論文 / 著書情報
Article / Book Information

論題(和文)	自然性と個人性に優れたF0パターン適応法
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2009年秋季講演論文集, , No. 1-2-7, pp. 249-250
Citation(English)	, , No. 1-2-7, pp. 249-250
発行日 / Pub. date	2009, 9

自然性と個人性に優れた F_0 パターン適応法*

◎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 [1][2] を開発してきたが，これらの韻律特徴量のモデル学習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．

F_0 は声の高さに対応し合成音の話者性を特徴付ける基本的な要素であるとともに，その変化は発話の抑揚に対応し合成音声の自然性や聞き取りやすさに大きく影響する．そのため本研究では，複数の話者による大量の音声から F_0 パターン生成モデルを学習し，さらに話者性を取り込むために少量の特定話者データから推定した F_0 平均値を用いる適応法を提案する．この手法は，複数の話者で F_0 パターン生成モデルを学習することで，日本語（標準語）として自然なイントネーションが学習され，特定話者に合わせて平均値を置換することで自然性が高くかつ個人に適応したモデルが作成できると考えられる．

本稿は，まず数量化 I 類による韻律情報制御法と提案する平均値置換手法について説明を行い，ついで本手法によって平均値置換したモデルの推定誤差の調査と主観評価実験の結果について述べる．

2 数量化 I 類を用いた F_0 パターン制御法

高度な韻律制御法として，統計的手法である数量化 I 類を利用した F_0 パターン制御法が提案されている [3]〜[5]．本研究でも，これらの手法と同様に数量化 I 類を用いた韻律制御を行う．

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．

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

$\bar{y}$ は全データの平均値，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．

F_0 パターンをモデル化する際の目的変数は，各モーラの母音，撥音，長音の中心時刻における F_0 値を，次式で対数変換した値（semitone）である．

$$p = 12\log_2(F_0/55) \quad (2)$$

合成時にはモーラごとの semitone 値を推定し，直線補間することで文全体の F_0 パターンを生成する．

推定対象モーラを M，モーラ M を含むアクセント句を W とする．モーラ M が，アクセント句 W の第 n モーラであるとすると，数量化 I 類のモデルは $n = 1, 2, 3, 4$，および 5 以上の，5 つの場合に分けて作成する．制御要因については，Table 1 に示す 27 個を用いた．Table 1 において P は，アクセント句 W が属する呼気段落（ポーズで区切られる音声区間）である．

Table 1 数量化 I 類による F_0 パターン推定に用いる制御要因．括弧内はカテゴリ数．

1	W のモーラ数（8）
2/3	P 内で W に先行/後続するモーラ数（9）
4/5/6	先行/当該/後続アクセント句のアクセント型（7）
7	P 内で W に先行するアクセント核を有する句の数（4）
8/9	W 前/後の音調結合の強さ（4）
10/11	W 前/後の句境界のポーズの長さ（9）
12	トーンパタン（5〜10：n により異なる．）
13	当該音素（母音，撥音，長音）の種類（8）
14/15	13 の音素の前/後の音素の種類（13）
16	M の W 内のモーラ位置（$n \geq 5$ の場合）（6）
17	M のアクセント句の位置（$n \geq 5$ の場合）（6）
18/19	W 前/後の句境界のポーズの長さ（9）
20/21	P 内で W に先行/後続するモーラ数（9）
22/23	W の 2 つ前/後の音調結合の強さ（5）
24/25	W の 3 つ前/後の音調結合の強さ（5）
26/27	13 の音素の 2 つ前/後の音素の種類（6）

3 平均値置換による F_0 パターン適応法

数量化 I 類による F_0 パターン生成モデルを，平均値置換によって話者適応する手法について述べる．まず話者独立な F_0 パターン生成モデルを，様々な話者のデータを混合して学習する．提案手法はこの話者独立 F_0 パターン生成モデルの平均値を，適応対象の話者に合うように値を置き換えることでモデル作成する．新しい平均値 $\bar{y}'$ は，推定二乗誤差を最小化するように次の式で求める．

$$\frac{\partial E}{\partial \bar{y}'} = \frac{\partial}{\partial \bar{y}'} \sum_i (\hat{y}'_i - y'_i)^2 = 0 \quad (3)$$

$$\Rightarrow \bar{y}' = \frac{1}{N'} \sum_i (y'_i - \sum_f \sum_c x_{fc}\delta_{fc}(i)) \quad (4)$$

$\hat{y}'_i$ 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，モーラ位置によって分けられた 5 つのモデルそれぞれに対して平均値を置き換えた．カテゴリースコア x_{fc} は話者独立 F_0 パターン生成モデルのものを使う．

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．F_0 は，STRAIGHT 法 [6] で窓幅 40ms フレーム周期 1ms で抽出した．

*An F_0 contour adaptation method for achieving naturalness and speaker individuality by Hosana Kamiyama, Takahiro Shinozaki (Tokyo Institute of Technology), Koji Iwano (Tokyo City University) and Sadaoki Furui (Tokyo Institute of Technology)

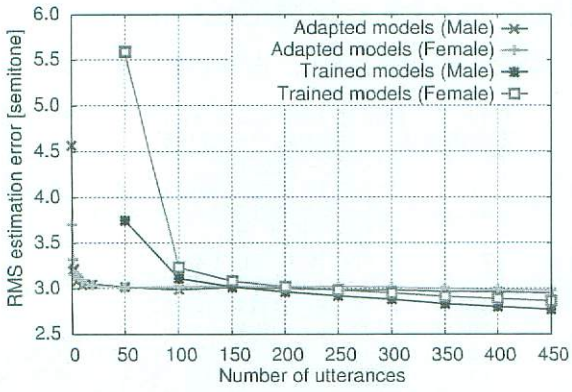

Fig. 1　使用文数と推定誤差の関係

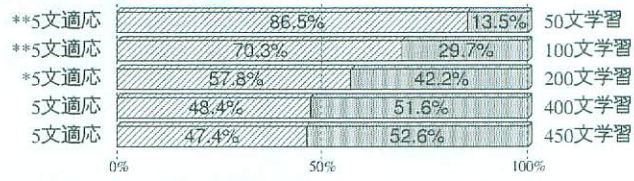

Fig. 2　自然性のプレファレンススコア

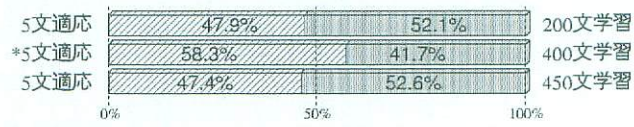

Fig. 3　話者性のプレファレンススコア

4.2　適応文数と推定誤差の調査

まず，本手法による適応法に必要な文数と推定誤差について調査をした．実験では，始めに適応対象の話者を除く各話者の A〜I セット（450 文）を用いて，男性話者 3 名による話者独立モデルと，女性話者 3 名による話者独立モデルを学習した．続いて，適応対象話者の A〜I セットから 1〜450 文をランダムに選んで本手法によって適応した．このとき，適応対象の話者が男性話者の場合は男性話者のみで学習した話者独立モデルから，適応対象の話者が女性話者の場合は女性話者のみで学習した話者独立モデルから適応を行った．また，50 文以上用いる場合については同様のデータを用いて学習したモデルを求めた．その上で，適応対象話者の J セット（53 文）に対する推定誤差を求めた．

使用文数と推定誤差の関係を Fig. 1 に示す．推定誤差は男性，女性モデルそれぞれについての話者平均値で示す．推定誤差は男性，女性とも適応データが 5 文のとき約 3.1[semitone] になり，その後はほぼ変化はなかった．また，5 文以上を用いて適応したモデルと 450 文を用いて学習したモデルの推定誤差にはほとんど差がなかった．したがって，本手法では 5 文程度で十分な適応効果が得られることが確認された．

4.3　主観評価実験

本手法によって適応したモデルの自然性と話者性について主観評価を行った．話者独立なモデルは，適応対象話者を除く男性話者女性話者 7 話者の A〜I セットを用いて学習した．

主観評価は，A〜I セットのうちから 50 文，100 文，200 文，400 文，450 文をランダムで選んで学習したモデルによる合成音と，ランダムな 5 文で適応したモデルによる合成音について比較を行った．音声は，男性話者 2 名（MHT, MYI）女性話者 2 名（FKS・FTK）について，J セットの文について合成し，このとき，ケプストラムと音素継続時間長，非周期性指標は合成する話者から抽出した値（正解値）を用いた．被験者は 12 名である．

4.3.1　自然性の評価

まず，被験者には学習モデルと適応モデルによる合成音のペアをランダムな順で提示し，どちらが自然性が高いかを評価してもらった．この主観評価の結果を Fig. 2 に示す．図中の *, **印は，二項分布に基づく有意検定を行ったときに，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している．これより，50 文・100 文で学習したモデルより 5 文で適応したモデルで推定した F_0 パターンの方が自然性で優れることが確認された．また，400 文・450 文で学習したモデルと 50 文で適応したモデルで推定した F_0 パターンに有意差は現れなかった．これより本手法によって適応したモデルによる合成音は，特定話者による大量の音声で学習したモデルとほぼ同程度の自然性が得られたと言える．

4.3.2　話者性の評価

続いて，被験者に学習モデルと適応モデルによる合成音のペアをランダムな順で提示した後，正解音声を提示して 2 つの音声から話者性が正解音声に近い方を選ぶ ABX テストを行った．この主観評価の結果を Fig. 3 に示す．図中の *印は，二項分布に基づく有意検定を行ったときに，有意水準 5% でスコア間に有意差が認められたことを示している．実験の結果，5 文適応モデルと 200 文，450 文学習モデルはほぼ同程度の評価が得られている．有意差がみられた 5 文適応と 400 文学習モデルの比較においても，5 文適応の合成音の方が高い評価を得ている．よって本手法における適応法によるモデルは，特定話者による大量の音声で学習したモデルと比較して十分な話者性が実現されていると言える．

5　まとめ

本稿では自然性が高くかつ個人に適応した F_0 パターン生成モデルを少量の音声データから作成するため，数量化 I 類の平均値を置換する話者適応法を提案した．この手法によって適応したモデルについて推定誤差を調査したところ，5 文程度で十分な適応が可能であることが確認できた．また主観評価実験を行ったところ，400〜450 文で学習したモデルによる合成音と 5 文で適応したモデルによる合成音で，ほぼ同程度の自然性と話者性が認められた．これより，適応手法が自然性と個人性に優れた話者適応法であることが確認された．

今後の課題としては，今回は F_0 の平均値のみの置換を行い，数量化 I 類のカテゴリースコアで表現される話者性は適応しなかったため，話者性や自然性に影響を与えるカテゴリーについてさらに調査し，適応する必要性について検討することが挙げられる．また，数量化 I 類による音素継続時間長モデルの適応法についても検討していく必要がある．

参考文献

[1] 山田 他，音講論，1-2-8, 2001.
[2] 外川 他，音講論，1-2-8, 2002.
[3] 酒寄 他，音講論，pp.245–246, 1986.
[4] 阿部 他，音学誌，vol.49, No.10, pp.682-690, 1993.
[5] 海木 他，信学論，vol.J83-D-II, no.9, pp.1853–1860, 2000.
[6] H. Kawahara et al., Speech Communication, vol.27, pp.187-207, 1999.