文節を基本単位とした基本周波数パターン生成過程モデルのパラメータ自動抽出*

☆橋本浩弥, 広瀬啓吉, 峯松信明 (東大)

1 はじめに

声帯振動の基本周波数の時系列パターンを表現するモデルとして、基本周波数パターン生成過程モデルがある [1]。このモデルは少数のパラメータで基本周波数パターンをよく記述することができ、生理的・物理的根拠に基づいており、言語情報とよく対応が取れるという特徴がある。そのため、このモデルパラメータを用いることにより、焦点制御やスタイル制御などを少数の学習コーパスから効率的に実現することができる [2, 3]。しかし、観測される基本周波数パターンからモデルパラメータを抽出することは逆問題であり、解析的に解くことができないという問題がある。そこで、HMM 音声合成で用いられているコンテキストラベルを利用して、モデルパラメータを高精度に自動抽出する手法を以前に提案した [4]。しかし、従来の HMM 音声合成で用いられているコンテキストラベルは、アクセント句を基本単位としているが、アクセント句は定義に曖昧性がある上、テキストだけではなく、話者の発話スタイルや発話速度などに依存するため、自動推定することが困難である。

本稿では、生成過程モデルの特にアクセント指令について、アクセント句ではなく、文節を単位として自動抽出することを試みる。そして、どのような場合に文節とアクセント指令が対応しないのか考察する。

2 基本周波数パターン生成過程モデル

基本周波数パターン生成過程モデルとは、喉頭の生理的・物理的特性に基づいて、声帯振動制御機構を定量的にモデル化したものである。

このモデルは、対数軸上で表現した基本周波数パターン（F0 パターン）が基底周波数、フレーズ成分、アクセント成分の和で表されとしている。概念図は Fig. 1 であり、数式は次式であらわされる。

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} A_{p,i} G_p(t - T_{p_s,i}) + \sum_{j=1}^{J} A_{a,j}\{G_a(t - T_{a_s,j}) - G_a(t - T_{a_e,j})\} \quad (1)$$

ここで、F_b は話者の発話スタイルに固有な値である基底周波数、$A_{p,i}, T_{p_s,i}$ は i 番目のフレーズ指令の大きさ、立ち上がり位置 $A_{a,j}, T_{a_s,j}, T_{a_e,j}$ は j 番目のアクセント指令の大きさ、立ち上がり位置、立ち下がり位置をあらわす。

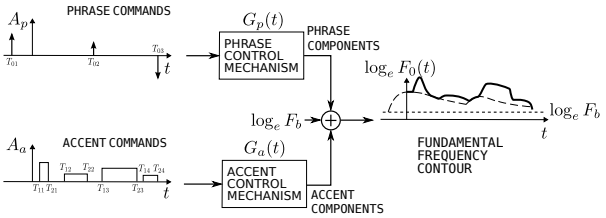

Fig. 1 基本周波数パターン生成過程モデル

また、$G_p(t)$ はフレーズ成分であり、次式

$$G_p(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & t \geq 0, \\ 0, & t < 0, \end{cases} \quad (2)$$

であらわされ、$G_a(t)$ はアクセント成分であり、次式

$$G_a(t) = \begin{cases} \min[1 - (1 + \beta t)\exp(-\beta t), \gamma], & t \geq 0, \\ 0, & t < 0, \end{cases} \quad (3)$$

であらわされる。ここで α, β, γ は日本人の話者の平均的な値として、経験的に、$\alpha = 3.0$, $\beta = 20.0$, $\gamma = 0.9$ とし得ることが知られている [5]。

このモデルでは、少数のパラメータで F0 パターンをよく表現しているが、実測される F0 パターンは、式 (1) に示す連続曲線ではなく、F0 が観測されない（無声）区間や、不規則な声帯振動や F0 抽出アルゴリズムの不完全性に起因する誤った F0 が存在するため、F0 パターンのみからパラメータを推定することが困難であるという問題がある。

3 文節を基本単位としたモデルパラメータの自動抽出

手法の詳細は [4] に譲るとし、ここでは、本文での文節を定義する。文節は自立語とそれに後続する 0 個以上の付属語として定義することが多いが、実際に文を文節に区分するに際しては、いくつか考慮しておく点がある。まず、複合名詞は 1 つの文節に含まれるとした。また、「運動する」のように形態素解析辞書 Unidic[6] において、「名詞」＋「非自立可能な動詞」となる場合は 1 つの文節とし、「…、という …」のようなケースは、読点には必ず文節句境界があると

*Automatic command extraction for the generation process model of F0 contours using “bunsetsu” as the basic unit of analysis by HASHIMOTO Hiroya, HIROSE Keikichi, and MINEMATSU Nobuaki (The University of Tokyo)

	accent phrase		bunsetsu	
	recall	precision	recall	precision
phrase command	96.7	80.4	96.4	79.6
accent command (onset)	91.6	97.9	90.6	96.2
accent command (reset)	89.2	95.4	88.3	93.8

Table 1　モデルパラメータの抽出性能

し、直後の「と」は自立語を持たない単独の文節句であるとして取り扱った。そして、文節句内の第一アクセントを基準にアクセント指令の推定を行う。

4　モデルパラメータ抽出性能に関する実験

生成過程モデルのモデルパラメータ自動抽出手法 [4] において、アクセント句を基本単位とする場合と、文節を基本単位とする場合を比較した。

4.1　実験条件

それぞれの手法で抽出したモデルパラメータと正解のモデルパラメータから、次式で定義される再現率（recall) と適合率 (precision) を求め、比較した。

$$recall = \frac{\text{正解数}}{\text{手動抽出数}} \times 100 \quad [\%] \quad (4)$$

$$precision = \frac{\text{正解数}}{\text{自動抽出数}} \times 100 \quad [\%] \quad (5)$$

ここで正解とは、手動で抽出した指令の該当するモーラから前後 1 モーラ以内に指令があるものを指す。そして、フレーズ指令の生起位置、アクセント指令の立ち上がり位置、立ち下がり位置についてそれぞれ再現率と適合率を求めた。ただし、指令の大きさが 0.1 以下のものについては、F0 パターンの再現性において重要ではないため、ここでは除外した。

音声データは ATR 日本語音声データベース [7] の B セットの中から、話者 MHT を選択した。両手法において、F0 は STRAIGHT[8] を用い、フレーム周期 5 [msec]、最小値 60 [Hz]、最大値 200 [Hz] で抽出し、基底周波数 F_b は 60 [Hz] とした。提案手法において、音素ラベルの時間情報は Julius[9] を用いて得て、アクセント情報は、HTS-2.1[10] のデモスクリプトに付属しているラベルを利用した。ただし、sp（ショートポーズ）ラベルは音声に合わせて修正した。

4.2　結果

結果を Table 1 に示す。大きな差はないものの、全体的に文節を基本単位とした方が自動抽出性能が低下しているが、その原因を述べる。まず、アクセント指令において、再現率が低下した主な原因は、「ばかり」「べき」などの助動詞や、「反」などの接頭辞においてアクセント核（すなわちアクセント成分）が自立語とは別に単独で存在するケースがあった。また、今回は複合名詞を 1 つの文節としてしまったため、複数のアクセントに分かれる名詞連続に対応ができなかった。これに対しては、名詞ごとに文節を分割することも考えられるが、接尾辞に近い役割を持つ名詞を考慮する必要がある。一方、適合率の低下については 1 つの指令を 2 つの指令で表現しているだけであり、特に問題はないと考えられる。そして、「とき」、「こと」などのほぼアクセントがないような（アクセント、フレーズ指令を必要としない）文節において、不要な指令を立ててしまっているケースがあった。

5　まとめ

本稿では、文節を基本単位として、基本周波数パターン生成過程モデルのモデルパラメータ自動抽出を試みた。文節を基本単位とした場合、アクセント句を基本単位とした場合に比べて自動抽出性能が若干低下したものの、大きな差はなかった。そして、文節とアクセント句が対応しないケースは、名詞連続の場合以外では、ほぼ特定の単語に限られることがわかった。文節を単位とすることで、発話スタイル、発話速度の変更への対処が容易になると考えられる。今回の実験結果を踏まえて、テキスト音声合成のための文節を利用した基本周波数パターンの加算モデルの構築を検討していく予定である。

参考文献

[1] H. Fujisaki, *et al,* Annual Report of Engineering Research Institute, Univeristy of Tokyo, vol. 28, pp. 53–60 1969.

[2] K. Hirose, *et al,* Proc. Speech Prosody, 6th International Conference, SS1-3, 2012.

[3] K. Hirose, *et al,* Proc. INTERSPEECH, 2011.

[4] H. Hashimoto, *et al,* Proc. INTERSPEECH, 2012.

[5] H. Fujisaki, *et al,* J. Acoust. Soc. Japan (E), vol. 5, no. 4, pp. 233–242, 1984.

[6] Unidic, http://www.tokuteicorpus.jp/dist/

[7] A. Kurematsu, *et al,* Speech Communication, vol. 9, pp. 357–363, 1990.

[8] H. Kawahara, *et al,* Speech Communication, vol. 27, no. 3–4, pp. 187–207, 1999.

[9] Julius, http://julius.sourceforge.jp/

[10] HTS, http://hts.sp.nitech.ac.jp/