5 種類の合成音をそれぞれ ATR データベース中の 10 文を基に作成した．ADM，Ref，Prop，PUM の合成音の作成においては健常者音声は ATR データベース 503 文，障害者音声は収録した同じデータベースの 429 文を使用した．モデル適応（ADPT）においては，学習データとして健常者音声 503 文，適応データとして構音障害者の音声 50 文を使用した．本研究では適応法には CSMAPLR[5] を用いた．特徴量についてはサンプリング周波数は 48kHz，フレームシフトは 5ms 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，50 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易さの 2 つの観点からの実験を試みた．10 人の日本人に対してヘッドホンで聴取実験を行った．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．

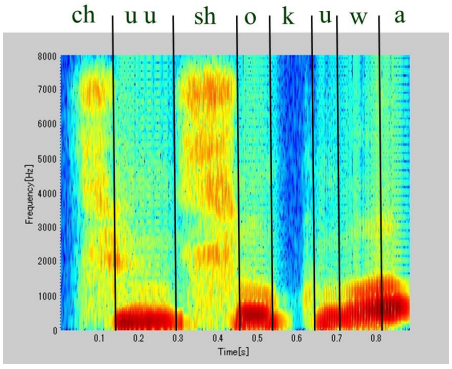
3.2 実験結果



(a) ADM スペクトル



(b) PUM スペクトル



(c) 修正後スペクトル

Fig. 3: 合成後スペクトルの一例
//ch u u sh o k u w a

Fig. 3a は構音障害者スペクトルモデルから生成したスペクトルであり，Fig. 3b は健常者スペクトルモデルから生成したスペクトルである．Fig. 3a の高周波成分は，Fig. 3b と比較して弱くなっており，これが子音の聞き取りにくさの原因となっている．Fig. 3c は Eq. (2) による補正後のスペクトルである．Fig. 3c よ

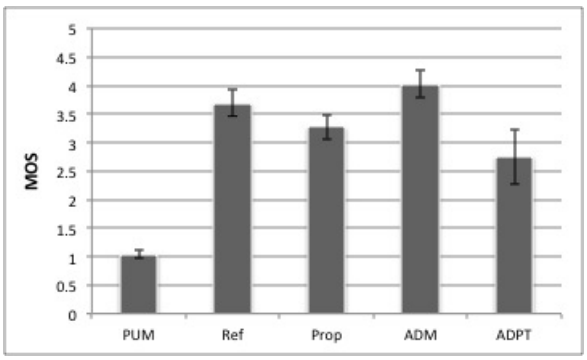
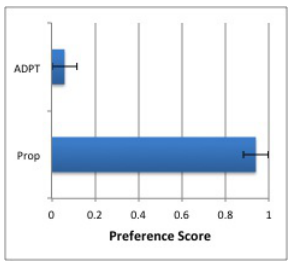
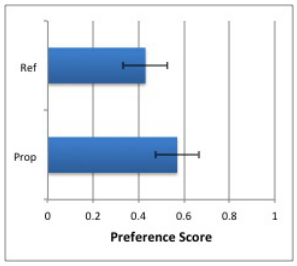


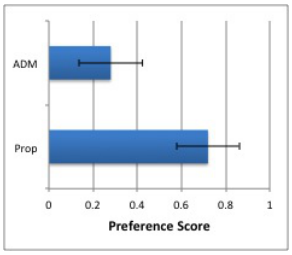
Fig. 4: 構音障害者の話者性に関する実験結果



(a) ADPT



(b) Ref



(c) ADM

Fig. 5: 聞き取り易さに関する比較

り，ADM の低周波成分は保ちつつ高周波成分を PUM によって補完できていることがわかる．Fig. 4 は話者性に関する実験結果である．ここでは ADM がもっとも良い値であり，次いで Ref，Prop という結果であった．有意差検定により，提案手法は既存の適応法に対しては話者性において優位であることがわかった．また，提案手法は ADM や Ref と比較して低い結果となったことから健常者特徴が増えるとその分話者性が下がってしまうことがわかった．Fig. 5 は聞き取り易さに関する実験結果である．Fig. 5 より，Prop が ADM，ADPT よりも聞き取りやすい音声であることが示された．しかし，Ref に対しては優位な結果とはならなかった．原因としては，今回実験に用いた合成音声 10 文のなかでスペクトル修正に該当する音素が全体の約 13% のみであり，聞き取りやすさに影響しづらかったと考えられる．

4 おわりに

本研究では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既存の話者適応手法よりも話者性を維持し聞き取りやすい合成音を実現出来ることが示された．しかしながら，本研究ではスペクトル修正が局所的であった為スペクトルの修正が聞き取りやすさ向上にそれほど寄与しなかった．今後は今回修正した音素に加えて低周波域にパワーを持つ音素の修正方法も検討していきたい．

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] 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.

[5] J. Yamagishi *et al.*, “Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained smaplr adaptation algorithm,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 17, no. 1, pp. 66–83, 2009.

[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.