論文 / 著書情報
Article / Book Information

論題(和文)	自然性と個人性に優れた音声合成のための音素継続時間長適応法
Title(English)	
著者(和文)	神山歩相名, 篠崎隆宏, 岩野公司, 古井貞熙
Authors(English)	Hosana Kamiyama, Takahiro Shinozaki, Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会2010年春季講演論文集, , No. 2-7-1, pp. 329-330
Citation(English)	, , No. 2-7-1, pp. 329-330
発行日 / Pub. date	2010, 3

自然性と個人性に優れた音声合成のための音素継続時間長適応法*

☆神山歩相名，篠崎隆宏（東工大），岩野公司（都市大），古井貞熙 (東工大)

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おり，音声合成の話者適応の研究が行われている [1].

当研究室ではこれまで数量化 I 類によって基本周波数（F_0）と音素継続時間長の韻律制御を行う TTS システム [2][3] を開発してきた．しかし，これらの韻律特徴量の自然性の高いモデル生成には大量の音声データを必要としてきた．そのため，応用の観点からは特定の話者からの少量のデータを使って，その話者の特徴を取り込んだ音声を合成することが望まれている．

我々は以前，F_0 パターン生成モデルを数量化 I 類の平均値を変換する話者適応法を提案した [4]．しかし，音素継続時間長モデルについての適応法は未だ検討されていなかった．そこで本研究では，音素継続時間長モデルにおいても，同様の平均値変換による話者適応を行いその性能について評価する．具体的には，複数の話者による大量の音声から学習した音素継続時間長モデルの平均値成分を，少量の特定話者データから推定した音素継続時間長の平均値に変換する．本稿では，まず数量化 I 類による音素継続時間長制御法と平均値変換による話者適応法について説明し，ついで本手法の客観評価実験と主観評価実験の結果について述べる．

2 数量化 I 類による音素継続時間長モデル

数量化 I 類とは，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．

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

$\bar{y}$ 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．

本研究における数量化 I 類の目的変数は音素継続時間長となる．音素継続時間長は，ケプストラム特徴量の triphone HMM で強制切り出しすることで抽出を行う．音素継続時間長は，当該音素の種類と当該音素の先行・後続する音素の種類，当該音素の 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 [5]．本研究においてもそれらの制御要因によって数量化 I 類のモデル化を行った．音素継続時間長モデルは，Table 1 に示す 13 種類の音素クラスごとに学習する．

Table 1　音素クラスの一覧

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

3 平均値成分の変換による話者適応法

数量化 I 類による音素継続時間長モデルを，平均値成分の変換によって話者適応する手法について述べる．本手法は，話者独立モデルの平均値成分を適応対象の話者に合うように置き換えることでモデルを生成する．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ように次の式で求める．

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

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

$\hat{y}'_i$ は，適応対象話者についての i 番目データの推定値，y'_i はサンプル値，N' はサンプル数である．この操作を 13 の音素クラスのモデルそれぞれに対して行い，平均値成分を置き換える．ただし，適応データの中に該当する音素クラスが存在しない場合は話者独立音素継続時間長モデルの平均値成分をそのまま使用した．

4 評価実験

4.1 使用データ

実験は ATR 日本語音声データベース中の男性話者 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．音素継続時間長の抽出に用いるケプストラム HMM 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

初期モデルとして使用する話者独立モデルは，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）を用いて学習した．続いて，合成対象話者の A〜I セット（450 文）と J01〜J20（20 文）からランダムに 1〜100 文を選んで本手法で話者適応を行った．比較に用いる話者依存モデルは，A〜I セット（450 文）と

*A phoneme duration adaptation method for achieving speech synthesis with naturalness and individuality by Hosana Kamiyama, Takahiro Shinozaki (Tokyo Institute of Technology), Koji Iwano (Tokyo City University) and Sadaoki Furui (Tokyo Institute of Technology)

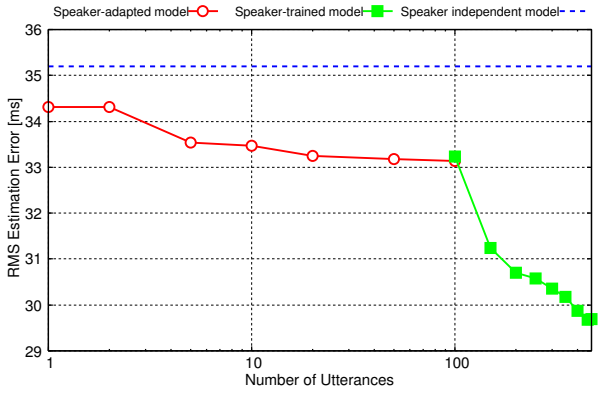

Fig. 1　使用文数と推定誤差の関係

J01〜J20（20 文）からランダムに 100〜470 文を選んで学習した．生成した話者独立モデル，話者適応モデル，話者依存モデルを用いて J21〜J53（33 文）について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，推定誤差についての客観評価実験を行った．Table 1 の モーラを決定付ける 1〜4 の音素クラスについて適応文数と推定誤差の関係を求め，本手法の適応効果について調査した．結果を，Fig. 1 に示す．実験の結果，適応効果は 5 文でほぼ飽和しており，100 文の話者依存モデルと同程度の推定誤差となった．150 文以上の話者依存モデルは，適応したモデルより推定誤差が小さく，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良いと考えられる．5 文で適応した場合と 470 文で学習したモデルとの推定誤差の差は，およそ 3.5 [ms] 程度であった．よって，本手法による話者適応は 150 文以上の学習には劣るものの，5 文程度による適応によって，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することができると言える．

4.4　主観評価実験

続いて，音素継続時間長モデルについて提案手法で適応した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．音声合成に用いるケプストラム，非周期性指標，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）で学習したモデルから生成した．結果のプリファレンススコアにおいて 5%，1%の有意水準で有意に高い評価が得られた方を *, **印で示す．被験者は 11 名である．

4.4.1　話者独立モデルとの比較

適応効果について調べるため，話者独立モデルと話者適応モデルの比較を行った．今回の実験は比較的話速の速い話者（MYI・MMY）と遅い話者（FTK・MTK）2 名ずつについて音声を合成した．その後，自然性についてはペアテスト，個人性については ABX テストを行った．ABX テストの正解音声は，正解値の音素継続時間長を使用した．結果を Fig. 2 に示す．

自然性の評価　話速の速い話者遅い話者とも話者独立モデルの方が高い評価が得られた．特に総合では 5%の有意水準で自然性が劣化していることが確認された．

個人性の評価　話速の速い話者においては，5 文適応では有意差が現れなかったが，話速の遅い話者，及び総合においては 1%の有意水準で適応効果が認められた．

4.4.2　話者依存モデルとの比較

続いて，話者適応モデルと特定話者の音声で学習した話者依存モデルの比較を行った．音声は 6 名の

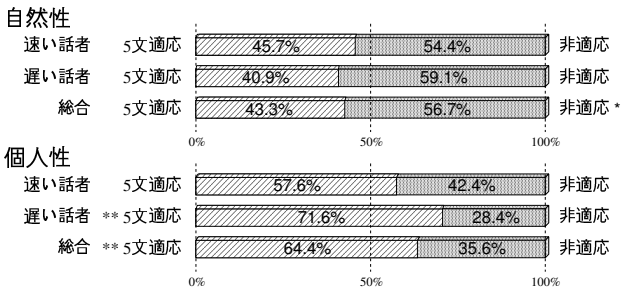

Fig. 2　話者独立モデルと比較した自然性と個人性の評価

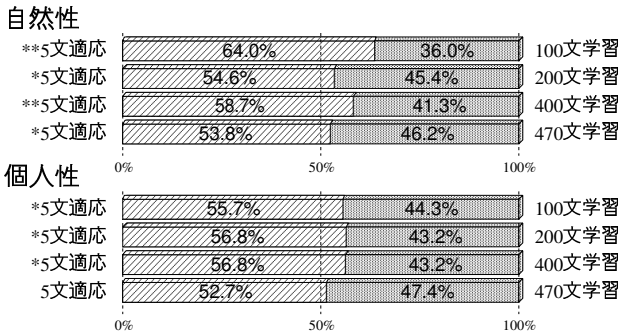

Fig. 3　話者依存モデルと比較した自然性と個人性の評価

話者（FKS・FTK・FYM・MYI・MMY・MTK）について合成した．その後，自然性についてはペアテスト，個人性については ABX テストを行い，正解音声は 4.4.1 節と同様の条件で合成した．結果を Fig. 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高い評価を得ることができた．話者独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同程度の自然な音声が合成でき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高い評価を得ることができた．特に，5 文による話者適応モデルが 400 文以下の話者依存モデルより有意に高かった．本手法による話者適応モデルは 470 文の音声で学習した話者依存モデルと同程度の個人性が実現できることが確認された．

5　まとめ

本稿では自然性が高くかつ個人に適応した音素継続時間長モデルを少量の音声データから作成するため，数量化 I 類の平均値成分の変換による話者適応法を行った．客観評価実験を行ったところ，150 文以上を用いた話者依存モデルよりは推定誤差が劣るものの，5 文程度の適応で効果が得られることを確認した．また，主観評価実験によって，本手法による話者適応の効果が認められ，5 文を用いた適応で，470 文で学習したモデルと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確認した．

今後の課題として，感情音声や話し言葉音声など読み上げ音声以外における韻律特徴量の分析を進め，適応する手法を検討する必要がある．

参考文献

[1] M.Tachibana et al., Proc. ICASSP, 2008.
[2] 山田 他，音講論，1-2-8, 2001.
[3] 外川 他，音講論，3-10-9，2002.
[4] 神山 他，音講論，1-9-22，2009.
[5] 岩野 他，信学技報，SP2002-73，2002.