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

山田 真裕　　岩野 公司　　古井 貞熙

東京工業大学大学院 情報理工学研究科 計算工学専攻
〒 152-8552 東京都目黒区大岡山 2-12-1
Email: {masahiro, iwano, furui}@furui.cs.titech.ac.jp

我々はテキスト音声合成システムの構築を進めている．その際，高精度な韻律制御を実現するため，統計的手法である数量化I類を用いた2種類のF_0パターン制御を実装した．そこで，これらの手法を主観評価実験により比較し，さらに，評価の高かった手法における制御要因の組み合わせと主観評価の関係を調べることで，どのような制御要因が自然なF_0パターンの生成に有効であるか検討した．その結果，トーンパタン，音素の種類，モーラ数，ポーズの長さなどに関する制御要因が，合成音声のF_0パターンの知覚上，特に重要であることが示され，モーラ・音素を単位とする情報も評価に影響を及ぼしていることがわかった．

A Study on F_0 Contour Generation Factors Using Categorical Multiple Regression

Masahiro Yamada, Koji Iwano, and Sadaoki Furui

Department of Computer Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
Email: {masahiro, iwano, furui}@furui.cs.titech.ac.jp

We have been developing a text-to-speech system. In order to realize high quality prosodic control, we implemented two statistical F_0 contour control methods using categorical multiple regression, and then compared these methods through subjective evaluation experiments. Furthermore, to research what kind of F_0 contour generation factors are effective for generating natural F_0 contours, we investigated relationship between combination of factors and subjective quality. Experimental results show that the following factors are particularly important for the perception of F_0 contours: tone patterns, kind of phonemes, number of mora, and pause length. This means that factors represented for mora/phoneme units affect the subjective quality.

1　はじめに

近年，さまざまな分野で音声合成の技術が用いられてきている．例えば，新聞や電子メールの読み上げといったものが挙げられるが，これらのような，読み上げるテキストの内容が限定されない場合は，テキストから自動的に音声を合成する必要がある．当研究室でも，任意のテキストを読み上げることのできるテキスト音声合成システムの構築を進めている．合成音声には高い自然性が望まれるが，人間の発声に近い自然なイントネーションをもつ合成音声を生成するためには，韻律的特徴の適切な制御が特に重要となる．そのため高精度な韻律制御として，数量化I類を用いてF_0パターンを統計的にモデリングする手法がすでに提案されている [1] [2] [3]．そこで本研究では，まず，これらの手法をもとに2種類のF_0パターン制御を実装する．さらに，実装した手法を主観評価実験により比較し，その性能

について検討を行う．具体的には，数量化 I 類における制御要因の組み合わせと主観評価の関係を調べることで，どのような制御要因が自然な F_0 パターンの生成に有効であるか検討する．

2　テキスト音声合成システム

本研究で用いるテキスト音声合成システムの構成を図 1 に示す．

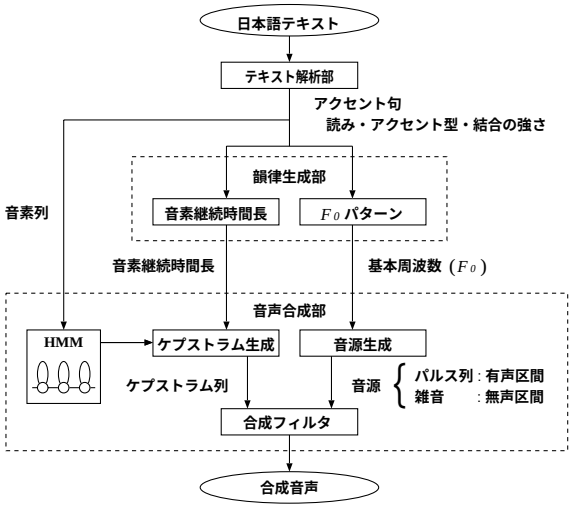

図 1: テキスト音声合成システムの構成

テキスト解析部では，入力となる日本語テキストを解析して言語情報を出力する．韻律生成部では，この情報から音素継続時間長，F_0 パターンといった韻律を生成する．さらに音声合成部において，HMM を用いてケプストラム列を生成，また，基本周波数情報から音源を生成し，これらを合成フィルタにかけることにより，合成音声を生成する．

以下では，テキスト解析部と音声合成部について説明する．

2.1　テキスト解析部

テキスト解析部では，NTT-IT 社のテキスト音声合成ソフトウェア「Hipervoice」[4] を用いて，任意の漢字かな混じり文をカナアクセント文に変換する．

カナアクセント文は 1 個以上のアクセント句から構成される．アクセント句は，読み，後続アクセント句との結合の強さ（音調結合またはポーズ），およびアクセント型から構成される．音調結合は，弱結合 ('/')，強結合 ('*')，ポーズは，短ポーズ (' ')，中ポーズ (',')，長ポーズ ('.') である [4]．一般にアクセント句は弱結合でつながり，後続句が弱いアクセント核をもつ場合は，強結合でつながる．

以下にテキスト解析の例を示す．

入力 : あらゆる現実を，すべて自分のほうへねじ曲げたのだ．

出力 : アラユル [/03] **ゲンジツオ** [,00] **スベテ** [01] **ジブンノ** [*00] **ホーエ** [/01] **ネジマゲタノダ** [.05]

なお，合成部では，中ポーズ，長ポーズのみに無音区間を割り当てている．

2.2　音声合成部

文献 [5] では，連続混合分布型 HMM で表現された音素モデルから尤度が最大となる音響パラメータ系列を生成し，これに適当な音源を与え，MLSA フィルタ [6] を用いて音声合成する手法が提案されている．当研究室では，この枠組において音素の持続時間を任意に与えたときに最尤の音響パラメータ系列を生成する手法を提案している [7]．本研究では，これらの手法を組み合わせて音声合成を行った．音響パラメータとして，フレーム長 25.6ms，フレーム周期 12.8ms で分析した 24 次のメルケプストラム，およびそのデルタ係数を用い，総状態数 1527，3 ループ 5 状態 4 混合 tied-state triphone の HMM を作成した．音源には，基本周波数情報をもとに，有声区間ではパルス列を，無音区間では白色雑音を用いている．

3　F_0 の推定

F_0 の推定には，次式に示すような，説明変数（制御要因）に定性的データを用いた線形重回帰手法である数量化 I 類 [1] [2] [3] [8] を

用いる．

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

ここで，$\hat{y}_i$ は i 番目のデータの推定値，$\bar{y}$ は全データの平均値，N はデータ数である．x_{fc} は制御要因 f のカテゴリ c の数量，$\delta_{fc}(i)$ は i 番目のデータの制御要因 f がカテゴリ c をとるときに1，それ以外のときに0を与える関数である．

以降では，システムに実装した数量化I類による2種類の F_0 パターン生成法について説明する．1つ目はアクセント句の F_0 パターンが決まったモデルに従うと仮定しパターンをあてはめる方法（以下，モデルベース），2つ目はモーラごとに F_0 の値を推定する方法（以下，モーラベース）である．

なお，基本周波数 (F_0) は，次式を用いて対数変換している (semitone)[9]．

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

3.1 モデルベース

この手法では，図2の台形点ピッチ近似モデル [10] を用いて，アクセント句ごとに自然音声の F_0 パターンから最小2乗近似により以下の値を求め，これを学習時の目標値とする．
・第1モーラの特定部分における F_0 の値 (Fs)
・最終モーラの特定部分における F_0 の値 (Fe)
・ストレス量 (S)

ここで，特定部分とは，各モーラを構成する特定の音素 (a, i, u, e, o, N, Q, :(: は長音)) の中心部のことである．

用いる制御要因とそのカテゴリは表1に示す通りである．() 内は制御要因のカテゴリ数である．ここでは，ポーズで区切られる区分を F_0 パターンの立て直しの制御単位（MG）としている．

なお，前後のポーズの有無により4つの場合に分けて推定した．

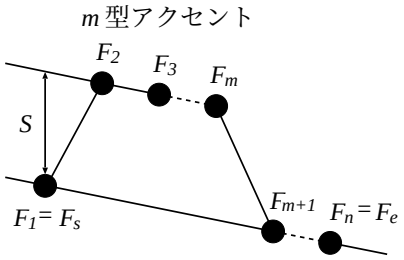

図 2: 台形点ピッチ近似モデル

表 1: 用いる制御要因とそのカテゴリ

1 当該アクセント句のモーラ数
▷ 2 以下，3，4，5，6，7，8，9 以上 (8)

2/3 当該アクセント句の属す MG 中における，当該句に先行/後続するモーラ数
▷ 2 以下，3，4，5，6，7，8，9，10 以上 (9)

4 先行アクセント句のアクセント型
▷ 文頭，平板型，1 型，2 型，3 型，4 型，その他 (7)

5 当該アクセント句のアクセント型
▷ 平板型，1 型，2 型，3 型，4 型，5 型，その他 (7)

6 後続アクセント句のアクセント型
▷ 文末，平板型，1 型，2 型，3 型，4 型，その他 (7)

7 当該アクセント句の属す MG 中において，当該句に先行する，アクセント核をもつ句の数
▷ 0，1，2，3 以上 (4)

8/9 当該アクセント句の前/後の結合の強さ
▷ 文頭/文末 or ’,’，’ ’，’/’，’*’ (4)

10/11 当該アクセント句の前/後の句境界の長さ（ポーズの長さ，単位 : [ms] ）
▷ <100, <200, <300, <400, <500, <600, <700, <800, $800\leq$ (9)

12/13 当該アクセント句の 2 つ前/後の結合の強さ
▷ なし or 文頭/文末，’,’，’ ’，’/’，’*’ (5)

14/15 当該アクセント句の 3 つ前/後の結合の強さ
▷ なし or 文頭/文末，’,’，’ ’，’/’，’*’ (5)

3.2 モーラベース

各モーラごとに，特定の音素 (a, i, u, e, o, N, Q, :) の中心部における F_0 の値を推定する．

制御要因は，モデルベースの手法で用いる

制御要因（ここでは,「当該アクセント句」などを「当該モーラの属すアクセント句」などといったように置き換える）のうち,「当該モーラの属すアクセント句のアクセント型」（制御要因5）以外のものに加え，表 2 に示す要因を用いる．ただし，制御要因「当該モーラの属すアクセント句のモーラ数」（制御要因1）のカテゴリを表中のように変更する．ここで，当該音素とは，当該モーラに含まれる特定の音素 (a , i , u , e , o , N , Q , :) のことである．

なお，アクセント句内のモーラ位置により 5 つ（1，2，3，4，5 以上）の場合に分けて推定した [2].

表 2: 用いる制御要因とそのカテゴリ

1 当該モーラの属すアクセント句のモーラ数
- ▷ 第 m モーラの場合
  - · (m+1) 以下，m+2，⋯，m+7，(m+8) 以上 ： $1 \leq m \leq 4$ (8)
  - · 6 以下，7，⋯，12，13 以上 ： $m \geq 5$ (8)

16 トーンパタン（先行・当該・後続モーラのトーンの 3 つ組） [2]
- ▷ 第 1 モーラの場合
  L*LH，H*LH，L*HL，H*HL，
  1 モーラアクセント句 (5)
- ▷ 第 2 モーラの場合
  HLL，HLL*，HLH*，LHL，LHH，
  LHL*，LHH* (7)
- ▷ 第 3 モーラ以降の場合
  HHL，HHH，HHL*，HHH*，HLL，HLL*，
  HLH*，LLL，LLL*，LLH* (10)

注: L，H はアクセント句内のトーンを，L*，H*はアクセント句が異なる場合のトーンを示す [2].

17 当該音素の種類
- ▷ a，i，u，e，o，N，Q，：(8)

18/19 当該音素の先行/後続音素の種類
- ▷ 母音 (a, i, u, e, o)，長音 (:)，促音 (Q)，無音 (#)，有声閉鎖音 (b, d, g)，無声閉鎖音 (p, t, k)，有声破擦音・摩擦音 (z, j)，無声破擦音 (ch, ts)，無声摩擦音 (f, h, s, sh)，鼻音 (N, m, n)，弾音 (r)，半母音 (w, y)，拗音 (by, dy, gy, py, ky, hy, ry, my, ny) (13)

20 アクセント句内のモーラ位置（句内のモーラ位置が 5 以上の場合のみ）
- ▷ 5，6，7，8，9，10 以上 (6)

21/22 当該音素の 2 つ前/後の音素の種類
- ▷ 母音 (a, i, u, e, o)，長音 (:)，撥音 (N)，促音 (Q)，無音 (#) or なし，その他 (6)

各アクセント句の F_0 パターンは，合成時に，各モーラごとの F_0 の値を直線補間する（点ピッチモデル）ことにより生成する．

4 評価実験

実験には ATR 日本語音声データベースの音素バランス文 503 文 [11]（話者 MHT）を用いた．このデータベースには，FFT ケプストラム法により自動抽出した F_0 を視察によって修正したデータが付与されており，これを学習時の正解の F_0 とした．学習には 493 文を用いた．なお，テキスト解析で誤りのあったものは手動で修正した．

各被験者に提示する評価文には，学習に用いた 493 文のうちテキスト解析で誤りのなかったものからランダムに 5 文 (close)，学習に用いていない残りの 10 文からランダムに 5 文 (open) を選択し，さらに使用データベース以外の文として，新聞記事から 5 文 (news) を用いた．

以下の実験では，被験者に合成音声の文を提示した上で，各方式で合成した音声を各文ごとに順序をランダムに入れ換えて提示し，自然に聞こえるかどうかを 5 段階（1. 非常に悪い 2. 悪い 3. 普通 4. よい 5. 非常によい）で評価してもらった．合成音声はヘッドホンを用いて提示した．被験者数は 10 名である．

各音素の継続時間長は，音声データの存在する close，open については合成部で用いる HMM による強制切り出し結果を利用し，news については，学習データから得られた，文脈を考慮した音素の平均時間長とした．具体的には，強制切り出し結果を用いて，データベース中の全 quinphone（前後 2 音素を考慮した音素）の平均時間長 (t_1) と全 monophone の平均時間長 (t_2) をあらかじめ求めておく．そして，同一の quinphone がある場合は t_1 を，ない場合は t_2 を各音素の継続時間長とした．

4.1 手法間の比較

前節で述べた数量化 I 類を用いた 2 種類の手法の比較・検討を行う．この際，統計的手法であることの有効性も検証するため，比較用に，ヒューリスティックなルールにより話調成分やアクセント成分の大きさなどを決める手法（以下，ルールベース）も実装し，これら計 3 種類の F_0 パターン生成法により作成した合成音声を用いて，被験者による比較評価実験を行った．

ルールベースの手法では，各 MG（F_0 の制御単位）に対して，結合の強さに応じて（強結合の場合，後続アクセントは押さえられることから），各アクセント句の各モーラのアクセントの大きさを 3 段階（高，中，低）に設定する．始端の F_0 と終端の F_0（話調成分に相当），アクセントの大きさといった値には経験的に定めた値を用い，点ピッチモデルにより F_0 パターンを生成する．

テキストには close，open，news を用いた．表 3 に実験結果を示す．

表 3: 3 手法の比較結果（スコアの平均）

手法	close	open	news	総合
ルールベース	2.94	2.78	2.68	2.80
モデルベース	2.98	3.50	3.08	3.19
モーラベース	3.80	3.82	3.42	3.68

close，open，news すべてにおいて，モーラベースの手法が最も高いスコアとなった．総合スコアの母平均の差について，危険率 1%で検定したところ，3 手法間のいずれにも有意差が認められた．

この結果から，ヒューリスティックなルールよりも，数量化 I 類を用いた 2 種類の手法のほうが有効であるといえ，統計的な手法を用いた効果が確認できる．さらに，モデルベースとモーラベースの手法を比較すると，モーラベースの手法がより有効であると判断できる．そこで，その理由を調べるため，この手法で用いる制御要因の影響について以降で検討する．

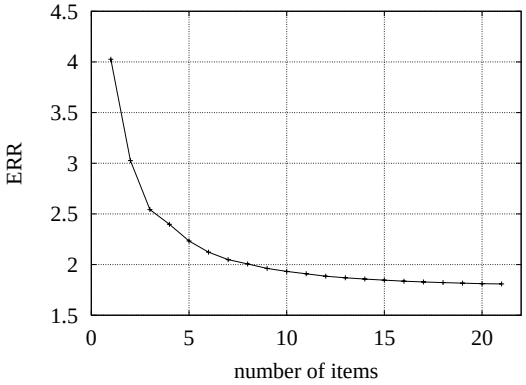

図 3: 制御要因数と誤差のグラフ

4.2 制御要因の組み合わせによる比較

モーラベースの手法において，制御要因が主観評価に与える影響を調べた．

21 個の制御要因から，最も重要でない制御要因（除いたときに誤差の増加が最小である制御要因）を順に除いて，制御要因数を減らしていったときの推定誤差（各モーラごとの推定値 (semitone) の平均誤差）の推移を図 3 に示す．除かれた制御要因は順に，4,14,21,6,22,15,7,10,9,17,12,20,3,8,18,19,13,1,2,11,16 となった．

制御要因の組み合わせと主観評価の関係を調べるため，誤差がほぼ等間隔となるような点を選び，そのときの制御要因数での学習結果を用いて作成した合成音声の比較評価実験を行うことにした．実際に点があるところだけに限ると，そのときの制御要因数は，1,2,3,5,8,21 となる．この 6 つの場合で作成した合成音声の比較評価実験を行った．テキストには close，open を用いた．表 4 に実験結果を示す．

総合では，制御要因数が 21 の場合が最も高いスコアとなった．総合スコアの母平均の差について，危険率 1%で検定したところ，制御要因数が 8 と 21 の場合を除いたすべての要因数間に有意差が認められた．8 と 21 の間には有意差が認められなかったことから，重要度の高さが 8 番目までの制御要因が，合成音声の F_0 パターンの知覚上，特に重要であると考

表 4: 制御要因数の違いによる比較結果（スコアの平均）

制御要因数	close	open	総合
1	1.96	2.06	2.01
2	2.70	2.78	2.74
3	3.26	3.18	3.22
5	3.72	3.42	3.57
8	4.00	3.90	3.95
21	4.08	3.88	3.98

えられる．具体的には，トーンパタン，音素の種類，モーラ数，アクセント句の結合の強さ，ポーズの長さに関する制御要因が重要である．このうち，トーンパタン，音素の種類に関する制御要因は，モーラベースの手法でのみ用いることができるものである．したがって， モデルベースとモーラベースの手法の評価に差が出たのは，このようなモーラ・音素を単位とする詳細な情報を利用していることが一因であると考えられる．

5 まとめ

2 種類の数量化 I 類を用いた F_0 パターンの統計的モデリングを行い，特にモーラベースの手法の有効性を確認した．さらに，これまでは制御要因の組み合わせに関する詳しい検討がほとんど行われていないため，この手法において，制御要因の組み合わせと主観評価の関係を調べた．その結果，トーンパタン，音素の種類，モーラ数，アクセント句の結合の強さ，ポーズの長さに関する制御要因が，合成音声の F_0 パターンの知覚上，特に重要であることがわかった．このうち，トーンパタン，音素の種類に関する制御要因は，モーラベースの手法でのみ用いることができるものであり，モデルベースとモーラベースの手法間の評価の差は，このようなモーラ・音素を単位とする詳細な情報を利用していることに起因していると考えられる．

今後は，音素継続時間長の制御の検討を行っていく．

参考文献

[1] 海木延佳, 勾坂芳典, “局所的句構造に基づく F_0 制御”, 信学論, Vol.J83-D-II, No.9, pp.1853–1860 (2000-9).

[2] 阿部匡伸, 佐藤大和, “音節区分化モデルに基づく基本周波数の 2 階層制御方式”, 音響誌, Vol.49, No.10, pp.682–690 (1993-10).

[3] 酒寄哲也, 佐々部昭一, 北川博雄, “規則合成のための数量化 I 類を用いた韻律制御”, 音講論, 3-4-17 (1986-10).

[4] `http://www.ntt-it.co.jp/goods/cts/onsei/hiper-v.html`

[5] 益子貴史, 徳田恵一, 小林隆夫, 今井 聖, “動的特徴を用いた HMM に基づく音声合成”, 信学論, Vol.J79-D-II, No.12, pp.2184–2190 (1996-12).

[6] 今井 聖, 住田一男, 古市千枝子, “音声合成のためのメル対数スペクトル近似 (MLSA) フィルタ”, 信学論 (A), Vol.J66-A, No.2, pp.122–129 (1983-2).

[7] 立和 航, 古井貞熙, “HMM による規則音声合成の検討”, 音講論, 2-3-7 (1999-3).

[8] 箱田和雄, 中嶌信弥, 広川智久, “文章音声の音調結合型導出規則の検討”, 信学技報, SP89-5 (1989-5).

[9] 小坂直敏, “文音声の発話方法と韻律情報との関係”, 音講論, 2-4-11 (1990-3).

[10] 箱田和雄, 中嶌信弥, 広川智久, “台形点ピッチ近似モデルによる文章音声のピッチパタン制御”, 音講論, 1-2-14 (1988-10).

[11] 阿部匡伸, 勾坂芳典, 桑原尚夫, “言語・韻律情報を持つ連続音声の基本周波数データベース”, 音講論, 2-3-22 (1989-10).