

基本周波数パターン生成過程モデルにおける フレーズ指令パラメータ抽出の高精度化*

☆千々岩圭吾, 鈴木雅之, 齋藤大輔, 峯松信明, 広瀬啓吉 (東京大学)

1 はじめに

音声の韻律的特徴の一つである、基本周波数 (F_0) パターンを表現するために様々な F_0 モデルが提案されている [1]。その中でも F_0 パターンを少数のパラメータによって定量的に記述するモデルとして、藤崎らによって提案された F_0 パターン生成過程モデルがある [2]。この生成過程モデルは、音声合成において広く利用されている [3]-[5]。また、外国語学習者の韻律分析 [6] にも利用されるなど、生成過程モデルは幅広い応用分野を持っている。

生成過程モデルを利用するためには、音声からモデルのパラメータを抽出する必要がある。手動によるパラメータ抽出も可能ではあるが、抽出には豊かな経験と大きな労力が必要になる。そのため、パラメータ抽出を自動的に行う手法が成澤らによって提案されている [7]。しかし、この精度は十分なものではない。

そのため、生成過程モデルのパラメータ自動抽出を、言語情報 (各種境界の時間情報や係り受けの情報) を用いることで高精度化する手法が古山らによって検討されている [8]。この手法は、母語話者の発声はおおよそ言語的制約に沿うと仮定することでモデルパラメータの抽出精度を改善させる手法であり、大幅な精度改善が実現されている。音声合成のためのデータ作成の自動化には、この手法は有効である。ところが、例えば母語話者以外の発音の誤りの評価を目的とした韻律分析という応用を考えた場合、母語話者が従う言語的制約をパラメータ抽出に導入するということは、望ましいものではない。非母語話者の発音の韻律的な間違いを定量的に評価及び指摘できるようにするためには、言語情報を制約として使わない自動抽出が求められる。

そこで本研究では、言語情報を用いず、より高い精度で韻律分析を行える手法を目指す。特に、成澤らの手法において、無音区間および無声区間付近におけるパラメータの脱落誤りや、挿入誤りが非常に多いことに着目し、生成過程モデルのフレーズ指令の自動抽出手法を高精度化する手法を提案する。提案手法を用いることで、従来手法に比べ再現率を下げることなく、挿入誤りを 21% 削減する精度向上が実現できた。

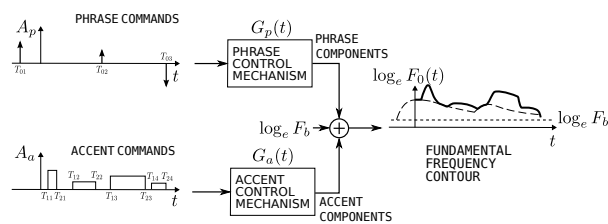


Fig. 1 基本周波数パターン生成過程モデル

2 基本周波数パターン生成過程モデル

生成過程モデルは、 F_0 パターンを句の始まりから終わりに至るまで緩やかに下降するフレーズ成分と、語に対応した局所的な凹凸で表されるアクセント成分、発声中は常に一定の値を保つベースライン成分 (基底周波数) の 3 つの成分の和として考えて表現したものである。生成過程モデルの概念図を Fig. 1 に示す。

生成過程モデルでは、フレーズ成分をインパルス状のフレーズ指令に対するフレーズ制御機構の応答、アクセント成分をステップ状のアクセント指令に対するアクセント制御機構の応答として表現し、両制御機構をそれぞれ臨界制動 2 次線形系で近似する。 F_0 パターンは式 (1) で表される。

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^I A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^J A_{aj} \{G_a(t - T_{1j}) - G_a(t - T_{2j})\} \quad (1)$$

ただし、ここで F_b は話者に依存する F_0 パターンの基底周波数である。また、 $G_p(t)$ 、および $G_a(t)$ はそれぞれフレーズ制御機構のインパルス応答、アクセント制御機構のステップ応答で、以下の式 (2)、式 (3) で表される。

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

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

ここで α はフレーズ制御機構の固有角周波数、 β はアクセント制御機構の固有角周波数であり、 γ は $\gamma \leq 1$ なる定数で、アクセント成分が有限時間内で一定値に達することを保証するものである。 α 、 β および γ

* An improved method for automatic extraction of phrase commands of the F_0 contour generation process model by CHIJIWA, Keigo and SUZUKI, Masayuki, SAITÔ, Daisuke, MINEMATSU, Nobuaki and HIROSE, Keikichi, (The University of Tokyo)

の話者ごと、発話ごとの変動は比較的小さいためそれぞれ、 $\alpha = 3.0\text{rad/s}$ 、 $\beta = 20.0\text{rad/s}$ 、 $\gamma = 0.9$ を用いることができる [9]。また、 A_{pi} と T_{0i} はそれぞれ i 番目のフレーズ指令 (インパルス) の大きさと生起位置、 A_{aj} と T_{1j} と T_{2j} はそれぞれ j 番目のアクセント指令 (ステップ) の振幅と立ち上がり位置、立ち下がり位置である。この生成過程モデルを用いれば、 F_0 パターンを少数のパラメータで定量的に表すことができる。

3 先行研究

生成過程モデルパラメータを発話音声から自動的に抽出しようとする研究がこれまでにいくつかなされてきている [7][8][10][11]。ここでは、本研究のベースとして参考にした成澤らの手法 [7] に付いて述べる。先行研究で提案されているモデルパラメータ抽出法の流れを、Fig. 2 に示す。以下、Fig. 2 にそって順に説明していく。

3.1 基本周波数パターンに対するの前処理

実測の F_0 パターンは、式 (1) で表されるような連続的な曲線ではなく、またモデルでは考慮されていない倍ピッチや半ピッチなどのピッチ抽出エラーや、無声区間前後での F_0 の急峻な変化であるマイクロプロソディーなどの要素も含まれている。そのため、まずそれらの検出及び除去を行う。これらの検出には倍ピッチ半ピッチエラーに関しては連続性などの制約条件、またマイクロプロソディーに関しては、無声区間前後での F_0 の微分係数の閾値処理などの手法を用いる。エラーを除去することにより、Fig. 2 における (3) のようなパターンが得られる。

次に、全体を 3 次曲線でスプライン補間する。これにより、Fig. 2 における (4) のような平滑化 F_0 パターンが得られる。

3.2 アクセント指令の推定

次に、平滑化 F_0 パターンからアクセント指令の位置及び大きさを推定する。アクセント成分は F_0 パターンの局所的な凹凸に相当するので、平滑化 F_0 パターンの傾きが大きい時点を見つけることで推定する。また、その大きさについては推定された時刻の微分値より求める。

前処理で、 F_0 パターンは 3 次式である平滑化 F_0 パターンで近似されているので、それ微分して 1 次微係数を算出し、算出した 1 次微係数の極大・極小値を発話開始時点から順に求める。ある閾値より大きい極値について、時間的に隣り合う同符号の極値をまとめてとりとして、それらのまとまりの中から最大値・最小値を選び出す。このように選定した 1 次微係数の正負の極値の位置をそれぞれ平滑化 F_0 パターンの立ち上がり候補点、立ち下がり候補点と定義する。アクセント指令は、この候補点の位置から推定する。隣接する立ち上がり、下がり候補点を 1 つの組として、立ち上がり候補点の位置より $1/\beta$ だけ先行する時点

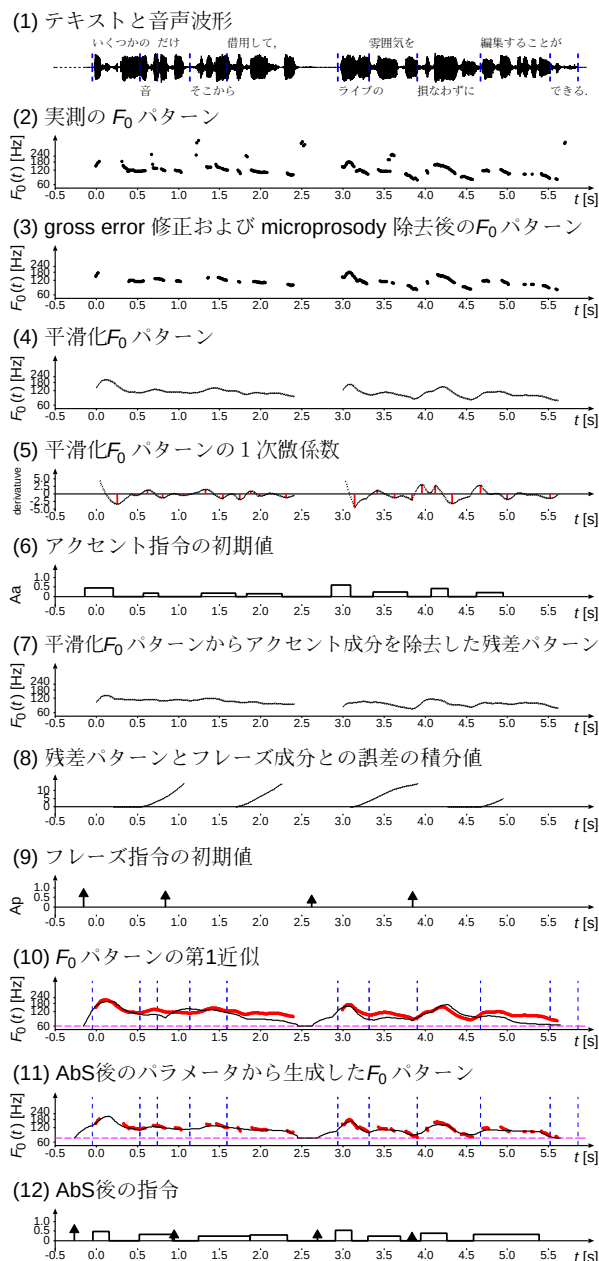


Fig. 2 自動抽出の流れ

をアクセント指令の立ち上がり位置とする。立ち下がりの位置も同様にして推定する。但し、発話の先頭の韻律語のアクセントが頭高型の場合は立ち上がり候補点が検出されずまず立ち下がり候補点が検出されるため、発話速度を 7[モーラ/sec] として、その立ち下がり候補点よりも 1 モーラ長 (=0.143[sec]) だけ先行する時点を実際のアクセント指令の立ち上がりとする。また発話の終わりの韻律語のアクセントが平板型の場合や、区間休止が存在する場合は立ち下がり候補点が検出されずに立ち上がり候補点が検出されるため、発話の終わりの位置を実際のアクセント指令の立ち下がりとする。

3.3 フレーズ指令の推定

フレーズ指令は、平滑化 F_0 パターンからアクセント指令による F_0 パターンを引き算した F_0 残差パターン (Fig. 2 の (7)) から推定する。

まず、フレーズ指令の一つ目の位置は、発話区間が始まる $1/\alpha$ 秒前とする。一つ目のフレーズ指令の大きさ A_{p1} は残差成分の最大値から式 (1) を展開した式 (4) で計算される。

$$A_{p1} = \frac{e}{\alpha} \{ \ln R_0(T_{R_{0max}}) - \ln F_b \} \quad (4)$$

を用いて計算する。次に、 F_0 残差パターンから一つ目のフレーズ成分を引き算する。そうして得られた新たな残差パターンから、二つ目のフレーズ成分を推定する。

アクセント成分と一つ目のフレーズ指令の成分を引いた残差パターンの成分は、残る二つ目以降のフレーズが担う成分である。その残差成分を積分していき、その値が閾値 θ_1 を越えた場合、その積分値が増え始める時点に遡って、そこに新たなフレーズ指令を立てる。ただし、新たなフレーズ指令を立てる位置がアクセント指令の内部に存在する場合は、アクセント指令の外で平滑化 F_0 パターンの 1 次微係数が負から正へと変化する点を新たなフレーズ指令の生起位置とする。このような点が存在しない場合は、便宜的に前後のアクセント指令に関する直前の立ち下がり直後の立ち上がりの中間を新たなフレーズ指令の立ち上がり点とする。大きさについては、一つ目と同様に式 (4) から計算される。この処理を、誤差パターンの積分値 (Fig. 2 の (8)) が閾値をこえる限り続けることで、二つ目以降のフレーズ指令も得る。

3.4 パラメータの最適化 (A-b-S 処理)

以上の操作によって推定されたパラメータを初期値として、パラメータを微小変化させ、平滑化された F_0 パターンの対数値とパラメータから生成される F_0 パターンの二乗誤差が最小になるように逐次近似していく。ただし、最適化を行う分析区間は、指令の生起位置から、パラメータの微小変化が生成される F_0 パターンに与える影響が無視できる区間までとする。またこの際、フレーズ指令の大きさが 0.1 以下となった場合は、大きさを 0 として削除する。

4 無音区間に着目した自動抽出の高精度化

上に述べた手法では、経験則に基づき各種計算を行うことでパラメータ初期値を推定し、それを A-b-S 法で最適化することでパラメータを抽出している。しかしながら、従来手法では本来抽出されるべきフレーズ指令が脱落してしまったり、また不要なフレーズ指令が多数挿入されてしまっている。これらの誤りは、特に前処理によって F_0 が内挿される無音区間および無声区間付近において多発している。

そこで、本研究では、パラメータの初期値推定時において、無音区間および無音区間付近に対しては

フレーズ指令をなるべく多く設定し、A-b-S 法でパラメータの最適化するときには、削除すべきフレーズ指令を検出することにより、脱落および挿入誤りを低減することを目指す。なお、今回の手法では、有聲区間とならなかった無音区間及び無声区間を同等に、無音区間として扱った。以下、無音区間及び無声区間を、無音区間として表記する。

4.1 初期値抽出の改善

まず、フレーズ指令の初期値抽出における改善手法を述べる。Fig. 2 の (7) から (9) にかけての処理について、無音区間付近の欠落誤りを減らすために、無音区間付近においてフレーズ指令が立ちやすいような条件を追加することで、脱落誤りの低減を目指す。

今回は、その単純な実装として、Fig. 2 の (8) において残差の積分を計算する際、無音区間の部分ではその区間の残差に、それ以外の部分より大きくなるように重みをつけることを検討した。無音区間の重み付けは、予備実験により、他の区間の 5 倍とした。このような処理を行うことで、フレーズ指令の初期値推定の段階で、無音区間が多い部分においてフレーズ指令候補が多く推定されることになり、無音区間付近の脱落誤りが低減することが予想される。

4.2 最適化手法の改善

前項で述べたように、初期値抽出の段階では、フレーズ指令を多く推定するように変更した。そのため、このままではフレーズ指令が立ちすぎてしまい、挿入誤りがさらに増加してしまう。そこで、Fig. 2 の (11) における A-b-S 法を用いたパラメータの最適化の段階で、挿入誤りと考えられるフレーズ指令を従来より多く削除する手法を提案する。

削除すべきフレーズ指令の検出方法は、フレーズ指令はショートポーズ付近に生じやすいということから [8]、比較的無音区間が少ない部分で多く削除を行う。初期値抽出が終わった後、まず Fig. 2 の (12) のように一度 A-b-S 法によってパラメータの最適化を行い、指令の位置を整える。次に、各フレーズ指令の前後を探索し削除すべきフレーズ指令の候補を検出し、その候補のうち指令の大きさが小さいものを削除する。ここで、各フレーズ指令の探索区間は予備実験により指令生起時点から前後 1 モーラ (計 2 モーラ) とし、無音区間の比率が探索区間の半分を下回っていた場合を削除すべきフレーズ指令の候補とする。その削除候補のうち大きさが閾値 $\theta_2 (> 0.10)$ を下回ったものを削除する。また、候補以外のフレーズ指令については、大きさ 0.10 以下の場合のみ、無意味な指令として削除される。最後に、削除すべきフレーズ指令を除いた後に、もう一度 A-b-S を行いパラメータの微調整を行う。

5 実験

無音区間に着目した自動抽出法の有効性を、実験を通して検証する。実験には、日本語音声資料とし

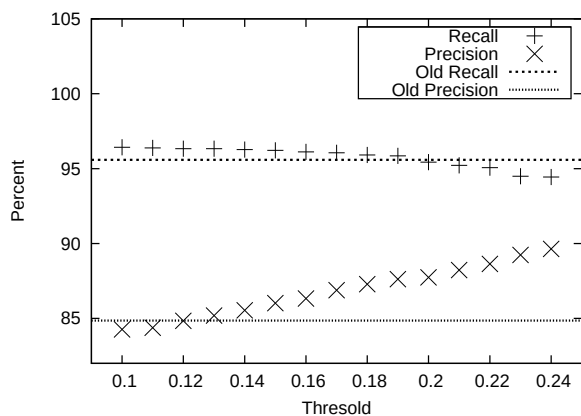


Fig. 3 提案手法の再現率、適合率 (破線は従来手法)

て ATR503 文 (男性話者 mht) を用いた。今回は、16bit/16kHz で収録された音声を 16bit/10kHz にダウンサンプリングして自動抽出を行った。また、指令の脱落、挿入誤りは、韻律研究に深く関わっているエキスパート (第五著者ら) が手動抽出した指令を正解として算出した (これらがすべて正しく、かつ他の指令はないものとする)。このとき、エキスパートによる手動分析結果のフレーズ指令が存在する韻律句内に、フレーズ指令を最終的に推定した場合を正解とした [7]。評価には、再現率 (recall) と適合率 (precision) を用いる。再現率と適合率は下式で計算される。

$$\text{再現率} = \frac{\text{正解数}}{\text{手動抽出数}} \times 100 \quad (5)$$

$$\text{適合率} = \frac{\text{正解数}}{\text{自動抽出総数}} \times 100 \quad (6)$$

また実験では、削除すべきフレーズ指令を検出する指令の大きさの閾値 θ_2 を、0.10 から 0.25 まで変化させて、抽出結果の再現率と適合率を見る。この閾値を上げるほど削除されるフレーズ指令は増え、適合率は上がるが、逆に正解の指令まで削除し、再現率が低下する可能性が高まる。 $\theta_2 = 0.10$ のときに、従来手法と同等となる。

結果を Fig. 3 に示す。まず、閾値 θ_2 が 0.10 の場合 (従来手法と同等) を見る。このときは、削除すべき指令を検出しないのと同様なので、フレーズ指令の初期値抽出法の改善 (第 4.1 節) の結果だけを見て取ることができる。無声区間付近でフレーズ指令を抽出しやすくなった結果、再現率は 0.8% 上がっているが、挿入誤りも増え適合率が 0.5% 下がってしまった。次に閾値 θ_2 を 0.10 以上に上げフレーズ指令を削除していくと、0.10 から 0.19 までは、0.10 のときと同様再現率はわずかに上昇し、適合率は大きく上昇した。つまり、正解数を下げることなく、挿入誤りを減らすことができたことになる。閾値を 0.19 としたときの再現率は 95.9%、適合率は 87.6% で、これは従来手法よりそれぞれ 0.3%、2.8% 上昇している。しかし、閾値を 1.9 以上に高くしてしまうと、正解指令まで除去してしまい、再現率の低下が起こる。

Table 1 探索の有無による再現率、適合率の比較

探索処理の有無	再現率	適合率
有	95.9%	87.6%
無	88.0%	95.9%
従来手法	95.6%	84.9%

また、第 4.2 節のように、まず削除する指令候補を検出してから閾値処理をするのではなく、従来手法のように単純にフレーズ指令の大きさの閾値処理だけでフレーズ指令を削除した場合、閾値を 0.10 から 0.19 に上げると適合率は 93.0% まで高まるが、再現率が従来手法よりはるかに低い 88.0% まで低下してしまう。再現率が大きく低下するのは、パラメータの自動抽出としては望ましくない。つまり、従来手法のような単純な閾値処理だけでは、高い再現率と適合率の両方を実現できないことが分かる。このことより、フレーズ指令の前後を探索し、無音区間の比率から削除すべき指令の候補を見つける操作が有効であると言える。

6 まとめ

本論文では、無音区間に着目してフレーズ指令の初期値抽出において脱落誤りを減らし、指令の最適化の段階では疑わしいフレーズ指令を削除する手法を提案した。実験により提案手法により、再現率を下げることなく、適合率をあげることに成功した。今後の予定としては、パラメータの最適化を非周期性などの特徴量を用いて、さらに高精度化する予定である。

参考文献

- [1] 山下洋一, 信学技報, SP2009-66, 2009.
- [2] 藤崎博也, 須藤寛, 日本音響学会誌, vol.27, No.9, 1971.
- [3] 石川泰, 信学技報, SP2000-72, pp. 27-34, 2000.
- [4] 越智景子, 他, 信学技報, SP2009-67, 2009.
- [5] K.Hirose, K.Sato, et al, *Speech communication*, vol.46, no.3-4, pp. 385-404, 2005.
- [6] 平野宏子, 他, 音講論 (秋), 331-332, 2007.
- [7] 成澤修一, 他, 情報処理学会論文誌, vol.43 no.7, 2002.
- [8] 古山悠介, 他, 信学技報, SP2003-70, 2003.
- [9] H.Fujisaki, K.Hirose, *J. Acoust. Soc. Jpn(E)*, vol.5, No.4, 233-242, 1984.
- [10] 木村太郎, 他, 信学技報, SP2006-82, pp. 37-42, 2006.
- [11] 布 社輝, 他, 信学技報, SP2004-172, pp.47-52, 2004.