社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

話者性を維持した構音障害者のためのHMM音声合成システム

上田　怜奈　　滝口　哲也　　有木　康雄

神戸大学大学院システム情報学研究科
〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
E-mail: reina_1102@me.cs.scitec.kobe-u.ac.jp, {takigu,ariki}@kobe-u.ac.jp

あらまし　本研究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．アテトーゼ型脳性麻痺から起こる構音障害は，意図した動作に緊張を起こさせるため彼らの動作はしばしば健常者と比べて不安定なものとなってしまう．また，口の筋肉を上手く動かせないことで発話も不明瞭になりやすい．このような構音障害を持つ人々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．そこで，本研究では音声合成システムを使用し構音障害者の発話の手助けをすることを試みた．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システムから出力される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．この問題を解決するため，彼らの話者性は維持しつつより聞き取りやすい音声を作り出すTTSシステムを構築する必要がある．本研究ではHMM音声合成システムをベースとし，健常者音声と構音障害者音声の両方を学習データとして利用し，構音障害者音声のスペクトル，ピッチ，話速のそれぞれを健常者成分を用いて修正を行った．評価実験の結果から，今回の提案法が障害者の話者性は維持しつつより聞き取りやすい合成音声を実現出来ていることを示す．

キーワード　音声合成，HMM, 構音障害者, 話者性，聞き取りやすさ

Individuality-Preserving HMM Speech Synthesis System for Articulation Disorders

Reina UEDA, Tetsuya TAKIGUCHI, and Yasuo ARIKI

Graduate School of System Informatics, Kobe University
1-1 Rokkodaicho, Nada-ku, Kobe, Hyogo, 657-8501 Japan
E-mail: reina_1102@me.cs.scitec.kobe-u.ac.jp, {takigu,ariki}@kobe-u.ac.jp

Abstract　This paper presents a speech synthesis method for a person with an articulation disorder resulting from the athetoid type of cerebral palsy. Cerebral palsy results from damage to the central nervous system, and the damage causes movement disorders. Because his/her rip movements are sometimes more unstable than usual due to the athetoid symptoms, their utterances (especially their consonants) are often unstable or unclear. This is why there is great need for speech synthesis system to aid them in their communication. In this paper, we propose an HMM-based speech synthesis method for articulation disorders. To generate the intelligible voice while preserving the speaker's individuality, our training data include the voice of a physically unimpaired person. Then we modified patient's spectrum, pitch and duration by using features of a physically unimpaired person. The experimental results demonstrate that our proposed method achieves the output synthesized signals which are intelligible and preserve the patient's individuality.

Key words　Speech Synthesis System，HMM，Articulation Disorders，Similarity，Intelligibility

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．これまでの研究ではピッチ，スペクトル，話速を健常者特徴を用いて修正し、聞き取りやすさに対するそれぞれの独立した効果を実験を通して確認してきた. [3] [4] [5] 本研究では構音障害者の話者性を維持しつつ聞き取りやすさを向上させるため，ピッチ，スペクトル，話速をすべて修正できる HMM 音声合成システムを提案する．学習データに構音障害者と健常者音声を使用し，聞き取りにくさの原因となる構音障害者音声の成分を健常者音声によって補完し，より聞き取りやすい音声を実現する．評価実験を通して本研究の提案法が障害者の話者性は維持しつつより聞き取りやすい合成音声を実現出来ていることを示す．

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 [6] を用いて二人の話者から 3 つの音声パラメータ (F0 概形，スペクトル包絡，非周期成分 (AP)) を抽出する．特徴量を抽出したのち，障害者の 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 モデルから生成される．AP パラメータは構音障害者のモデルから生成する．スペクトルパラメータは構音障害者と健常者両方のモデルから生成され，その後スペクトル修正を行う (2.3 節)．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，2.3 節では F0・音素継続長・スペクトルに関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．

$$\hat{y}_t^{(pit)} = \frac{\sigma^{(y_{pit})}}{\sigma^{(x_{pit})}}(x_t^{(pit)} - \mu^{(x_{pit})}) + \mu^{(y_{pit})} \quad (1)$$

Eq. (1) において，$x_t^{(pit)}$ は健常者の t フレーム目の対数 F0，

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

表 1: 学習後音素継続長モデルの平均，分散

$\mu^{(x_{pit})}$, $\sigma^{(x_{pit})}$ は健常者の F0 系列の平均･分散, $\mu^{(y_{pit})}$, $\sigma^{(y_{pit})}$ 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 音素継続長の修正

構音障害者の話速は健常者のものと比べて全体的に間延びしたものとなっており，音素ごとの音素長にもばらつきが見られる．このことが，聞き取りにくさの原因となっている．Fig. 4 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. 4 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

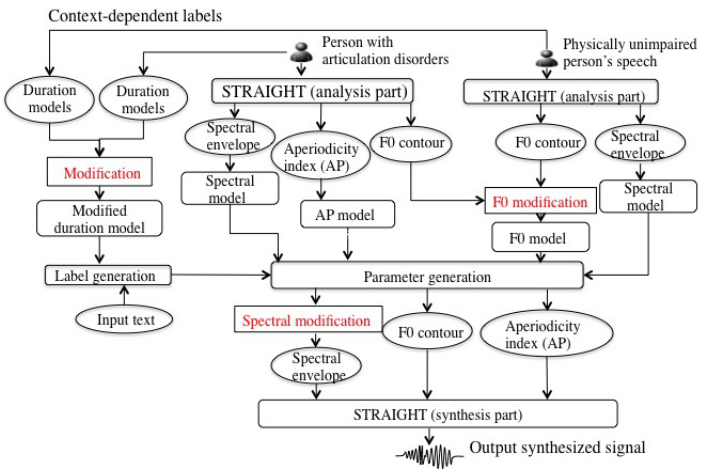
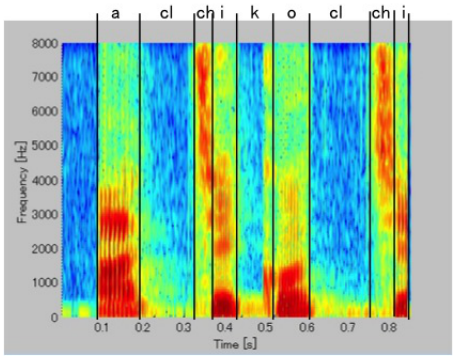
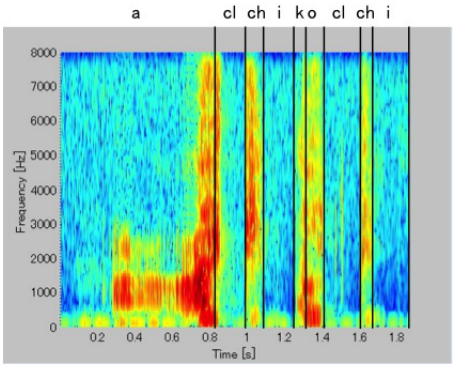


図 1: 構音障害者のための HMM 音声合成手法の概要



(a) 健常者



(b) 構音障害者

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

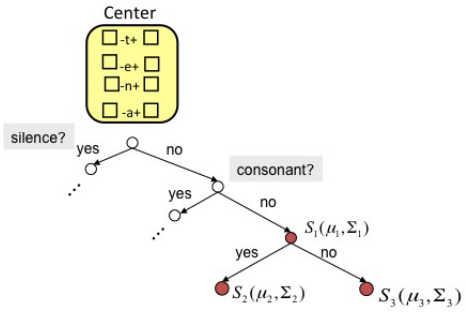


図 3: 健常者音素継続長モデルの修正

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．

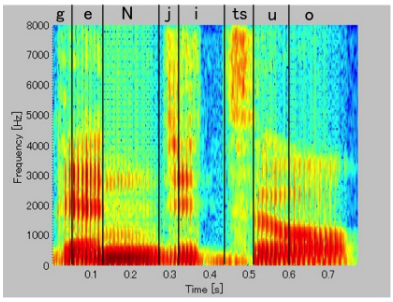
$$\hat{y}_i^{(dur)} = x_i^{(dur)} - \mu^{(x_{dur})} + \mu^{(y_{dur})} \tag{2}$$

$$\mu^{(x_{dur})} = \frac{\sum_{i=1}^{I} \mu_i^{(x_{dur})}}{I} \tag{3}$$

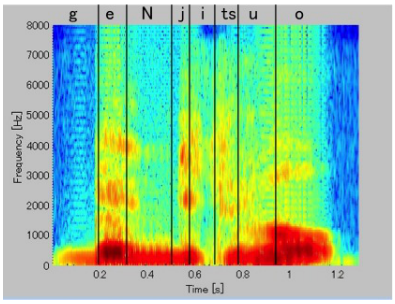
$$\mu^{(y_{dur})} = \frac{\sum_{i=1}^{I} \mu_i^{(y_{dur})}}{I} \tag{4}$$

Eq. (2) において，$x_i^{(dur)}$ 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，$\mu^{(x_{dur})}$，$\mu^{(y_{dur})}$ 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，$\mu_i^{(x_{dur})}$ は健常者モデルの i 番目の母音ノードの平均値，$\mu_i^{(y_{dur})}$ 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．

2.3 スペクトル系列の修正



(a) 健常者



(b) 構音障害者

図 4: 元音声スペクトルの一例// g e N j i ts u o

Fig. 4 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. 4 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．

$$\hat{S}^{(ij)} = f_{PU}^{(j)} S_{PU}^{(ij)} + f_{AD}^{(j)} S_{AD}^{(ij)} \tag{5}$$

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，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．

$$f_{PU}^{(j)} = \frac{1}{1+e^{(-j+c)}}, f_{AD}^{(j)} = \frac{1}{1+e^{(j-c)}} \tag{6}$$

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. (5) を用いることにより，高周波領域では健常者の

表 2: 健常者と構音障害者の類似度

	SpD
FKN	3.72.E-03
FKS	2.87.E-03
FTK	2.97.E-03
FYM	3.60.E-03
MMY	3.02.E-03
MTK	**2.79.E-03**
MHO	3.62.E-03
MHT	3.79.E-03
MSH	3.46.E-03
MYI	3.65.E-03

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. (7) によって閾値が 4000Hz になるように設定する．

$$c = \frac{4000}{f_s} \times D \tag{7}$$

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 予備実験

本研究においては，構音障害者の音素継続長・F0・スペクトル特徴に対して健常者成分を用いて修正を行うことから，障害者の話者性と大きく異なる健常者音声を採用すると，生成される合成音の話者性も低下してしまう恐れがある．よって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の音声を選ぶことが望ましいと考えられる．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ラムの話者間の距離をそれぞれ算出した．類似度の算出には ATR データベースセット B 中の１０文を使用した．

Table 2 は構音障害者と健常者１０名それぞれの類似度の算出結果である．Table 2 より，使用する構音障害者音声には 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者音声には MTK の音声データを採用した．

3.2 実験条件

学習データには構音障害者の男性１名，健常者男性１名（MTK）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中の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は，２５次元のメルケプストラムとその 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．提案法の有効性を示すため本研究では話者性と聞き取りやすさの２つの観点から実験を試みた．

表 3: 実験で比較した合成音の生成条件

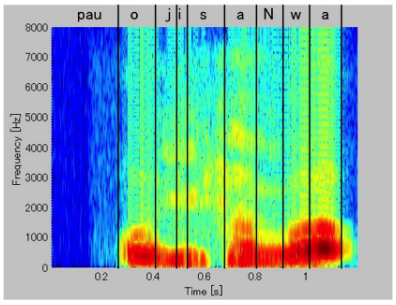
Type \ Model	Duration	F0	AP	Spectral
ADM	AD	AD	AD	AD
Dur(健)	PU	AD	AD	AD
Dur	Mod	AD	AD	AD
Dur_F0	Mod	Mod	AD	AD
Dur_Spe	Mod	AD	AD	Mod
Dur_F0_Spe	Mod	Mod	AD	Mod

（**AD**: 構音障害者，**PU**: 健常者，**Mod**:修正有り）

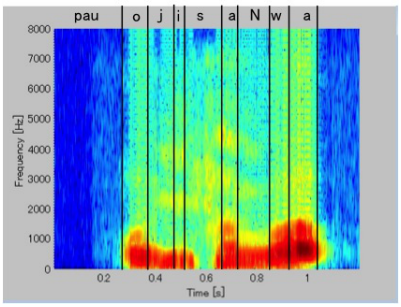
１０人の日本人に対してヘッドホンで聴取実験を行った．話者性に関する実験には本研究では DMOS(Degradation Mean Opinion Score) テストを実施した．このテストではリファレンス音声と評価対象音声を聞き比べ，評価対象音声がどれだけ劣化しているかを５段階（5:劣化が全く認められない 4:劣化が認められるが気にならない 3:劣化がわずかに気になる 2:劣化が気になる 1:劣化が非常に気になる）で評価した．聞き取りやすさの実験は一対比較法で行った．聴取実験は二回に分けて行い，一回目の実験では音素継続長修正の効果を，二回目の実験では F0・スペクトル修正の効果をそれぞれ検証した．

3.3 実験結果

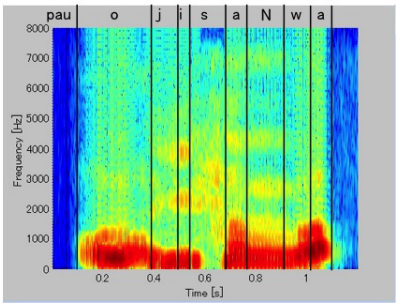
一回目の実験では音素継続長修正の効果について検証した．実験にあたり，Table 3 のうち ADM，Dur(健)，Dur の３種類の合成音を用意した．Fig. 5 は３種類の条件を基に作成した合成音のスペクトルである．テキストは全て同じ発話である．Fig. 5c は構音障害者の音素継続長モデルで作成したスペクトルである．発話は全体として長くなっており，特に発話中の “o"が間延びしていることが分かる．Fig. 5b は健常者の音素継続長モデルで作成したスペクトルである．Dur(健) と Dur(構) を比較すると Dur(健) の方が発話時間が短く音素の間延びも見られないことが分かる．Fig. 7a 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．Fig. 8a は話者性に関する実験結果である．Fig. 8a より，Dur の方が Dur(健) よりも優位な結果となった．このことから，健常者の音素継続長モデルをそのまま使うよりも修正後音素継続長モデルを使う方が話者性が保たれることが分かった．聞き取りやすさに関する実験では ADM と Dur，Dur と Dur(健) をそれぞれ比較した．Fig. 9a より，ADM よりも Dur のほうが優位であるとわかる．ここから，修正後音素継続長モデルを利用するほうが，構音障害者の音素継続長モデルを利用するよりも聞き取りやすさは向上することがわかった．Fig. 9b より，Dur と Dur(健) がほぼ同じ評価となった．ここから，健常者の音素継続長モデルを利用した時と，構音障害者の音素継続長モデルを利用したときでは聞き取りやすさはほぼ同じで遜色はないことが分かる．以上の結果より音素継続長モデルの修正が話者性と聞き取りやすさの両面から見て有効であることがわかった．



(a) Dur



(b) Dur(健)



(c) ADM

図 5: 合成スペクトルの一例// pau o j i s a N w a

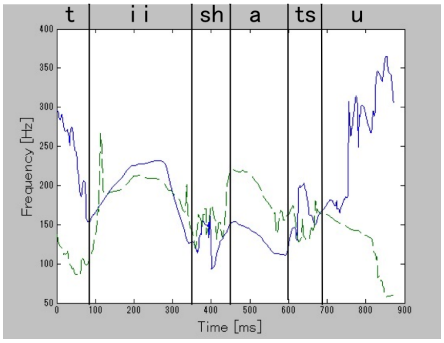
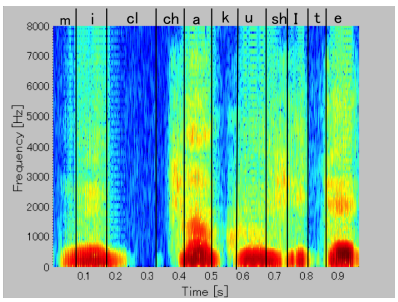
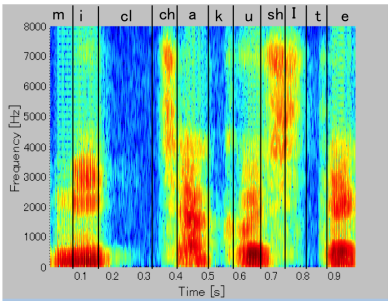


図 6: F0 修正前と修正後の比較
//t ii sh a ts u 実線:修正前 破線:修正後

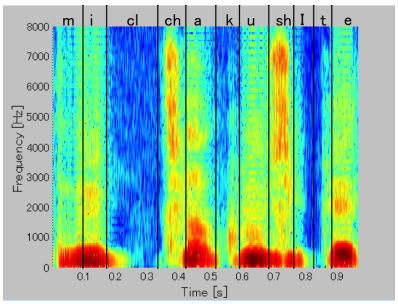
二回目の実験では F0 修正，スペクトル修正の効果について検証した．F0 修正について，Fig. 6 は Dur と Dur_F0 の合成音のピッチを表示したものである．実線が Dur（修正前），破線が Dur_F0（F0 修正後）である．どちらも同じ修正後音素継続長モデルを使用しているためフレーム数は一致している．Fig. 6 では日本語で “T シャツ” と発話している．Dur において



(a) 障害者合成音スペクトル



(b) 健常者合成音スペクトル



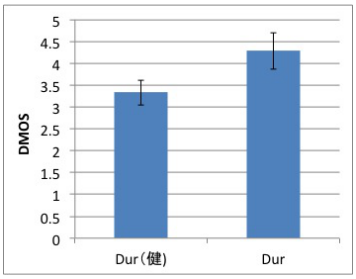
(c) 修正後スペクトル

図 7: 合成後スペクトルの一例
//m i cl ch a k u sh I t e

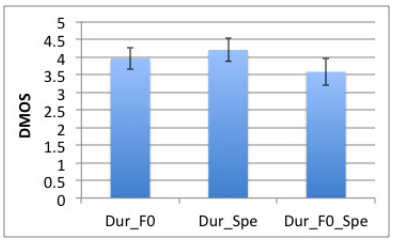
“t” や ‘u” の部分のピッチが不自然に上がっているのに対して，Dur_F0 では正しいイントネーションに修正されていることがわかる．

Fig. 7 は合成音のスペクトルであり，パラメータ生成にあたりいずれも修正後音素継続長モデルを用いていることからフレーム数は一致している．テキストはすべて同じ発話内容で日本語で “密着して” と発話している．Fig. 7a 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. 7b 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. 7a の高周波成分は，Fig. 7b 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. 7c は Eq. (5) による補正後のスペクトルである．この修正により，修正該当音素である，ch/sh の高周波域が PUM によって補完されていることがわかる．

Fig. 8b は話者性に関する実験結果である．ここでは Dur をリファレンスとして 5 段階 DMOS で Dur_F0，Dur_Spe，Dur_F0_Spe を評価した．Fig. 8b より，Dur_Spe，Dur_F0，Dur_F0_Spe の順で話者性を保持出来ていることが分かる．



(a) リファレンス音声:ADM



(b) リファレンス音声:Dur

図 8: 話者性に関する比較

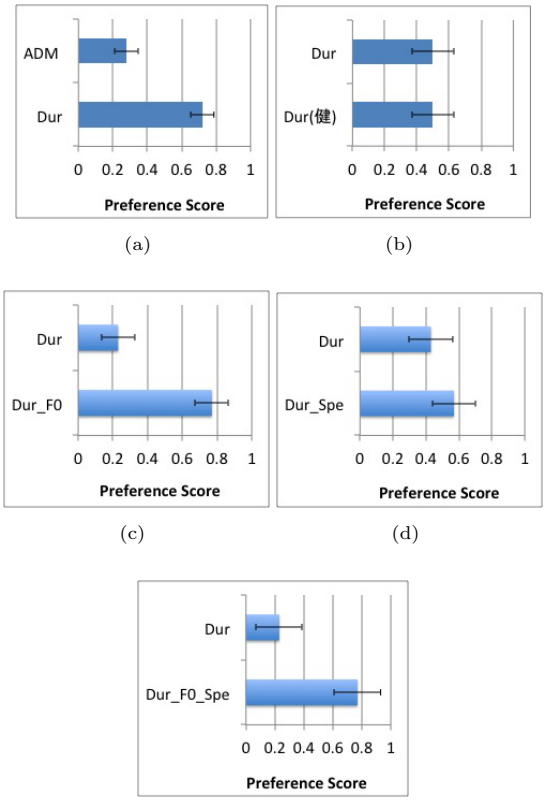


図 9: 聞き取り易さに関する比較

Dur_F0，Dur_Spe の間にはそれほど大きな差は見られなかったが，Dur_F0_Spe のスコアは他の 2 条件と比べて低い値となっている．このことから F0 修正，スペクトル修正の両方を適用すると被験者は話者性が下がったと感じることが分かる．聞き取りやすさに関する実験では，Dur と Dur_F0，Dur と Dur_Spe，Dur と Dur_F0_Spe をそれぞれ聞き比べてよりどちらがより聞き取りやすいかを評価した．Fig. 9c より，Dur と Dur_F0 を比較したとき Dur_F0 のほうが大幅に優位な結果となった．これは F0 修正が聞き取りやすさに対して大きな効果があることを示している．Fig. 9d は Dur と Dur_Spe を比較したときの実験結果である．Fig. 9d において，Dur と Dur_Spe 間で有意差は確認できなかった．これは，スペクトル修正の対象となる音素が非常に少ないことが一因していると考えられる．Fig. 9e は Dur と Dur_F0_Spe を比較したときの実験結果である．Fig. 9e より，Dur と Dur_F0_Spe を比較したとき，Dur_F0_Spe のほうが優位な結果となった．このことから，F0 修正，スペクトル修正の両方を適用すると被験者は聞き取りやすさが向上したと感じたことが分かる．

4. おわりに

本研究では構音障害者のための話者性を維持した HMM 音声合成手法を提案した．実験を通して F0・音素継続長修正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．今後はより広い周波数域に対応したスペクトル修正法を検討していきたい．

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．

文　献

[1] C. Veaux, J. Yamagishi and S. King: “Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders”, Proc. of Interspeech (2012).

[2] 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, **33**, 1, pp. 1–5 (2012).

[3] R. Ueda, R. Aihara, T. Takiguchi and Y. Ariki: “Individuality-preserving spectrum modification for articulation disorders using phone selective synthesis”, 6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), pp. 118–123 (2015).

[4] R. Ueda, T. Takiguchi and Y. Ariki: “Individuality-preserving voice reconstruction for articulation disorders using text-to-speech synthesis”, Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 343–346 (2015).

[5] 上田怜奈, 滝口哲也, 有木康雄：“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”, 日本音響学会 2016 年秋季研究発表会, No. 3-Q-17, pp. 229–232 (2016).

[6] 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, **27**, pp. 187–207 (1999).