Vol.2010-SLP-80 No.12
2010/2/13

HMM 音声合成における自然性と個人性に優れた韻律モデル適応法の検討

神山 歩相名[†1] 篠崎 隆宏[†1]
岩野 公司[†2] 古井 貞熙[†1]

本稿では，HMM 音声合成における音素継続時間長モデルの話者適応法を提案する．提案手法は，数量化 I 類に基づくモデルの平均値を変換することで適応化を行う．客観評価実験を行ったところ，提案手法は 5 文程度で適応効果が収束することが確認された．また主観評価実験を行ったところ，5 文で適応したモデルと 470 文で学習したモデルでほぼ同程度の自然性と個人性が認められた．さらに，音素継続時間長モデル，F_0 パターン生成モデルを平均値変換によって適応し，ケプストラムモデルを SMAPLR 適応したモデルから音声合成を行い主観評価実験を行った．その結果，20 文で話者適応したモデルによる音声と 470 文で学習したモデルによる音声に，ほぼ同程度の自然性と個人性が認められた．これより，提案手法が自然性と個人性に優れた韻律モデル適応法であることが確認された．

A Prosody Adaptation Method for HMM-based Speech Synthesis Achieving High Naturalness and Individurity

Hosana Kamiyama,[†1] Takahiro Shinozaki,[†1]
Koji Iwano[†2] and Sadaoki Furui[†1]

This paper proposes a phoneme duration adaptation method for HMM-based speech synthesis. The proposed method converts mean values of the duration models based on the Quantification Theory (Type I). Objective evaluation results for the models made by the adaptation method confirm that around five sentences are enough for adaptation. Subjective evaluation results confirm that naturalness and inviduality of the synthesized speech using models adapted by five sentences is almost equivalent to that of the synthesized speech using models trained by 470 sentences for a specific speaker. Finally, we synthesized speech using F_0 contour generation models, duration models made by the mean adaptation method and cepstrum HMM adapted by the SMAPLR method. Subjective evaluation results confirm that naturalness and inviduality of the synthesized speech using models adapted using 20 sentences is almost equivalent to that of the synthesized speech using models trained by 470 sentences for a specific speaker. These results indicate that the proposed method for the prosody models can effectively produce synthesized speech with naturalness and inviduality.

1. はじめに

近年，Web コンテンツや電子メールの読み上げなどの様々な分野でテキスト音声合成（Text-to-speech: TTS）の技術が用いられるようになりつつある．これらの応用が進むにつれ，合成音声には聞き取りやすさとともに様々な話者の個人性を表現することが求められるようになってきている．

当研究室は，図 1 に示すような隠れマルコフモデル（HMM）に基づいた TTS システムを開発してきた[1)2)]．この TTS システムはケプストラム及び非周期性指標を HMM で，音素継続時間長及び基本周波数情報（F_0 パターン）を数量化 I 類でモデル化しており，各特徴量をこれら統計的モデルから推定することで音声合成を行う．しかし，これらの統計的モデルの学習には大量の音声データが必要となるため，応用の観点からは特定の話者の少量の音声データから話者の特徴を示すモデルを生成することが望まれている．HMM を用いたケプストラム及び F_0 の推定についてはこれまでに大量のデータから学習した話者独立モデルを少量の話者音声によって適応化する手法が提案されており，4 文程度の音声で個人性に優れた音声合成が実現されている[3)4)]．しかし数量化 I 類を用いたモデルに対する適応手法は，これまでほとんど研究が行われてこなかった．そこで，文献[5)] において数量化 I 類によってモデル化された F_0 パターン生成モデルの平均値成分を変換する話者適応法を提案し，5 文程度の話者適応化により 400〜450 文で学習したモデルとほぼ同程度の自然性と個人性の音声合成が実現できることを示した．しかし，音素継続時間長の話者適応法については，未だ検討が行われていなかった．

本研究では，同様に数量化 I 類を用いてモデル化した音素継続時間長モデルに対して平均値変換による話者適応化を応用し，客観評価実験と主観評価実験によってその性能について

†1 東京工業大学 情報理工学研究科 計算工学専攻
Department of Computer Science, Tokyo Institute of Technology
†2 東京都市大学 環境情報学部 情報メディア学科
Faclty of Environmental and Information Studies, Tokyo City University

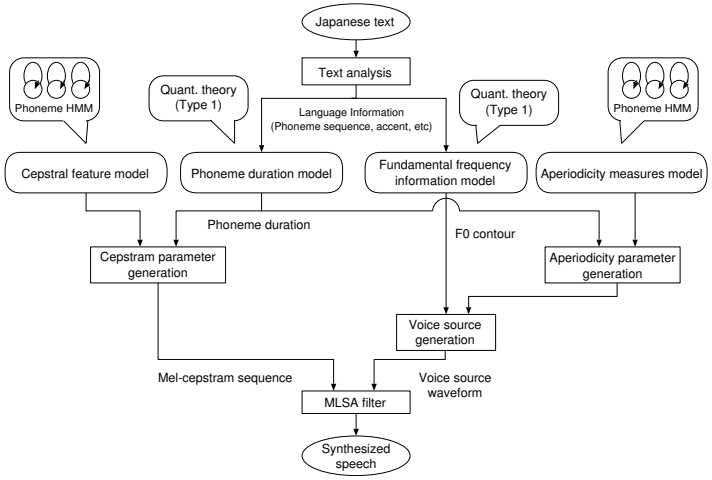

図 1 HMM に基づくテキスト音声合成システム
Fig. 1 HMM-based text-to-speech system

評価する．複数の話者の大量の音声で学習した話者独立モデルは，個人性は失われるものの様々な話者の音素長や話速が平均化され日本語（標準語）として自然な音素継続時間長をモデル化したモデルであると考えられる．提案手法は，この話者独立モデルの平均値成分を特定の話者に合わせて置き換えることで，自然性の劣化を抑えつつ個人性を取り込む話者適応化を目指す．

以下では，第 2 章においてまず数量化 I 類と先行研究[5] で提案した平均値変換による話者適応法について説明する．ついで，第 3 章において音素継続時間長の数量化 I 類を用いた制御法について説明し，その平均値変換による話者適応法について提案する．その後，第 4 章において客観評価実験と主観評価実験により音素継続時間長モデルの話者適応法の性能を評価する．第 5 章において少量の話者音声を用いて音素継続時間長と F_0 パターン生成モデルの平均値変換による話者適応と，ケプストラムモデルの SMAPLR 法による話者適応を同時に行う．適応した各モデルを用いて音声合成を行い，主観評価実験によって自然性と個人性について評価する．最後に第 6 章でまとめと今後の課題について述べる．

2. 数量化 I 類と平均値変換による話者適応法

本章では，数量化 I 類によるモデル化手法ついて説明し，ついで F_0 パターン生成モデルの話者適応[5] を目的として提案した平均値変換による話者適応法について説明する．

2.1 数量化 I 類

数量化 I 類[6] とは，質的説明変数（制御要因）と目的とする量的変数を線形重回帰分析に基づいてモデル化する手法である．数量化 I 類では，制御要因（アイテム）内の質的説明変数の選択肢をカテゴリーといい，以下の式で定式化される．

$$\hat{y}_i = \bar{y} + \sum_f \sum_c x_{fc}\delta_{fc}(i) \quad (i = 1, \cdots, N) \tag{1}$$

$\bar{y}$ は平均値成分，N はサンプル数である．$\delta_{fc}(i)$ は i 番目のデータのアイテム f がカテゴリー c に属する場合に 1，それ以外のときに 0 を与える関数である．重み x_{fc} はアイテム f カテゴリ c の数量（カテゴリースコア）であり，推定二乗誤差 $E = \sum_i (\hat{y}_i - y_i)^2$ を最小化するように求められる．

2.2 平均値変換による話者適応法

数量化 I 類は平均値成分とカテゴリースコアによって定式化され，音声合成においては，これらの数量が音声の自然性や話者の個人性を特徴づける．しかし学習データ数が少ない場合は，カテゴリースコアの過学習が生じる場合が多く，自然性劣化の要因となる．そこで，平均値変換による話者適応法[5] では適応前の初期モデルの平均値成分を適応対象に合うように値を置き換えることで話者適応化を行う．このときカテゴリースコアは適応前初期モデルのものをそのまま使うため，自然性の劣化を抑えることができる．置換する新しい平均値成分 $\bar{y}'$ は，適応データに対する推定二乗誤差を最小化することにより求める．

$$\frac{\partial E}{\partial \bar{y}'} = \frac{\partial}{\partial \bar{y}'} \sum_i (\hat{y}'_i - y'_i)^2 = 0 \tag{2}$$

$$\Rightarrow \bar{y}' = \frac{1}{N'} \sum_i (y'_i - \sum_f \sum_c x_{fc}\delta_{fc}(i)) \tag{3}$$

$\hat{y}'_i$ は，適応対象話者についての i 番目データの推定値，y'_i は適応データのサンプル値，N' はサンプル数である．この式 (3) で推定される平均値成分はカテゴリースコアを用いて推定するため，適応データの平均値とは異なる値となる．

3. 音素継続時間長制御法と話者適応の提案

本章では，音素継続時間長の数量化 I 類による制御法について説明し，平均値変換による話者適応法の音素継続時間長モデルに対する応用を提案する．

表 1　音素クラスの一覧
Table 1　List of phoneme classes.

音素クラス	音素
1. 母音（vowel）	/a/,/i/,/u/,/e/,/o/
2. 撥音（syllabic nasal）	/N/
3. 促音（choked sound）	/Q/
4. 長音（long vowel）	/-/
5. 有声破裂音（voiced stop）	/b/,/d/,/g/
6. 無声破裂音（unvoiced stop）	/p/,/t/,/k/
7. 有声摩擦音（voiced fricative）	/z/,/j/
8. 無声破擦音（affricate）	/ch/,/ts/
9. 無声摩擦音（unvoiced fricative）	/f/,/h/,/s/,/sh/
10. 鼻音（nasal consonant）	/m/,/n/
11. 流音（liquid）	/r/
12. 半母音（semi vowel）	/w/,/y/
13. 拗音（palatalized consonant）	/by/,/dy/,/gy/,/py/, /ky/,/hy/,/ry/,/my/,/ny/

3.1　音素継続時間長制御法

音素継続時間長は，表 1 に示す 13 の音素クラスごとの数量化 I 類によって学習する[7]．音素継続時間長は，ケプストラム特徴量を triphone HMM で強制切り出しすることにより抽出する．音素継続時間長は，当該音素と先行・後続する音素，及び 2 つ前・後の音素の種類が強く影響していることが明らかにされており，これらを制御要因として用いる手法が有効であることが示されている[7]．本研究においてもそれらの制御要因によって数量化 I 類のモデル化を行う．

3.2　平均値変換による話者適応法の音素継続時間長モデルへの応用

音素継続時間長モデルは，表 1 に示した音素クラスによって 13 種類のモデルが存在するため，13 種類のモデルそれぞれに対して式 (3) によって平均値成分を求めることで話者適応化を行う．ただし，音素クラスによって分類をしたときに，適応データが存在しない場合（$N' = 0$ の場合）は，初期モデルの平均値成分をそのまま使用する．

4.　音素継続時間長モデルにおける話者適応法の評価実験

音素継続時間長モデルについて話者適応化を行い，推定誤差についての客観評価実験と自然性と個人性についての主観評価実験を行った．

4.1　使用音声データベース

実験には ATR 日本語音声データベース[8] 中の男性話者 4 名（MHT, MYI, MTK, MMY）と，女性話者 4 名（FKS, FKN, FKS, FYM）による 503 発声（A〜I セット各 50 発声，J セット 53 発声）を用いた．音声合成のためのケプストラム，非周期性指標及び F_0 の抽出は STRAIGHT 法[9] を用いて，窓幅 16ms，フレーム周期 1ms で抽出したのち，5ms 周期に特徴量を取り出して使用した．音素継続時間長の抽出に用いるケプストラムモデルは，話者 8 名の A〜I セット（450 文）で学習した話者独立モデルを使用した．

4.2　実験の流れ

適応対象話者を除く男性話者・女性話者計 7 名の A〜I セット（450 文× 7 話者）と J01〜J20（20 文× 7 話者）の計 470 文× 7 話者を用いて話者独立モデルを学習した．続いて，適応対象話者の H21〜H50（30 文）と I セット（50 文）と J01〜J20（20 文）の計 100 文からランダムに 1〜100 文を選び，提案手法によって話者適応モデルを生成した．比較に用いる話者依存モデルは，A〜I セット（450 文）からランダムに選んだ 80〜450 文と J01〜J20（20 文）の計 100〜470 文を選んで学習した．生成した各モデルから J21〜J53（53 文）の音素継続時間長を推定し，客観評価実験及び主観評価実験を行った．

4.3　客観評価実験

まず，客観評価実験として適応文数と推定誤差の関係を求めた．その際，モーラを決定付ける表 1 中の 1〜4 のクラスに属する音素を対象とした．結果を図 2 に示す．実験の結果，話者適応モデルは 5 文以上の適応において推定誤差にほとんど変化がなく，100 文の話者依存モデルと同程度の推定誤差となった．つまり，提案手法は 5 文でほぼ適応効果が収束することが分かった．一方，150 文以上で学習した話者依存モデルは 5 文以上で生成した話者適応モデルより推定誤差が小さくなっており，その差はおよそ 2〜3[ms] 程度であった．つまり，150 文以上の音声データがある場合は，適応を行うより学習をした方が推定誤差が小さくなることが分かった．よって，提案手法による話者適応法は 150 文以上で学習した話者依存モデルよりは劣るものの，5 文程度による適応によりある程度適応対象の話者に近づいた音素継続時間長モデルを生成できることが分かった．

4.4　主観評価実験

次に，提案手法で生成した話者適応モデルの自然性と個人性について主観評価実験を行った．

4.4.1　実 験 条 件

今回の実験では，話者適応モデルは話者独立モデルを 5 文で平均値変換したものを使用した．また，話者依存モデルは 100 文・200 文・300 文・400 文・470 文を用いて学習したものを使用した．音声は男性話者 3 名（MYI，MMY，MTK）女性話者 3 名（FKS，FTK，FYM）について合成した．このときケプストラム，非周期性指標及び F_0 パターンの特徴量は，合成対象話者の A〜I セット（450 文）と J セットの 20 文（J01〜J20）の計 470 文

Vol.2010-SLP-80 No.12
2010/2/13

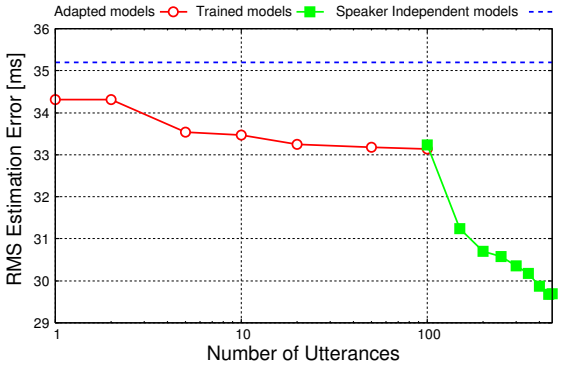

図 2　音素継続時間長モデルの使用文数と推定誤差の関係

Fig. 2　Relationship between the number of utterances used for training or adaptaion, and the average RMS estimation error in the duration modeling

で学習した話者依存モデルから生成した．評価実験は，比較を行う 2 つの合成音のペアをランダムな順に提示し自然性が高い方を選ぶペアテストと，ペアの提示の後に正解音声を提示し正解音声に個人性が近い方を選ぶ ABX テストについて行った．ABX テストで使用する正解音声は，抽出した音素継続時間長（正解値）から合成した．被験者は 11 名で音声の受聴にはヘッドフォンを使用した．

4.4.2　話者独立モデルとの比較

話者独立モデルと話者適応モデルの比較評価実験は，比較的話速の速い話者（MYI・MMY）と遅い話者（FTK・MTK）2 名ずつについて行った．結果を図 3 に示す．

自然性の評価　全てのペア間で話者適応モデルの方が自然性において低い評価を得た．総合では 5%の有意水準で有意差が確認され，話者適応することで自然性が若干劣化することが確認された．

個人性の評価　全てのペア間で話者適応モデルの方が個人性で高い評価を得た．特に，話速の遅い話者と総合においては 1%の有意水準で適応効果が認められた．

4.4.3　話者依存モデルとの比較

話者適応モデルと話者依存モデルの比較評価実験は，6 名の話者（FKS・FTK・FYM・MYI・MMY・MTK）について行った．結果を図 4 に示す．

自然性の評価　全てのペア間で話者適応モデルの方が自然性で高い評価を得た．話者独

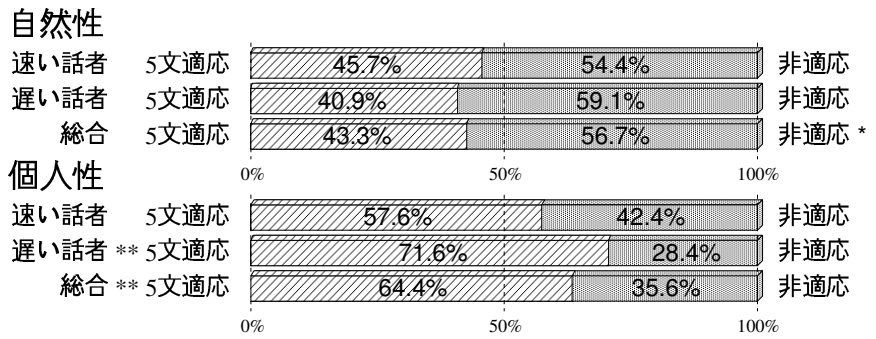

図 3　音素継続時間長の話者適応モデルと話者独立モデルを比較したプリファレンススコア．* 及び **印はそれぞれ有意水準 5%, 1%でスコア間に有意差が認められたことを示す．

Fig. 3　Preference scores of synthesized speech produced by speaker-adapted and speaker-independent duration models. “*” and “**” indicate that differences are statistically significant at 5% and 1% significance levels, respectively.

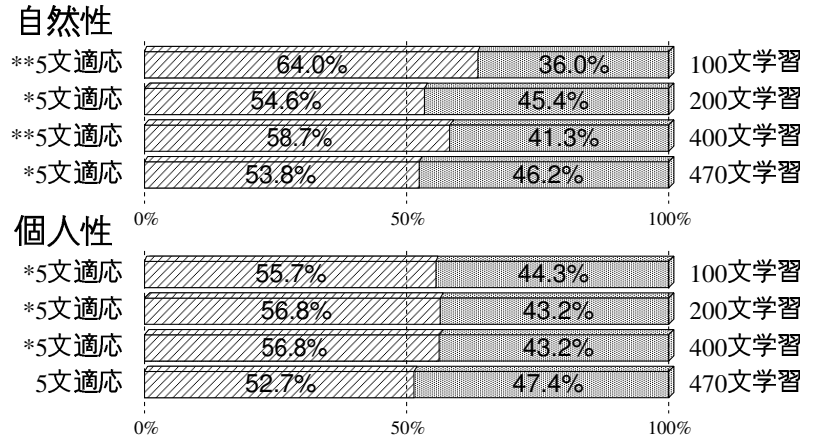

図 4　音素継続時間長の話者適応モデルと話者依存モデルを比較したプリファレンススコア．* 及び **印はそれぞれ有意水準 5%, 1%でスコア間に有意差が認められたことを示す．

Fig. 4　Preference scores of synthesized speech produced by speaker-adapted and speaker-trained duration models. “*” and “**” indicate that differences are statistically significant at 5% and 1% significance levels, respectively.

立モデルとの比較では話者適応することで自然性の劣化が確認されたが，470 文の話者依存モデルと比較した場合は，話者適応した音声の方が自然性に優れた評価が得られた．

個人性の評価　全てのペア間で話者適応モデルの方が個人性で高い評価を得た．5 文による話者適応モデルと 400 文以下の話者依存モデルの比較においては，話者適応モデルの方が有意に高い結果となり，少ない発話で高い適応効果が確認された．

5. 全モデルを適応化した音声の自然性と個人性の評価

先行研究[5] 及び前章での実験結果から，F_0 パターン生成モデルと音素継続時間長モデルをそれぞれ 5 文によって平均値変換したモデルが，400〜470 文で学習した話者依存モデルとほぼ同程度及びそれより優れた自然性と個人性の韻律特徴量を生成できることが確認できた．本章では，音素継続時間長モデルの適応に加えて，F_0 パターン生成モデルとケプストラムモデルについても話者適応を行い，少量の話者音声から合成した音声の自然性と個人性について評価を行う．

5.1 モデル生成の流れ

実験に用いる音声合成モデルの生成には，4.1 節と同様の音声データベースを使用し，同様の条件で STRAIGHT 分析による特徴量抽出を行った．まず，適応対象話者を除く男性話者・女性話者計 7 名の話者の A〜I セット（450 文× 7 話者）と J01〜J20（20 文× 7 話者）の計 470 文× 7 話者の音声データを使用して話者独立なケプストラムモデル，非周期性指標モデル，F_0 パターン生成モデル及び音素継続時間長モデルを学習した．このとき，音素継続時間長の抽出には生成した話者独立ケプストラムモデルを使用した．続いて，適応対象話者の J01〜J20 のランダムな 5 文，20 文を STRAIGHT 分析し，ケプストラム特徴量と F_0 パターンを抽出した．さらにケプストラム特徴量は，話者独立ケプストラムモデルで強制切り出しを行い音素継続時間長を抽出した．その後，抽出した音素継続時間長と F_0 パターンを使用して提案手法で話者適応化し，ケプストラム特徴量を SMAPLR 法で話者適応化した．ただし，非周期性指標は話者適応をせず，音声合成時には話者独立モデルを使用した．比較に用いる話者依存モデルは同話者の A〜I セット（450 文）からランダムに選んだ 80 文・180 文・380 文・450 文と J01〜J20（20 文）の計 100 文・200 文・400 文・470 文を選んで，それぞれケプストラムモデル，非周期性指標モデル，F_0 パターン生成モデル及び音素継続時間長モデルを学習した．その際，音素継続時間長の抽出には生成した話者依存ケプストラムモデルを使用した．

5.2 主観評価実験

生成した話者独立モデルと話者適応モデルを使用して男性話者 2 名（MYI・MTK）と女性話者 2 名（FTK・FKS）の J21〜J53（33 文）について音声合成し，主観評価実験を行った．評価実験は，自然性についてのペアテストと個人性についての ABX テストを行った．ABX テストに用いる正解音声は STRAIGHT 分析にて抽出したケプストラム，非周期性指標，F_0 の特徴量を再合成したものを使用した．被験者は 16 名で音声の受聴にはヘッドフォンを使用した．ペアテストによる自然性の評価の結果を 図 5 に，ABX テストによる個人性の結果を図 6 に示す．以下，自然性及び個人性の評価結果について説明する．

自然性の評価　実験の結果，5 文適応は 100 文学習より優れ，200 文学習とほぼ同程度，400 文・470 文学習よりは劣る評価が得られた，また 20 文適応は 100 文・200 文・470 文学習より優れ，400 文学習とほぼ同程度の評価が得られた．

個人性の評価　20 文適応と 470 文学習とのペア以外は自然性の評価とほぼ同様の結果が得られた．具体的には，5 文適応は 100 文学習より優れ，200 文学習とほぼ同程度，400 文・470 文学習よりは劣る評価が得られた，また 20 文適応は 100 文・200 文学習より優れ，400 文学習とほぼ同程度の評価が得られた．20 文適応と 470 文学習については自然性と異なりほぼ同程度の評価が得られた．

自然性・個人性の評価ともに，5 文による話者適応化では 400 文・470 文学習と比べて十分な適応効果が得られていないことが確認された．F_0 パターン生成モデル及び音素継続時間長モデルの主観評価実験では，それぞれ 5 文による話者適応化で 400 文・470 文学習とほぼ同程度，もしくはそれより優れた音声が合成されていたことから，ケプストラムモデルの適応が 5 文程度では少なく，十分な適応効果が得られなかったためと考えられる．しかし，20 文適応は 400 文・470 文学習とほぼ同程度の評価を得たことから，ケプストラムモデルは 20 文程度あれば十分適応効果が得られることがわかる．この結果から，本音声合成システムでは 20 文の話者音声から適応することで，400〜470 文で学習した話者依存モデルとほぼ同程度の自然性と個人性の音声合成が実現できることが確認された．

6. ま と め

本稿では，少量の話者音声によって自然性と個人性に優れた音声合成を実現するために，先行研究[5] の F_0 パターン生成モデルと同様の手法で音素継続時間長モデルについても平均値変換による話者適応を行い，その性能について客観評価実験及び主観評価実験を行った．客観評価実験を行ったところ，5 文以上による話者適応モデルが 150 文以上の話者依存モデ

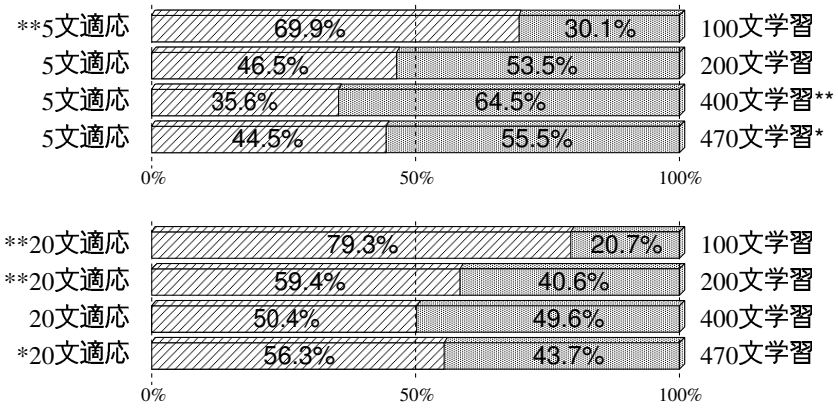

図 5 話者適応モデルと話者依存モデルを比較した自然性のプリファレンススコア．* 及び **印はそれぞれ有意水準 5%, 1%でスコア間に有意差が認められたことを示す．

Fig. 5 Preference scores of synthesized speech produced by speaker-adapted and speaker-trained models in naturalness. "*" and "**" indicate that differences are statistically significant at 5% and 1% significance levels, respectively.

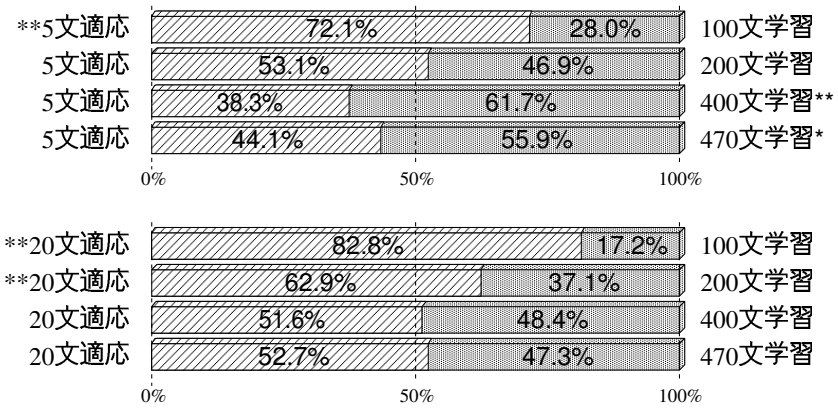

図 6 話者適応モデルと話者依存モデルを比較した個人性のプリファレンススコア．* 及び **印はそれぞれ有意水準 5%, 1%でスコア間に有意差が認められたことを示す．

Fig. 6 Preference scores of synthesized speech produced by speaker-adapted and speaker-trained models in individuality. "*" and "**" indicate that differences are statistically significant at 5% and 1% significance levels, respectively.

ルより推定誤差において劣るものの，主観評価実験では 5 文程度の適応で 470 文で学習した話者依存モデルと同程度の自然性と個人性が実現できることを確認した．さらに，少量の話者音声から音素継続時間長モデルと F_0 パターン生成モデルを平均値変換によって話者適応化し，ケプストラムモデルを SMAPLR 法で話者適応化したモデルを用いて音声合成を行い，自然性と個人性について主観評価実験を行った．その結果，20 文で話者適応化したモデルが 100 文・200 文学習の話者依存モデルより自然性と個人性で優れることを確認し，400 文・470 文学習の話者依存モデルとほぼ同程度であることを確認した．以上の結果より，平均値変換が F_0 パターン生成モデル，音素継続時間長モデルにおいて，自然性と個人性に優れた話者適応法であることを確認した．

今回の実験に用いた音声は朗読調の読み上げ音声であったため，それ以外の感情音声や話し言葉音声合成では，韻律特徴が感情や話者により朗読調の読み上げ音声に比べて大きく変化すること考えられる．そこで，今後の課題としてより多様で自然性と個人性について優れた音声合成を実現するために，様々な感情や話者ごとに韻律特徴量の分析を進め，話者適応化する手法を検討していく必要がある．

参 考 文 献

1) 山田真裕，岩野公司，古井貞熙，“数量化 I 類による F_0 パターン生成の制御要因に関する検討，” 情報処理学会研究報告，vol.2001, no.100, pp.15–20, 2001.
2) 外川太郎，山田真裕，岩野公司，古井貞熙，“HMM 音声合成における数量化 I 類を用いた発話速度制御法，” 秋季音講論，vol.1，pp.345–346，2002.
3) 橘誠，小林隆夫，“平均声モデルを用いる合成音声の話者性とスタイルの同時多様化の検討，” 電気情報通信学会技術研究報告，vol.107，pp.7–12，2007.
4) 田村正統，益子隆史，徳田恵一，小林隆夫，“HMM に基づく音声合成におけるピッチ・スペクトルの話者適応，” 電気情報通信学会論文誌，vol.J85-D-II，pp.545–553，2002.
5) 神山歩相名，篠崎隆宏，岩野公司，古井貞熙，“自然性と個人性に優れた F_0 パターン適応法，” 日本音響学会講演論文集，1–2–7，pp.249–250，2009.
6) C.Hayashi, “On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statical point of view, ”Ann. Inst. Statist. Math. , vol.3, no.2, pp.69–98, 1952.
7) 岩野公司，山田真裕，外川太郎，古井貞熙，“HMM に基づく音声合成における様々な発話速度の実現，” 電子情報通信学会技術研究報告，vol.102, no.292, pp.11–16, 2002.
8) 阿部匡伸，匂坂芳典，梅田哲夫，桑原尚夫，“研究用日本語データベース利用解説書（連続音声データ編），” TR-I-0166，ATR 自動翻訳電話研究所, 1990.
9) H. Kawahara et al.，Speech Communication，vol.27，pp.187-207, 1999.