論文 / 著書情報
Article / Book Information

論題(和文)	数量化Ｉ類を用いたピッチパターン生成における制御要因の検討
Title(English)	
著者(和文)	山田 真裕, 岩野 公司, 古井 貞熙
Authors(English)	Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会 2001年秋季講演論文集, Vol. , No. 1-2-8, pp. 221-222
Citation(English)	, Vol. , No. 1-2-8, pp. 221-222
発行日 / Pub. date	2001, 10

数量化 I 類を用いたピッチパターン生成における制御要因の検討 *

◎山田 真裕　岩野 公司　古井 貞熙 (東工大)

1　はじめに

当研究室では，現在テキスト音声合成システムの構築を進めている．ピッチパターンの制御には，2 種類の数量化 I 類による手法を用いた．そこで，ピッチパターンについて，ヒューリスティックなルールに基づく手法との，被験者による比較評価実験を行った．また，数量化における制御要因の検討を行った．

2　テキスト音声合成システム

2.1　テキスト解析部

テキスト解析には NTT-IT 社のテキスト音声合成ソフトウェア「Hipervoice」[1] 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．

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

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，これに適当なピッチを与えて，MLSA フィルタ [2] を用いて音声合成する手法が提案されている [3]．当研究室では，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 [4]．本研究では，この手法によりパラメータ生成を行った．

3　ピッチの推定

ピッチの推定には，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 類を用いる [5] [6] [7] [8]．

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

ここで，i 番目のデータの推定値を $\hat{y_i}$，全データの平均値を $\bar{y}$，データ数を N とする．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，数量化 I 類を用いた 2 種類のピッチ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ピッチ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ピッチの値を推定する方法 (以下，モーラベース) である．

なお，ピッチは，基本周波数 (F_0) を次式のように対数変換 [9] した値 (semitone) を用いている．

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

3.1　モデルベース

この手法では，図 1 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ピッチパターンから最小 2 乗近似により以下の値を推定し，これを学習段階での目標値とする．

・第 1 モーラの母音中心部におけるピッチの値 (Fs)
・最終モーラの母音中心部におけるピッチの値 (Fe)
・ストレス量 (S)

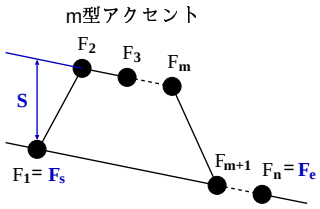

図 1. 台形点ピッチ近似モデル

用いた制御要因は以下のものである．() 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ピッチ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．

1	当該アクセント句のモーラ数 (8)
2/3	当該アクセント句の属す MG 中の先行/後続グループのモーラ数 (9)
4	先行アクセント句のアクセント型 (7)
5	当該アクセント句のアクセント型 (7)
6	後続アクセント句のアクセント型 (7)
7	アクセント核をもつ先行アクセント句数 (4)
8/9	前/後の結合の強さ (4)
10/11	前/後の句境界の長さ (ポーズの長さ)(9)
12/13	2 つ前/後の結合の強さ (5)
14/15	3 つ前/後の結合の強さ (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, :(:は長音)) の中心部におけるピッチの値を推定する．

制御要因はモデルベースで用いる制御要因のうち，当該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，以下のものを用いる．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 [6]．

---

* A study on pitch contour generation factors using categorical multiple regression.
By Masahiro Yamada , Koji Iwano and Sadaoki Furui (Tokyo Institute of Technology)

16	トーンパタン [6](モーラ位置により異なる．5〜10)
17	当該音素の種類 (8)
18/19	当該音素の先行/後続音素の種類 (13)
20	モーラ位置 (モーラ位置が 5 以上の場合のみ．6)
21/22	当該音素の 2 つ前/後の音素の種類 (6)

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 (話者 MHT) を用いた．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．

評価文には，学習に用いた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)，ニューステキスト 5 文 (news) を用いた．

以下の実験では，被験者にテキストを提示した上で，各方式で合成した音声を被験者ごとに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，close，open については合成部で用いる HMM により強制切り出しした結果をそのまま利用し，news については，文脈を考慮した音素の平均時間長とした．

4.1 3 手法の比較

テキスト解析結果からヒューリスティックなルールに基づいて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) と，数量化 I 類を用いた 2 つの手法の計 3 つのピッチパターン生成法を用いて合成音声を作成し，評価実験を行った．

テキストには close，open，news を用いた．表 1 に実験結果を示す．

表 1. 3 手法の比較の結果 (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とから，モーラベースの手法が最も有効であると判断できる．

4.2 制御要因の組み合わせによる比較

最もスコアの高かったモーラベースの手法において，制御要因の組み合わせと主観評価の関係を調べた．

21 個の制御要因から，最も重要でない制御要因 (除いたときに誤差が最小になる制御要因) 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 2 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．

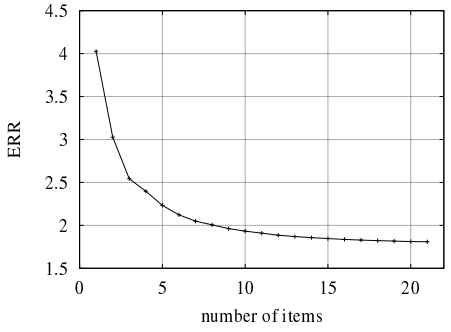

図 2. 制御要因数と誤差のグラフ

表 2. 推定誤差による比較の結果 (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

制御要因数が 1,2,3,5,8,21 の場合の学習結果を用いて合成音声を作成し，この 6 種類の合成音声の比較実験を行った．テキストには close，open を用いた．表 2 に実験結果を示す．

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．このことから，重要度の高さが 8 番目までの制御要因が，ピッチパターンの知覚上，特に重要であると考えられる．具体的には，トーンパタン，モーラ数，音素，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．

5 まとめ

数量化 I 類を用いたピッチパターン生成の有効性を確認した．また，ピッチパターンの知覚上，どの制御要因が重要であるかについて検討した．今後は，4.2 節で有意差が認められた部分について検討を行っていく．

参考文献

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

[2] 今井 他, 信学論, Vol.J66-A, No.2, pp.122–129 (1983-2).

[3] 益子 他, 信学論, Vol.J79-D-II, No.12, pp.2184–2190 (1996-12).

[4] 立和 他, 音講論, 2-3-7 (1999-3).

[5] 海木 他, 信学論, Vol.J83-D-II, No.9, pp.1853–1860 (2000-9).

[6] 阿部 他, 音響誌, Vol.49, No.10, pp.682–690 (1993-10).

[7] 箱田 他, 信学技報, SP89-5 (1989-5).

[8] 酒寄 他, 音講論, 3-4-17 (1986-10).

[9] 小坂 他, 音講論, 2-4-11 (1990-3).

[10] 箱田 他, 音講論, 1-2-14 (1988-10).