社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2003－ＳＬＰ－49　(36)
2003／12／18

F_0モデル・パラメーターの自動決定方法についての考察

布　社輝†　板橋　秀一††　山本　幹雄††

†筑波大学 システム情報工学研究科　〒305-8573 茨城県つくば市天王台 1-1-1
††筑波大学 電子・情報工学系　〒305-8573 茨城県つくば市天王台 1-1-1

E-mail: †bushehui@milab.is.tsukuba.ac.jp, ††{itahashi,myama}@is.tsukuba.ac.jp

あらまし　本研究は、実測の F_0 パターンから、提案された F_0 モデルのパラメーターを自動的に抽出する方法について議論する。提案されたアルゴリズムは動的計画法および最小二乗法に基づいている。計算時間を減少させるために、このアルゴリズムは二ステップに分割する。最初に、フレーズ成分の境界を決定し、次に、フレーズおよびアクセント・コマンドの最適なパラメータを決定する。さらに、フレーズ・コマンドの最適数を自動的に決定するために、最小二乗誤差の減少量 $c(n)$ を提案する。最後に、日本語の音声例を対象として、本提案手法を用いて分析を行い、本手法の性能を実験的に検討する。

キーワード　F_0 パターン、F_0 モデル、藤崎モデル、最小二乗法、動的計画法 (DP)、平均最小二乗誤差 (LMSE)。

Considerations on a method of automatic determination of F_0 Model Parameters

Shehui BU†, Suichi ITAHASHI††, and Mikio YAMAMOTO††

† Graduate School of Systems & Information Science, University of Tsukuba
1-1-1 Tennodai. Tsukuba. 305-8573 Japan
†† Institute of Information Sciences & Electronics, University of Tsukuba
1-1-1 Tennodai. Tsukuba. 305-8573 Japan

E-mail: †bushehui@milab.is.tsukuba.ac.jp, ††{itahashi,myama}@is.tsukuba.ac.jp

Abstract　In this paper, an automatic method to extract the discrete parameters of an F_0 model is proposed which is an extended version of what is called Fujisaki model. This method is based on the dynamical programming(DP) and the least mean square error(LMSE) methods. It is divided into two main steps in order to decrease the calculation time exhausted by the DP method. Furthermore, in order to detect the optimal number of phrase commands automatically, decrease of LMSE c(n) is used. From the results of the experiment on a set of 11 sentences spoken by four Japanese speakers, we obtained about 84% correct rate of phrase component detection by the proposed method.

Key words　F_0 pattern, F_0 model, Fujisaki model, Dynamic programming(DP), Least mean square error, Two-step algorithm.

1. はじめに

音声の基本周波数パターン (以下 F_0 パターン) は音声コミュニケーションに重要な役割を果す。また、このパラメーターは主に韻律の情報によって決定される。人間の耳は、他の音声信号パラメーターの変化よりも F_0 の変化に、より敏感である。近年、コンピューターとマルチメディア技術の開発によって、実際の音声 F_0 パターンを自動的に分析し近似することが必要とされるようになった。

F_0 の生成のプロセスを数学的にパターン化することを説明する適切なモデルを得ていれば、F_0 パターンと韻律学の情報の関係は、定量的に分析することができる。藤崎モデル、Downstep モデルなどのようないくつかのモデルがこの問題に対処するために提案された [7]。多くの研究により、藤崎モデルが日本語等の言語についてイントネーションの変化を非常によく記述することができることが示された。一方、F_0 モデルのパラメーターを自動的に抽出する方法が必要である。幾つかの方法がこの問題を解決するために提案された。しかしながら、これらの方法では、良い初期値が必要とされることから、複雑な予処理を要求される [9、10]。

本研究は、実測の F_0 パターンから、提案された F_0 モデルのパラメーターを自動的に抽出する方法について議論する。提案されたアルゴリズムは動的計画法および最小二乗法に基づいている。計算時間を減少させるために、このアルゴリズムは二ステップに分割する。最初に、フレーズ成分の境界を決定し、次に、フレーズおよびアクセント・コマンドの最適なパラメータを決定する。さらに、フレーズ・コマンドの最適数を自動的に決定するために、最小二乗誤差の減少量 $c(n)$ を提案する。最後に、日本語の音声例を対象として、本提案手法を用いて分析を行い、本手法の性能を実験的に検討する [1、4]。

2. F_0 モデル

このモデルは 1970 年代に藤崎らによって提案された。このモデルは、フレーズ成分とアクセント成分、二つの成分から構成される。文献 [2、8、9、10] によれば、藤崎モデルは以下のように表現される:

$$\ln\left(\hat{F}_0(t)\right) = \log_e F_{min} + \sum_{i=1}^{I} A_{pi} G_{pi}(t - T_{0i}) + \sum_{j=1}^{J} A_{aj}\left\{G_{aj}(t - T_{1j}) - G_{aj}(t - T_{2j})\right\} \quad (1)$$

ここで、

$$G_{pi}(t) = \begin{cases} \alpha_i^2 t exp(-\alpha_i t) & : t \geq 0 \\ 0 & : t < 0 \end{cases} \quad (2)$$

$$G_{aj}(t) = \begin{cases} min\left[1 - (1 + \beta_j t)\, exp(-\beta_j t), \gamma\right] & : t \geq 0 \\ 0 & : t < 0 \end{cases} \quad (3)$$

方程式 (2) は、フレーズ制御機構の関数、また、方程式 (3) は、アクセント制御機構の関数を示している。方程式 (1)、(2) および (3) 中の記号は次の量を示す: I: フレーズ・コマンド数; J: アクセント・コマンド数; A_{pi}: 第 i 番フレーズ・コマンドの振幅; A_{aj}: 第 j 番アクセント・コマンドの振幅; T_{0i}: 第 i 番フレーズ・コマンドの時刻; T_{1j}: 第 j 番アクセント・コマンドの開始時刻; T_{2j}: 第 j 番アクセント・コマンドの終了時刻;

発話された音声の F_0 パターンは、個々の音韻による影響も加わって、非常に複雑な働きを示している。一般に、F_0 パターンの基本となる成分は、吸気を伴うポーズの後、肺からの呼気圧の自然な減少に伴い、一定の減少率で降下するものとみなされるが、一定の高さから出発して一定の高さに向かって制御されつつ降下する特性を有する [5]。F_0 パターンをよく近似するために、傾きを考慮し、式 (1) の中の項 $\log(F_{min})$ を、直線 $b_i(t - T_{0i}) + c_i$ に置き替える [1,2]。

$$\ln\left(\hat{F}_0(t)\right) = b_i(t - T_{0i}) + c_i + \sum_{i=1}^{I} A_{pi} G_{pi}(t - T_{0i}) + \sum_{j=1}^{J} A_{aj}\left\{G_{aj}(t - T_{1j}) - G_{aj}(t - T_{2j})\right\} \quad (4)$$

α と β はそれぞれの制御機構の固有角周波数、γ はアクセント成分が有限時間内に一定値に達することを保証する相対飽和値である。α、β 及び γ の話者ごと、発話ごとの変動は比較的小さいため、初期値としては、それぞれ $\alpha = 3.0 rad/s$、$\beta = 20.0 rad/s$)、$\gamma = 0.9$) を用いることができる。

3. 分析方法

3.1 パラメータ値の決定

文献 [1、3、4] によって、方程式 (4) の中で示されるような関数 $\hat{F}_0(t)$ によって、$F_0(t)$ を近似することを仮定すると、我々は以下の関数によって平均二乗誤差を得ることができる:

$$\varphi(t) = \frac{1}{T}\sum_{t=1}^{T}\left\{\hat{F}_0(t) - F_0(t)\right\}^2 w(t) \quad (5)$$

誤差は有声と判断された区間に対してのみ考慮するので、有声ならば1、無声ならば0をとる関数$w(t)$をかける。

各モデルの持つ変数ごとについての偏微分を取り、これらを0とする連立方程式を解くことで、$\varphi(t)$を最小にする近似関数を得ることができる。

$$\frac{\partial\varphi(t)}{\partial A_{pi}}=0 \quad (6) \qquad \frac{\partial\varphi(t)}{\partial b_i}=0 \quad (8)$$

$$\frac{\partial\varphi(t)}{\partial A_{aj}}=0 \quad (7) \qquad \frac{\partial\varphi(t)}{\partial c_i}=0 \quad (9)$$

ここで、b_iは負の値を取る;また一方A_{pi}、A_{aj}およびc_iは負の値を取らないとする。

3.2 パラメータ入力時刻の決定

モデルのパラメータ入力時刻を決定するために、動的計画法(DP法)[1、3、4、5、11]を導入する。動的計画法の原理は多段間題を2段間題にすることと言える。その代わり、前の段階の最適の値を保持するための記憶テーブルが総計算量を減らすために必要となる。

最初に、長さTの発話を等間隔Nに分割し、フレーム番号$n=(1,2,...,N)$を仮定する。$0\equiv n_0<n_1<n_2<n_3<\cdots<n_K\equiv N$に対して、$\varphi_k(n:n_k:a^{(k)})$が$k$区間目の二乗誤差を表すとすると、

$$\varphi_k\left(n:n_k,a^{(k)}\right)=$$

$$(1/N_k)\sum_{n=n_{k-1}+1}^{n_k}\left\{\left(\hat{F}_{0k}\left(n-n_{k-1},a^{(k)}\right)-F_0(n)\right)^2 w(n)\right\} \quad (10)$$

ただし、$n_0=0$、$n_K=N$、$N_k=n_k-n_{k-1}$、また、$a^{(k)}$がパラメータA_p、A_a、b及びcを表す。ここで、最小化の手続を二つに分けることができる。1つはk番目の区間を最小化すること、また、もう一つは、1番目、2番目、、、、$k-1$番目を最小化することである。ここで、、$g_k(n_k)$はk番目の段階における0とn_kの間の区間の最適解、pはパラメータの数を表す。

$$g_k(n_k)$$

$$=\min\left(a_1^{(k)},a_2^{(k)},\cdots,a_p^{(k)};n_1,\cdots,n_k\right)\sum_{k=1}^{K}\varphi_k\left(n:n_k,a^{(k)}\right) \quad (11)$$

$$=\min\left(a^{(k)};n_1,\cdots,n_k\right)\left\{\varphi_k(n:n_k,a^{(k)})+g_{k-1}(n_{k-1})\right\}$$

3.3 二段アルゴリズム

DP法によると、4つのパラメーターA_p、A_a、bおよびcの計算は相当な計算時間を要する[1,3]。計算時間を減小するために、二段アルゴリズムを提案した[1,4,5]。このアルゴリズムは、図に示されるように、最初にDP法を用いてフレーズコマンドの時刻を自動的に決め、次に1フレーズ成分内で最適な4パラメータを決定する。

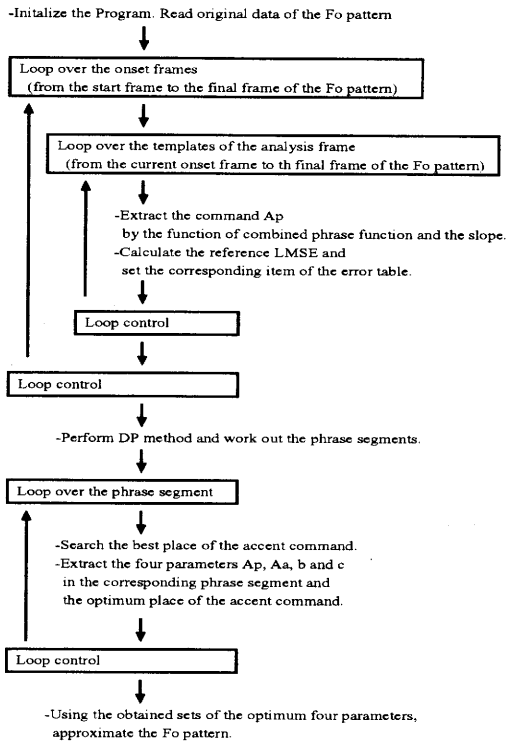

図1 二段アルゴリズムの概略図

第一段:フレーズ成分式(2)および直線$b_i(t-T_{0i})+c_i$の結合を最小二乗法とDP法を単純に使用して、フレーズ・コマンド時刻を安定かっ適切に抽出することができる。第二段:第一段の結果によって、発話文をフレーズ・セグメントに分割する。その次に、各フレーズ・セグメントの処理は二つの手続きから構成される。一つは対応するフレーズ・セグメント中のアクセント・コマンドの最良の境界を探索することで、もう一つは最小二乗法によりフレーズ・セグメント内の4つのパラメーターを算出することである。

3.4 フレーズ個数の自動判定

実際の分析では、提案された自動アルゴリズム中のフレーズ・セグメントの最適の数を決定すること

が必要である。ここでは、最小二乗誤差の減小量 $c(l)$

$$c(l) = |E(l) - E(l-1)| \tag{12}$$

と最小二乗誤差の減小率 $d(l)$

$$d(l) = |E(l) - E(l-1)/E(l)| \tag{13}$$

をとりあげる。

ここで、$E(l)$ は、フレーズ・セグメント l 個の場合、実測 F_0 と近似 $\hat{F}_0$ の平均最小二乗誤差；l はフレーズ・セグメントの個数である。そして、我々はそれぞれ、$c(l)$ と $d(l)$ の最高減少はフレーズ・セグメントの最適の数の最良の候補と考える。次の実験はこの仮定を検討するために実行された。

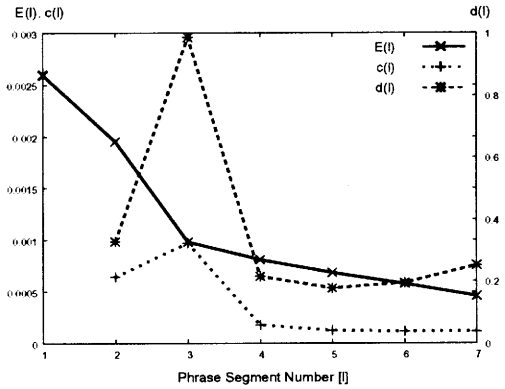

図 2 日本語音声「青い葵の絵はある」 */anoaoiaoinoewaaru/* の最小二乗誤差 (LMSE) $E(n)$、最小二乗誤差の減少量 $c(n)$ 及び最小二乗誤差の減少率 $d(n)$。

4. 分析実験

4.1 実験データ

実験データとして、表3で示されるように、日本語11文を一セットにとして選択した [2]。話者は20歳代の日本人男女各二名 (男性：M1、M2；女性：F1、F2) であり、防音室で2回 (一回目練習) 録音された。録音され音声データは 16kHz にサンプリングされ、16ビットで量子化された。AMDF方法によって F_0 パターンを抽出した。フレーム間隔は $10ms$ である。表3の項目 F(仮定されたフレーズ個数) は、実際の発音音声の F_0 パターンおよび言語的な意味を考慮して決定された個数である。実際に、発話者の個人性と言語的な曖昧性があるのために、このセット中のある文のフレーズ個数を唯一に決定することができない場合もある。

4.2 実験結果

4.2.1 分析一例

提案された方法を評価するために、実験を行った。分析内容は表3に示された男性 M1 によって発話された「あの青い葵の絵はある」 */anoaoiaoinoewaaru/* である。図2は最小二乗誤差 (LMSE)$E(n)$、最小二乗誤差の減少量 $c(n)$ および最小二乗誤差の減少率 $d(n)$ をそれぞれ表示している。図2の中で示された $c(n)$ および $d(n)$ によって、フレーズ個数が3である場合 $c(3)$ および $d(3)$ が最大値であることが分かる。これから、フレーズの最適な個数が3であると考えることができる。

フレーズ個数が3である場合の実験結果を図3に示す。この図から、モデルの F_0 パターンが実際の F_0 パターンを十分良く近似していることが分る。この場合、フレーズ個数3の実験結果がよく適合していることが分る。

4.2.2 比較結果

$c(n)$ と $d(n)$ の有効性を評価するために、収録音声データに対して実験を行った。表1と表2で、項目 $c(l)$ 及び $d(l)$ が判定個数を、項目 C が判定結果を表す。

No.	M1		M2		F1		F2	
	$c(l)$	C	$c(l)$	C	$c(l)$	C	$c(l)$	C
1	2	○	2	○	2	○	2	○
2	3	○	2	○	2	○	2	○
3	3	○	3	○	3	○	3	○
4	3	×	3	×	2	○	2	○
5	2	○	3	○	2	○	2	○
6	2	○	2	○	2	○	2	○
7	3	○	3	○	3	○	3	○
8	3	○	4	○	2	×	4	○
9	4	○	2	×	2	×	2	×
10	2	○	2	○	3	○	2	○
11	2	×	3	○	3	○	3	○
	81.8%		81.8%		81.8%		90.9%	

表 1 The number of Phrases decided by the decrease of LMSE $c(l)$. C:(○: correct; ×: incorrect.)

$c(l)$ と $d(l)$ は、それぞれ式 (12) と (13) を用いて、$c(l)$ と $d(l)$ が最大値となるときに、判断された最適なフレーズ個数である。C は、$c(l)$ 及び $d(l)$ と表3の項目 F を比べて、同じ場合正解 ○、異なる場合は、誤り × と判断された。話者 M1、M2、F1 および F2 に対応する $c(l)$ の正答率は 81.8% と 90.9% の間にある。対照的に、$d(n)$ の正答率は 45.5% から 72.7% までに及んでいる。$c(n)$ の平均正答率は 84.1% で、

また、$d(n)$のそれは59.1%である。これらの結果によって、我々は、$c(n)$の正答率が$d(n)$より高いと言うことができる。

No.	M1		M2		F1		F2	
	$d(l)$	C	$d(l)$	C	$d(l)$	C	$d(l)$	C
1	3	×	2	○	2	○	4	×
2	3	○	2	○	3	○	2	○
3	3	○	2	×	3	○	3	○
4	3	×	3	×	2	○	4	×
5	4	×	3	○	2	○	2	○
6	5	×	6	×	2	○	2	○
7	3	○	5	×	2	×	3	○
8	3	○	4	○	5	×	4	○
9	4	○	4	○	2	×	3	×
10	5	×	2	○	2	○	3	○
11	5	×	7	×	3	○	5	×
	45.5%		54.5%		72.7%		63.6%	

表 2 The number of Phrases decided by the decreasing rate of LMSE $d(l)$. C : (○: correct; ×: incorrect.)

5. まとめ

本研究では、提案されたF_0モデルに基づいてそのパラメーターを自動抽出する方法を提案した。また、フレーズ個数の自動判定方法の評価を行った。実験結果によって、本提案手法の有効性を示した。今後の課題としては、アルゴリズムの改善と高速化、より一般的な音声データによる検証、話者の増加が上げられる。

文献

[1] S. Bu, M. Yamamoto, S. Itahashi, *"A method of automatic extraction of F_0 Model parameters,"* Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition 2003, Tokyo, Japan, pp. 227-230, Apr. 2003.

[2] S. Bu, M. Yamamoto. S. Itahashi, *"Considerations on automatic parameters estimation of F_0 Model,"* Prep. Autumn Meeting of the Acoust. Soc. of Jpn, Paper 1-8-23, pp. 227-228, Sept. 2003. (in Japanese)

[3] H. Fujisaki, K.Hirose, *"Analysis of voice fundamental frequency contours for declarative Sentences of Japanese,"* Jour. Acoust. Soc. Jpn. (E)Vol.5, No.4, pp. 233-242, 1984.

[4] S. Furui, *"Digital speech processing, synthesis and recognition, second edition. revised and expanded,"* Marcel Dekker Inc., 2001.

[5] K. Hakoda, H. Sato, *"Prosodic Rules in Connected Speech Synthesis,"* IEICE Trans. , Vol. J63-D, No. 9 pp. 715-722. (in Japanese)

[6] S. Itahashi, *"Description of speech data pattern by several functions with applications to formant and fundamental frequency trajectories,"* STL-QPSR 2-3, pp. 1-22, Oct. 1978.

[7] H. Kubozono, *"The organization of Japanese prosody,"* Kurosio Publishers, Tokyo Japan, 1993.

[8] H. Mixdorff, *"A novel approach to the full automatic extraction of Fujisaki model parameters,"* ICASSP2000, Vol. 3, pp. 1281-1284, 2000.

[9] S. Narusawa, N. Minematsu, K. Hirose, H. Fujisaki, *"A method for automatic extraction of the fundamental frequency contours generation model,"* IPSJ Jour. , Vol. 43, No. 7, pp. 2155-2168, Jul. 2002. (in Japanese)

No.	文	ローマ字表示	F
1	青い葵	/aoiaoi/	2
2	青い葵の絵	/aoiaoinoc/	2,3
3	あの青い葵の絵	/anoaoiaoinoc/	3
4	葵の絵はある	/aoinocwaaru/	2
5	青い葵の絵はある	/aoiaoinocwaaru/	2,3
6	葵の絵は家にある	/aoinocwaicniaru/	2,3
7	あの青い葵の絵はある	/anoaoiaoinocwaaru/	3,4
8	青い葵の絵は家にある	/aoiaoinocwaicniaru/	3,4
9	あの青い葵の絵は家にある	/anoaoiaoinocwaicniaru/	4
10	葵の絵は山の上の家にある	/aoinocwayamanoucnoicniaru/	2,3
11	青い葵の絵は山の上の家にある	/aoiaoinocwayamanoucnoicniaru/	3

表 3　音声試料 (F : 各文のフレーズセグメント数)

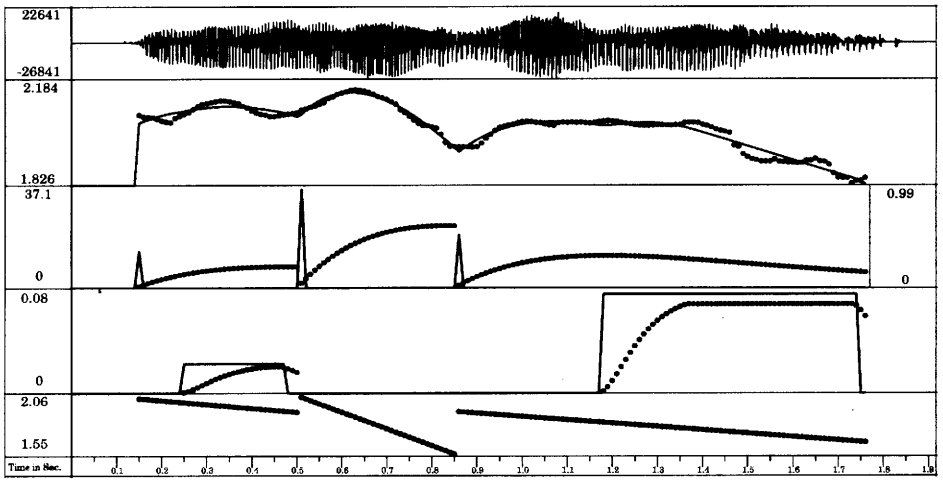

図 3　フレーズ個数が 3 とき、日本語音声「青い葵の絵はある」の分析結果。上から順に、音声波形；F_0 パターン (点線) と近似パターン (実線)；フレーズ指令 (実線) とフレーズ成分 (点線)；アクセント指令 (実線) とアクセント成分 (点線)、直線成分 (点線) をそれぞれ示す。