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，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いる．この障害は以下のタイプに分けることができる．1) 痙攣性, 2) アテトーゼ, 3) 失調性, 4) 弛緩性, 5) 硬直性，そしてこの混合型が存在する [1]．アテトーゼ型脳性麻痺から起こる構音障害者の人々の場合，彼らの動きはしばしば普通と比べて不安定なものとなる．そのことによって，彼らの発声（特に子音）はこの障害のため，しばしば不安定で不明瞭なものとなることから，このような構音障害者のコミュニケーションの手助けとなるような音声システムの構築が急がれている．HMM音声合成システム [2] はテキスト合成 (TTS) システムの一種であり，テキストを入力として音声を出力するものである．構音障害者の発話の手助けとしてTTS システムは非常に便利なものであると考えられる．障害者支援のための研究領域において，Khan *et al.* [3] は喉頭摘出後の患者のための TTS システムを提案し，Veaux *et al.* [4] は HMM 音声合成を用いて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 [5] 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．Creer *et al.* [6] は健常者の人々の音声から作成した平均声モデルに対して複数の構音障害者をそれぞれ適応させることで聞き取り易さの改善を試みた。

本研究では，構音障害者のための HMM 音声合成システムを提案する．彼らの発話はしばしば不安定なものであるため，収録した音声やそれを基に作成したTTS システムの合成音声は聞き取りにくさの原因となる様々な問題を孕んでいる．このため，彼らの話者性は維持しつつより聞き取り易い合成音を作り出せる TTS システムを構築する必要がある．これを実現するため，本研究では学習データとして構音障害者と健常者両方の音声を使用する．また発話長に関して，構音障害者の発話長は健常者と比較して間延びしたものとなっているため，HMM 音声合成中の音素継続長モデルに関しては健常者のラベリングデータから学習したものを使用する．それに加えて，彼らの抑揚 (F0 概形) は健常者と比べて不安定なものとなっているため，F0 モデル構築の際の学習データは健常者のF0 系列を構音障害者の特徴へ線形変換したものを使用する．構音障害者のスペクトルにおいては，子音成分が弱く不明瞭なものが多くそのことが聞き取りにくさの原因となっていることが考えられる．このような子音の問題を解決するため，本研究では彼らのスペクトルの子音部と母音部に対して別々の処理を行う．子音部に対しては，本研究では健常者のスペクトルモデルから生成したスペクトル系列を主に使用する．また，母音部に対しては，話者性を維持するために構音障害者のスペクトルモデルから生成したスペクトル系列を主に使用する．

2 HMM 音声合成

2.1 HMM 音声合成の概要

Fig. 1 はHMMを用いた音声合成システムの学習・合成部の概要である．学習部においては，まず特徴量（スペクトル，Ｆ０，ＡＰ成分）を抽出する．これらの特徴量はコンテキスト依存HMMによってモデル化される．また，音素継続長モデルを導入することにより，それぞれの特徴量と継続長を統一的枠組みでモデル化することが可能である．合成部においては，まず入力テキストをテキスト解析することによってコンテキスト依存ラベル系列を得る．そのコンテキスト依存ラベル系列を基にしてコンテキスト依存HMMを結合することによって文HMMを生成する．そのときHMM状態系列 $q = [q_1, \cdots, q_T]$ が Eq. (1) によって継続長モデルから決定される．

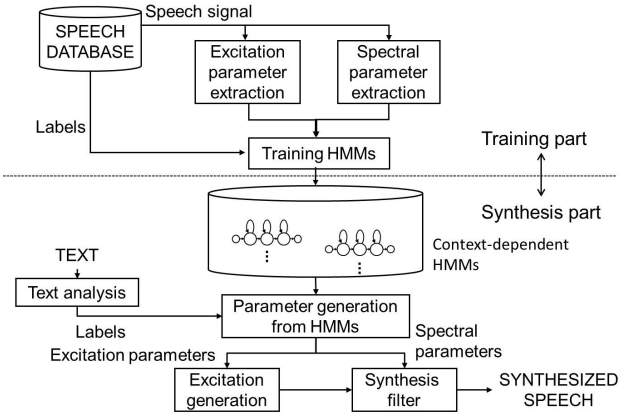


Fig. 1: HMM 音声合成システムの概要

$$\hat{q} = \arg\max_{q} P(q|\lambda) \tag{1}$$

このとき，変数 T はフレーム数，q_t はフレーム t におけるHMM状態インデックス，λ はHMMのパラメータセットである．静的・動的特徴量の間の明示的な制約の下で，パラメータセットはHMM尤度が最大となるように生成される [7].

$$c = \arg\max_{c} P(\mathbf{W}c|\hat{q}, \lambda) \tag{2}$$

Eq. (2) において，$c = [{c_1}^\mathsf{T}, \cdots, {c_t}^\mathsf{T}, \cdots, {c_T}^\mathsf{T}]^\mathsf{T}$ は音声パラメータ系列を示し，$c_t = [c(1), \cdots, c(D)]^\mathsf{T}$ は t フレームでの音声パラメータ系列，D は次元数，$\mathbf{W}$ は動的特徴量を計算に用いて構築した重み行列を示している [8]．パラメータ生成の後，MLSA フィルタ [9] を用いてパラメータ系列から音声が合成される．

2.2 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易い合成音を作成した．Fig. 2 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 [10] 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 2.3 節にあるように修正する．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

* Individuality-Preserving Voice Reconstruction for Articulation Disorders Using Text-to-Speech Synthesis

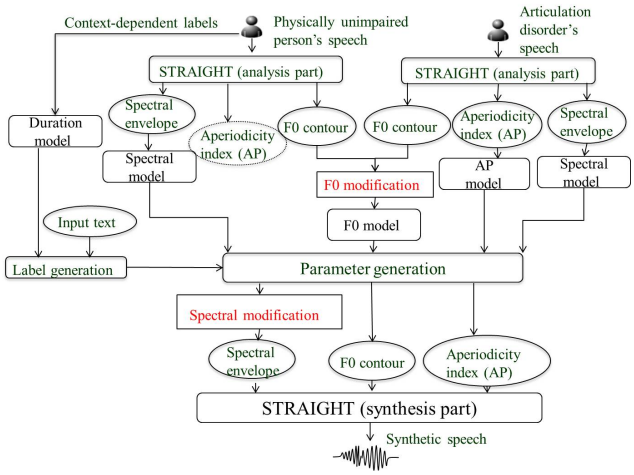
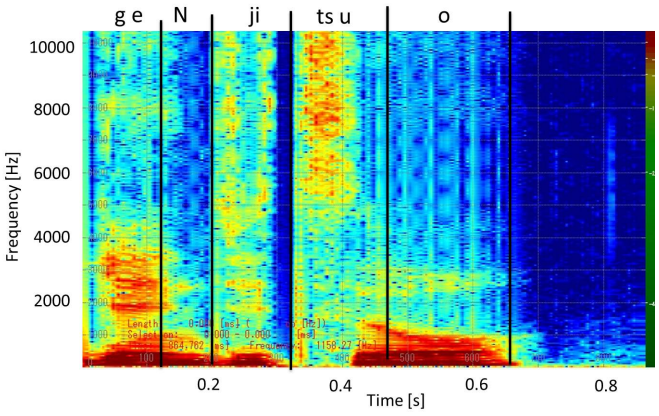
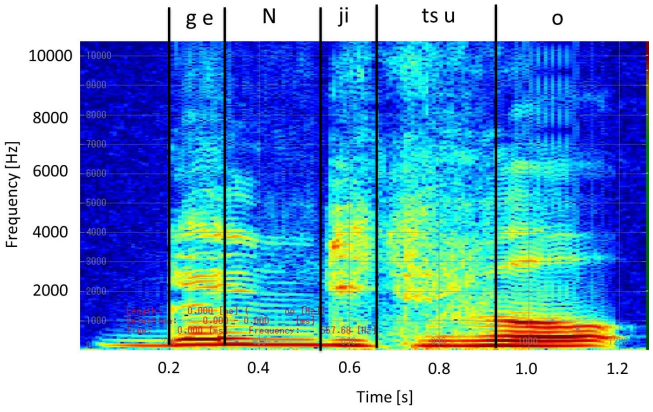


Fig. 2: 構音障害者のための HMM 音声合成手法の概要



(a) 健常者



(b) 構音障害者

Fig. 3: 元音声スペクトルの一例 // g e N j i ts u o

理を行う．音素継続長モデルについて，構音障害者の発話長は健常者に比べて長くなっているので，音素継続長モデルは健常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 パラメータ系列が生成される．F0 系列は修正した F0 モデルから生成，AP パラメータ系列は構音障害者の AP モデルから生成する．スペクトルパラメータ系列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．スペクトルを生成した後，障害者のスペクトル系列は 2.4 節のように修正される．最後に，STRAIGHT によって最終的な合成音が生成される．2.3 節，2.4 節ではスペクトルと F0 系列に対する処理の詳細を記述する．

2.3 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 系列を学習データとして学習するので，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. (3) のような線形変換を利用する．

$$\hat{x}_t = \frac{\sigma_y}{\sigma_x}(x_t - \mu_x) + \mu_y \tag{3}$$

Eq. (3) 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．

2.4 スペクトラム系列の修正

Fig. 3 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. 3 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．そこで構音障害者のスペクトルの子音部分に関しては補正を行う必要がある．補正の方法としては Fig. 2 のように，入力テキストが与えられた後，構音障害者と健常者それぞれのスペクトルモデルからスペクトルパラメータ系列を生成する．そして，低域部分には構音障害者のスペクトルを，高域部分には健常者のスペクトルを使って繋ぎ合わせたスペクトルパラメータ系列を作成する．繋ぎ合わせる際，境界部分を滑らかにするため，以下の式を使って繋ぎ合わせる．

$$\hat{S}^{(ij)} = f_m^{(j)} S_m^{(ij)} + f_g^{(j)} S_g^{(ij)} \tag{4}$$

Eq. (4) において，S_m は健常者のスペクトル，S_g は障害者のスペクトル，$\hat{S}$ は補正後のスペクトル，i はスペクトルのフレームのインデックス，j は次元のインデックスをそれぞれ表している．重み関数 f_m，f_g は以下のように定義される．

$$f_m^{(j)} = \frac{1}{1 + e^{(-j+S)}} \tag{5}$$

$$f_g^{(j)} = \frac{1}{1 + e^{(j-S)}} \tag{6}$$

このとき，f_m は健常者スペクトルに対する重み関数，f_g は構音障害者に対する重み関数，S は制御変数をそれぞれ表している．Eq. (4) 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．スペクトルの補正は Eq. (4) による処理をフレームごとに行うことによって実現し，周波数の閾値を制御する変数 S は母音部と子音部ごとに Eq. (6) によって設定する．

$$S = \begin{cases} \frac{4000}{f_s} \times D & consonant \\ \frac{5000}{f_s} \times D & vowel \end{cases} \tag{7}$$

Table 1: 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
Ref1	PU	AD	AD	AD
Prop	PU	convPU	AD	MIX
Ref2	PU	convPU	AD	AD
PUM	PU	PU	PU	PU

Note
ADM: 構音障害者モデル
Prop: 提案手法
PUM: 健常者モデル
AD: 構音障害者
PU: 健常者
convPU: Eq. (3) により，F0 を線形変換
MIX: Eq. (4) により，スペクトルを補正

Eq. (7) 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

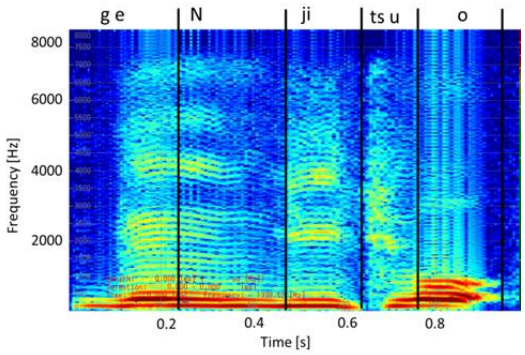
3.1 実験条件

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 文である．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 を用いて抽出した．スペクトルパラメータ系列としては，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 [11] とその Δ，ΔΔ を使用し，学習，合成には５状態のコンテキスト依存 HMM を使用した [2]．

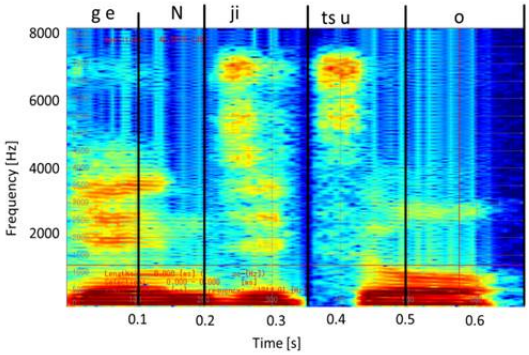
提案法の有用性を示すため，本研究では話者性と聞き取り易さの２つの観点からの実験を試みた．Table 1 のような，５つの条件のもとで５種類の合成音をそれぞれ ATR データベース中の 10 文を基に作成した．８人の日本人に対して聴衆実験を行った．話者性に関する実験には本研究では MOS(Mean Opinion Score) テスト [12] を実施した．このテストではそれぞれの音に対して５段階で評価をした．(5:非常に似ている, 4: とても似ている, 3: まあまあ似ている, 2: 似ていない, 1: 全く似ていない）聞き取りやすさの評価には一対比較法を行い２つの音声のうちより聞き取り易い方を選んで評価した．

3.2 実験結果

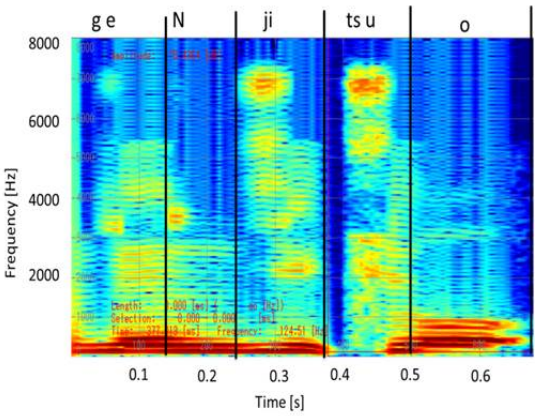
Fig. 4a は構音障害者スペクトルモデルから生成したスペクトルであり，Fig. 4b は健常者スペクトルモデルから生成したスペクトルである．Fig. 4a の高周波成分は，Fig. 4b と比較して弱くなっており，これが子音の聞き取りにくさの原因となっている．Fig. 4c は Eq. (4) による補正後のスペクトルである．Fig. 4c より，ADM の低周波成分は保ちつつ高周波成分を PUM によって補完できていることがわかる．Fig. 5 は話者性に関する実験結果である．ここでは ADM がもっとも良い値であることがわかる．これは ADM が最も構音障害者成分を多く使っているからだと考えられる．Prop はスペクトル補正を行ったことにより Ref1, Ref2 に比べて少し値が下がる結果となった．Fig. 6 は聞き取り易さに関する実験結果である．Fig. 6 より，Prop が ADM，Ref1，Ref2 の中で最も聞き取り易いということがわかる．これはスペクトルの補正が聞き取り易さに関して有効であることを示している．



(a) ADM スペクトル



(b) PUM スペクトル



(c) 補正後スペクトル

Fig. 4: 合成後スペクトルの一例

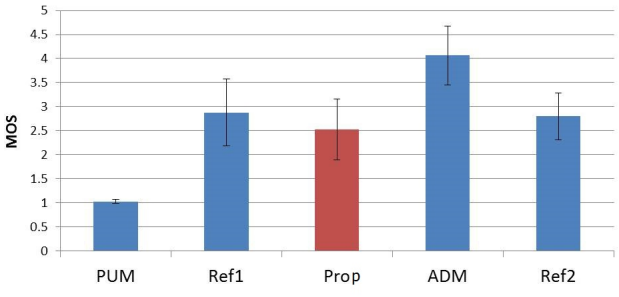


Fig. 5: 構音障害者の話者性に関する実験結果

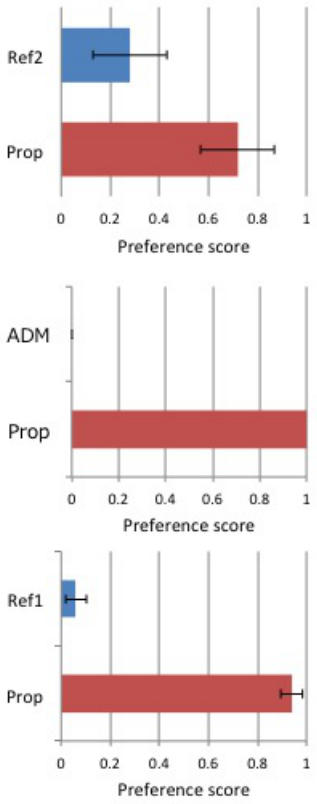


Fig. 6: 聞き取り易さに関する比較

4 おわりに

本研究ではスペクトルの補正によって聞き取り易さは大幅に改善したものの，話者性に関しては少し下がってしまうという結果になった．今後は子音部分の補正を学習段階で行い話者性の保持を目指していきたいと考えている．

参考文献

[1] T. Canale and W. C. Campbell, *Campbell's operative orthopaedics*. Technical report, Mosby Year Book, June 2002, vol. 12.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. of Eurospeech*, 1999, pp. 2347–2350.

[3] Z. Ahmad Khan, P. Green, S. Creer, and S. Cunningham, "Reconstructing the voice of an individual following laryngectomy," *Augmentative and Alternative Communication*, vol. 27, no. 1, pp. 61–66, 2011.

[4] C. Veaux, J. Yamagishi, and S. King, "Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders," in *Proc. of Interspeech*, 2012.

[5] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction," *Acoustical Science and Technology*, vol. 33, no. 1, pp. 1–5, 2012.

[6] S. Creer, S. Cunningham, P. Green, and J. Yamagishi, "Building personalised synthetic voices for individuals with severe speech impairment," *Computer Speech & Language*, vol. 27, no. 6, pp. 1178–1193, 2013.

[7] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. of ICASSP*, 2000, pp. 1315–1318.

[8] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, pp. 1039–1064, 2009.

[9] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, pp. 10–18, 1983.

[10] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigné, "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.

[11] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT." in *Proc. of MAVEBA*, 2001, pp. 59–64.

[12] I. T. Union, "ITU-T Recommendation P.800.1: Mean Opinion Score (MOS) terminology," International Telecommunication Union, Tech. Rep., July 2006.