音声の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法

成澤修一[†] 峯松信明[†]
広瀬啓吉[††] 藤崎博也[†††]

藤崎らによる音声の基本周波数パターン（F_0 パターン）生成過程のモデルは，少数のパラメータから実測の F_0 パターンにきわめて近いパターンを生成しうることが知られており，音声合成に広く用いられている．一方，実測の F_0 パターンからモデルのパラメータを抽出することは解析的には解けない逆問題であり，初期値を与え逐次近似を行う必要がある．この場合，高精度のパラメータを迅速に抽出するには適切な初期値の設定が不可欠であるが，従来はこれを人手によって行っていたため，大量の音声資料の自動的処理は困難であった．本論文では，実測の F_0 パターンからパラメータの初期値を自動的に決定し，さらにそれに基づいて高精度のパラメータ抽出を自動的に行う手法を提案する．この手法は，実測された F_0 パターンをいたるところで連続かつ微分可能な曲線によって近似するための処理，得られた曲線からアクセント指令とフレーズ指令のパラメータの初期値を決定するための処理，さらにそれらの初期値をもとに逐次近似によりパラメータの最適値を求める処理，の3段階の処理からなる．共通日本語の男性・女性話者各1名の朗読音声を対象とした実験の結果，男性の朗読音声について，以前に提案された手法では，パラメータ抽出の性能として，指令の再現率78%，精度67%であるのに対し，提案手法によればそれぞれ82%，80%であった．また，女性の朗読音声については，従来手法では再現率60%，精度51%であるのに対し，提案手法ではそれぞれ83%，72%であった．この結果から，本手法の有効性が実証された．

A Method for Automatic Extraction of Parameters of the Fundamental Frequency Contour Generation Model

Shuichi Narusawa,[†] Nobuaki Minematsu,[†] Keikichi Hirose[††] and Hiroya Fujisaki[†††]

The model for the generation process of the fundamental frequency contours (F_0 contours) of speech by Fujisaki et al. is known to be capable of generating F_0 contours quite close to observed natural contours, and is widely used in speech synthesis. The extraction of model parameters from an observed F_0 contour, however, is an inverse problem that cannot be solved analytically, and requires an iterative process starting from a set of initial parameter values. In order to guarantee a rapid convergence to an optimum solution, the process requires appropriate initial values. These initial values have usually been given manually, making it difficult to analyze a large amount of speech material. The present paper proposes a method for automatically extracting the parameter values from a given F_0 contour. The method consists of three steps: approximation of an observed F_0 contour by a curve that is continuous and differentiable everywhere, extraction of initial values for the parameters from the curve, and optimization of the parameters by successive approximation. Analysis of read speech material of common Japanese by a male speaker showed that the recall and precision rates of model command estimation reached respectively 82% and 80% by the proposed method, while the rates obtained by a previous method were 78% and 67%, respectively. The recall and precision rates obtained for a female speaker were respectively 83% and 72% by the proposed method, but were respectively 60% and 51% by the previous method. These results demonstrate the validity of the current approach.

† 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo
†† 東京大学大学院新領域創成科学研究科
Graduate School of Frontier Sciences, The University of Tokyo
††† 東京大学名誉教授
Professor Emeritus, The University of Tokyo

1. はじめに

音声の韻律的特徴は，本来文字言語にも含まれる語義・統語・意味・談話などの情報（言語情報）のみならず，文字言語には陽に含まれない話者の意図や態度に関する情報（パラ言語情報）や，話者の個体差，性別，年齢，感情などに関する情報（非言語情報）をも含んでいる．したがって，高度な音声情報処理を実現するためには，これらの情報と韻律的特徴との関係についての知識の獲得が不可欠であり，そのためには大量の音声データの分析結果が必要である．特に，近年のコーパスベース手法に関連して，韻律的特徴に関する定量的な記述のなされた音声データベースの重要性が増している．

音声の韻律的特徴を表現する客観的な物理量として，特に，基本周波数パターン（以下，F_0 パターン）は，構文や意味の伝達に重要な役割を果たしている．F_0 パターンは日本語を含む多くの言語の抑揚を表し，一般に単語レベルのアクセントに対応する局所的で急激な起伏と，句・節・文レベルのより広い範囲にわたる緩やかな起伏とからなるが，この F_0 パターンが生成される過程のモデル化は，Fujisaki らによりなされている[1]．このモデル（以下，生成過程モデル）は，発話の言語学的情報と密接な関係にある少数のパラメータを与えることによって，実測の音声 F_0 パターンをきわめてよく近似するパターンを生成することが可能で[2]，その音声言語処理における有効性は，音声合成に広く利用されていることからも明らかである．

一方，このモデルを，さらに音声認識を含めてより広く音声言語処理に利用するためには，観測された F_0 パターンを最良近似するモデルのパラメータを抽出し，抽出されたパラメータと言語情報などとの関係を調べ，把握することが重要である．特に，最近のコーパスベース手法を考慮すると，大量の音声データについてのモデルのパラメータ抽出が必要である．しかしながら，観測された F_0 パターンからモデルのパラメータを抽出することは，いわゆる逆問題であって，この場合には解析的に解くことはできず，モデルのパラメータの初期値を出発点とした逐次近似を必要とする．パラメータ値を与えてパターンを計算し，実測パターンとの誤差を何らかの基準の下に評価し，それが最小化するまでパラメータ値を変更して繰り返し計算する手法は，一般に Analysis-by-Synthesis 法（以下，AbS 法）と呼ばれ，解析的な求解が不可能な問題に対して有効であるが，探索するパラメータ空間が多次元の場合には，通常，誤差の極小点が多数存在し，いわゆる局所的な最小値に収束するという危険性がある．このため，的確なモデルパラメータの解を迅速に求めるためには，適切な初期値を与えることが重要となる．このような初期値の設定は，従来，モデルの特徴に関する豊富な知識を有するエキスパート（本論文では以下，単にエキスパートと呼ぶ）によることが多く，大量の音声データの処理は困難であった．

もし，適切な初期値を自動的に求めることが可能になれば，大量の音声データの分析が容易となり，生成過程モデルに基づく大規模な韻律コーパスの構築も可能になると考えられる．このような観点から，すでにモデルパラメータの推定手法がいくつか提案されている．これらについては 3 章に概説するが，基本的に，観測 F_0 パターンから，F_0 の抽出誤りや子音調音による F_0 の乱れなど，モデルで想定していない現象をいかに修正し，あるいは除去するか，F_0 の存在しない無音・無声区間をどのように取り扱うか，アクセント成分とフレーズ成分をいかに分離するかに努力が払われている．上記の F_0 の誤りや乱れを修正・除去した F_0 パターンからパラメータ抽出を行うには，アクセント指令に関するパラメータについては微係数に着目し，推定したアクセント指令から生成するアクセント成分を F_0 パターンから差し引いた差分を用いて，フレーズ指令に関するパラメータを求めるという手順が有効である[3]．ここで問題なのは，従来の手法では，F_0 の抽出誤りや子音調音による F_0 の乱れの除去手法が，F_0 パターンの局所的な連続性に着目したもので不十分であること，またパラメータ抽出プロセスを考慮せずに無音・無声区間を単に線形補間あるいは 2 次スプライン補間[4]していることである．もちろん，アクセント成分とフレーズ成分とを分離することも重要な課題であるが，フレーズ成分の変化は，アクセント成分の変化と比較して緩やかであるため，両者を分離せずにアクセント指令のパラメータ推定を行うことは可能であると考えられる．

このような観点から，本論文では，(1) 観測 F_0 パターンからアクセント指令の抽出に適した曲線を求めるプロセス（以下，前処理），(2) その曲線からパラメータ初期値を抽出するプロセス，さらに (3) 初期値から AbS 法によってパラメータの最適値を求めるプロセス，を一体化した手法を提案し，実際に分析を行ってその有効性を確認する．

以下，まず，2 章で生成過程のモデルを説明し，3 章で従来の手法を整理し，その問題点を明らかにする．次に 4 章で，観測 F_0 パターンに含まれる種々の変動要因について述べた後，5 章で手法の提案を行う．6

章で提案手法によるパラメータの抽出実験を行い，7章でまとめを述べる．

2. F_0 パターン生成過程モデル

藤崎らによる F_0 パターンの生成過程モデルは，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（**図 1**）．

一般に F_0 パターンは，句頭から句末へ向けて緩やかに下降する成分と，語句に対応して急激に上昇下降する成分とからなると解釈される．生成過程モデルではこれを F_0 の対数値をとって定式化する．すなわち，韻律句ごとの下降（文中の句では短区間の上昇が先行）に対応するフレーズ成分，韻律語ごとのアクセント型に対応して上昇・下降するアクセント成分☆，および発話のベースライン成分の和として F_0 パターンを表現する．F_0 パターンを対数軸上で定式化することの正当性は，声帯の伸びと F_0 の対数の変化値が比例関係にあることからすでに示されている[5]．さらに，同一の言語的内容を持つ音声の F_0 パターンの動的形状が，性別や年齢などの話者条件によらずほぼ一定になることからも，F_0 パターンを対数軸上で定式化することは妥当な表現である．したがって本論文では，以下，F_0 パターンとは $\log_e F_0(t)$ を指すものとする．ただし，$F_0(t)$ は時刻 t における基本周波数の値である．

この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．したがって，F_0 パターンは式 (1) で表される：

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

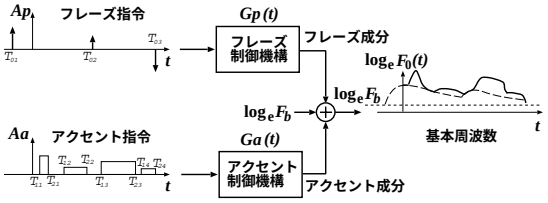

図 1 F_0 パターン生成過程モデル
Fig. 1 A functional model for the process of generating F_0 contours.

☆ 日本語の場合であって，他言語では対応する現象に多少の差異がある．

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り，$G_p(t)$，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あり，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，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 および 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 $\alpha = 3.0\text{rad/s}$，$\beta = 20.0\text{rad/s}$，$\gamma = 0.9$ を用いることができる[2]．さらに，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）の振幅と立上り位置，立下り位置である．

3. パラメータ抽出に関する従来の研究例

現在までに発表されているパラメータ抽出法のうち，生成過程モデルまたはそれに類似したモデルを用いている手法について，**表 1** にその特色と問題点を示す．

これらの手法のうち，“left-to-right AbS[6]” は，発

表 1 これまでに試みられた主なパラメータ抽出法
Table 1 Previous approaches to extraction of model parameters.

手法	特色	問題点
F_0 パターンの 1 次微係数の利用	1 次微係数の最大・最小を検出	フレーズの推定に弱い
left-to-right AbS	2 種の成分を同時推定	イベント検出の信頼度に問題
制御機構の逆フィルタの利用	出力結果が指令そのもの	逆フィルタの最適化が困難
ローパスフィルタの利用	フィルタの設計が容易	遮断周波数の最適化に問題
F_0 パターンに対する LPC 分析	全制御機構を 1 線形系と仮定	残差信号の分析が困難
J-ToBI ラベルの利用	J-ToBI ラベル情報を利用	ラベル付けが人手によることが問題
ハイパスフィルタの利用	フィルタの設計が容易	2 種の成分の完全な分離が困難

話開始からの有限長の AbS 処理を行い，その誤差の累積値がある閾値を超えるごとに新たな韻律イベントを検出し，それを利用して，アクセント指令またはフレーズ指令を推定するが，複雑な構造の文音声に対して，その有効性を示すための十分な実験は行われていない．

“制御機構の逆フィルタの利用[7]” は，アクセント制御機構の逆フィルタとフレーズ制御機構の逆フィルタとを並列に用いて，両指令を同時に推定するものであるが，その有効性を立証するための十分な実験は行われていない．

“ローパスフィルタの利用[8]” は韻律句境界の検出を目指したものであり，必ずしも生成過程モデルによる F_0 パターンの自動分析のためのパラメータ抽出を目指したものではない．

“F_0 パターンに対する LPC 分析[9]” は，予測残差からアクセント指令に類似したパラメータを推定しており，全制御機構を 1 つの線形系で仮定している点においては，生成過程モデルに類似したモデルであるといえるが，フレーズ成分，アクセント成分から構成されるという観点はない．

“J-ToBI ラベルの利用[10]” は生成過程モデルのパラメータ抽出を目指した有用な手法であるが，J-ToBI ラベル付けされた音声データを仮定しているため，提案手法との直接的な比較はできない．

ここでは以下，本手法の出発点となっている “F_0 パターンの 1 次微係数の利用[3]” と，比較的最近発表された “ハイパスフィルタの利用[11]” について，その概要を述べる．なお，6 章ではこの 2 手法と提案手法の有効性を実験によって比較する．

3.1 F_0 パターンの 1 次微係数を用いる手法[3]

パラメータ推定に先立ち，F_0 パターンに窓幅 130 ms の移動平均処理を施して全体を平滑化する．ただし，無声区間の近傍にある F_0 値は無声子音などの影響を受けるため，信頼性の低い値であるとして重みを小さくする．

まず，実測の F_0 パターンを 130 ms にわたって 2 次曲線で近似し，その 1 次微係数を求め，その大きさから，アクセント成分の平坦部，立上り部，立下り部を推定する．続いて，これらの情報を用いて，アクセント指令の大きさ，始点・終点および固有角周波数を推定する．最後に，推定したアクセント成分の寄与を平滑化した F_0 パターンから差し引いて残差 F_0 パターンを得，アクセント指令の始点と，フレーズ指令の立上りおよび極大の時間関係とを用いて，フレーズ成分のパラメータを発話開始時点から順に推定する．

平滑化した F_0 パターンの 1 次微係数の最大・最小の位置をもとにアクセント指令の生起位置を推定することは，F_0 パターンの変曲点を利用することと同等である．しかしながら，元来変曲点を持たない 2 次曲線を用いて平滑化することに問題がある．また，アクセント成分の推定誤りから，残差 F_0 パターンに多くの局所的な変動が残ることがあり，大局的なフレーズ成分に関するパラメータの推定誤りの原因となる．

3.2 ハイパスフィルタを用いる手法[11]

まず，実測の F_0 パターンに対し 2 次のスプライン近似を行い，補間・平滑化する．次に，平滑化後の F_0 パターンを遮断周波数が 0.5Hz のハイパスフィルタにかけ，高周波数成分曲線（以下，HFC）を抽出する．この HFC をアクセント成分と見なし，その形状からアクセント指令の立上り位置と立下り位置を推定する．また，アクセント指令の大きさは，その指令から生成されるアクセント成分が HFC の各山の極大値と一致するように決定する．その後，平滑後の F_0 パターンから HFC を差し引き，低周波数成分曲線（以下，LFC）を求める．LFC をフレーズ成分と見なし，その極小値の位置をフレーズ指令の生起位置とする．フレーズ指令の大きさは，その指令から生成されるフレーズ成分が LFC の起伏の極大値と一致するように決定する．

アクセント成分とフレーズ成分では，前者は相対的に高周波成分が多く，後者は低周波成分が多いが，その周波数成分は重なり合うため，理想的な分離は困難である．また，フィルタ出力の形状からパラメータを推定しているため，アクセント指令の中にフレーズ指令を推定したり，隣接するアクセント指令どうしが重なったりする．そのため，その後の AbS 処理が困難となる．

4. パラメータ抽出を困難にする種々の要因とそれらへの対応の基本的な考え方

どのような手法を用いてパラメータ抽出を行ったとしても，生成過程モデルで考慮されていない種々の要因が実測の F_0 パターンに含まれたままでは，抽出したパラメータに誤りが生ずるのを避けることはできない．したがって，ここでは，パラメータ抽出にとって有害である種々の要因を明白にし，その要因を取り除くための基本的な考え方について述べる．なお，音声波形からの F_0 の抽出法としては，従来から自己相関関数やケプストラムを利用した手法が報告され，最近では精度の高い手法として瞬時周波数に基づく手法も提案されているが，本研究では，文献 12) の遅れ時間

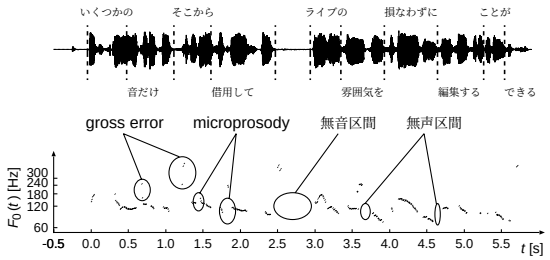

図 2 F_0 パターンの実測値に含まれる種々の変動
Fig. 2 Various types of variations included in the extracted F_0 contour.

比例窓長の自己相関関数を用いた手法を採用し，F_0 を抽出する．

4.1 抽出した F_0 パターンに含まれる変動

実測される F_0 パターンは，式 (1) に示す連続曲線ではなく，測定フレームごとの F_0 値の揺らぎや量子化雑音に加えて，**図 2** に示す種々の変動が存在する．

まず，声帯振動の不規則性や F_0 抽出アルゴリズムの不完全性に起因する F_0 抽出値の局所的な不連続性，音声録音時のノイズの影響などによる無音・無声区間での偽 F_0 などがあるが，パラメータ抽出で特に問題となるのは，F_0 値がこれらの原因によって近傍の値から大きく外れる場合である．これは，孤立して起こる場合と，比較的長い区間にわたって倍ピッチ・半ピッチなどが連続して起こる場合とがある．厳密にいえば，声帯振動の不規則性によって実際に半ピッチなどが生ずる場合もあり，必ずしもすべてが F_0 の抽出誤りではないが，ここでは簡単のため，そのような原因による F_0 の不連続現象も含めて，gross error と呼ぶこととする．

次に，無声破裂・摩擦などの影響による子音と隣接の母音との間の過渡部での不連続かつ起伏の激しい F_0 変動（以下，microprosody）などがある．

さらに，母音の無声化や無声子音による無音・無声区間，句間休止に対応した比較的長い無音区間も存在する．

4.2 パラメータ抽出のための処理についての基本的な考え方

上に述べた種々の要因による変動は生成過程モデルに含まれていないため，パラメータ抽出の妨げとなる．したがって，これらの変動の影響を取り除くための前処理が必要となる．一方，フレーズ成分の時間的変化はアクセント成分よりもはるかに緩やかであるため，F_0 パターンの変曲点の位置はアクセント成分のそれとほぼ一致する．したがって，実測の F_0 パターンから上述の変動要因の影響を除去する前処理を行ったうえで，それを至るところ連続かつ微分可能な 3 次曲線で区分的に近似すれば，その変曲点の位置は，その曲線の 1 次導関数の極値の位置（2 次導関数が 0 となる位置）として，容易に求めることができる．これは，解析的には解けなかった逆問題が，このような近似を行うことによって，1 次方程式の解を求める問題に帰着されたことを意味する．なお，アクセント指令の立上り・立下り位置は，式 (3) の関数の 2 次微分から明白なように，変曲点の位置から $1/\beta$ だけ先行する時点となる．また，アクセント指令の振幅は，変曲点での勾配（1 次微係数）から推定することができる．さらに，これらの手法ですべてのアクセント指令の位置と振幅の初期値を求めたのち，それらの指令によって生成されるアクセント成分を前処理後の F_0 パターンから差し引いたものは，近似的にフレーズ成分と一定の基底値とからなり，また，フレーズ成分は半無限の現象であるので，発話開始から left-to-right 的にフレーズ指令を推定することができる．

5. パラメータ抽出の手順

提案手法は，以下の 4 つの段階からなる．

(1) **実測の F_0 パターンに対する前処理**
gross error を除去し，microprosody を修正した後，短い無音・無声区間を補間する．さらに，その処理後の F_0 パターンを至るところ連続かつ微分可能な区分的 3 次曲線で近似する．本論文では，以下，これを平滑化 F_0 パターンと呼ぶこととする．

(2) **アクセント指令のパラメータの初期値の推定**
平滑化 F_0 パターンを微分し，得られた微分パターンの極大・極小値の位置からアクセント指令の立上り・立下り位置の初期値を求める．アクセント指令の大きさの初期値は，立上り・立下り位置の 1 次微係数から算出する．

(3) **フレーズ指令のパラメータの初期値の推定**
推定したアクセント指令を用いて生成したアクセント成分を平滑化 F_0 パターンから差し引いたパターンからさらに，すでに推定が完了したフレーズ指令を用いて生成したフレーズ成分を差し引いた残差パターンの積分値を計算する．算出した積分値が閾値を超えたら計算開始時点まで戻り，新たにフレーズ指令を挿入し，再びフレーズ成分と残差パターンとの差の積分値を計算する．この処理を発話終了時点まで left-to-right 的に行う．

(4) **パラメータの最適化**
推定した指令から F_0 パターンを生成し，その推定した F_0 パターンと実測の F_0 パターンとの平均二乗誤

差が最小となるように，AbS 法を用いてパラメータを微小変化させる．

以下に，これらの具体的な方法について述べる．

5.1 前処理法

以下で扱う F_0 値は，上述したように，遅れ時間比例窓長の自己相関関数を用いた手法[12]により算出した値で，自己相関関数の値の大きい順に抽出した第 1〜第 5 候補の F_0 値のうち，いずれかの候補値が選定されているものとする．

5.1.1 gross error の修正

(1) 正規化自己相関関数の値の閾値処理による F_0 候補値の選別

全有声フレームにおいて，正規化自己相関関数が一定の閾値（本手法では 0.4）以下の値を持つような F_0 の候補値は，抽出誤りであると見なして，候補値から除外する．

(2) 連続した誤りの修正

F_0 が抽出されている区間 $[i,j]$ の内部に存在するある区間 $[m,n]$ において，$F_0(m)$ が $F_0(m-1)$ の倍ピッチであり，$F_0(n+1)$ が $F_0(n)$ の半ピッチである場合，区間 $[m,n]$ 中のすべての F_0 について，それぞれの第 1〜第 5 候補値と $F_0(m-1)$ との対数尺度での誤差の絶対値を計算し，誤差が最小となる候補値を新たな F_0 値とする．

(3) 孤立した誤りの修正

$(2N+1)$ 個（本手法では $N=2$）の F_0 が連続して抽出されている部分を分析区間とする．また，第 i 分析区間に含まれる F_0 の対数尺度での中央値を $F_0(M|i,2N+1)$，区間内の各 $F_0(k)$ の第 n 候補値を $F_0(S_n|k)$ と表記する．

$F_0(M|i,2N+1)$ と $F_0(S_n|k)$ との対数尺度での誤差 $E(S_n|k)$ を次式

$$E(S_n|k)=\left|\frac{\log_e F_0(S_n|k)}{\log_e F_0(M|i,2N+1)}-1\right| \quad (4)$$

を用いて算出し，この誤差が最小である候補値 $F_0(S_n|k)$ を新たに $F_0(k)$ とする．以上の処理を，F_0 の標本値を N 個だけ移動しながら繰り返し発話終了時点まで行う．

(4) 穴埋め

有声/無声判定の誤りにより，有声区間であるにもかかわらず抽出されなかった数点の F_0 は，母音の無声化や句間休止などによる中断ではなく，抽出誤りによる欠落と見なし，欠落区間前後の F_0 値から対数尺度上で直線的に補間する．

5.1.2 microprosody の除去

無声区間を $[i+1,j-1]$ とする．また，有声区間において，$\log_e F_0(t)$ の 1 次微係数を $G_0(t)$ として以下の処理を行う．

(1) 移動平均処理

F_0 パターンのどの部分が microprosody であるかを的確に検出するために，窓幅 50 ms の移動平均処理を施し，有声区間に含まれるすべての F_0 値を平滑化する．

(2) microprosody の検出・除去

microprosody は無声区間前後で生じる F_0 の急激な局所的変化であるので，無声区間近傍で抽出された N 個（本手法では $N=10$）の $\log_e F_0(t)$ の変化の様子をとらえれば，どの箇所が microprosody であるか検出できる．したがって，以下の処理を施す．

$|G_0(i-n_1)|<|G_0(i)|/2$ を満たす $F_0(i-n_1)$ が存在し，かつ $0<n_1\le N$ である場合，区間 $[i-n_1, i-n_1+1,\cdots,i-1,i]$ に含まれる F_0 を microprosody と見なして除去する．同様にして，$|G_0(j+n_2)|<|G_0(j)|/2$ を満たす $F_0(j+n_2)$ が存在し，かつ $0<n_2\le N$ である場合，区間 $[j,j+1,\cdots,j+n_2-1,j+n_2]$ に含まれる F_0 を microprosody と見なして除去する．ただし，無声区間前後にある有声区間内の F_0 の総数 $Nsum$ が N 個以下である場合は，$N=Nsum/3$ とする．

5.1.3 短い無音・無声区間の補間

microprosody を除去した後の無声区間を $[i+1,j-1]$ と再定義する．無声区間前後の N 個（本手法では $N=5$，$N<5$ の場合は $N=2$）の F_0 を用いて，区間 $[i-N,j+N]$ での $\log_e F_0(t)$ の値を以下の 3 次曲線

$$\log_e F_0(t)=a_0+a_1t+a_2t^2+a_3t^3 \quad (5)$$

で近似し，無声区間 $[i+1,j-1]$ を補間する．また，係数 $[a_0,a_1,a_2,a_3]$ は，無声区間直前の F_0 値の対数，すなわち $\log_e F_0(i)$ とその近傍に存在する連続した N 個の F_0 値の対数，すなわち $\log_e F_0(n)$ $(n=i,i-1,\cdots,i-N)$ の平均の傾き $\overline{G_0(i)}$，および無声区間直後の F_0 値の対数，すなわち $\log_e F_0(j)$ とその近傍に存在する連続した N 個の F_0 値の対数，すなわち $\log_e F_0(m)$ $(m=j,j+1,\cdots,j+N)$ の平均の傾き $\overline{G_0(j)}$ を用いた以下の連立方程式

$$\begin{cases}\log_e F_0(i)=a_0+a_1i+a_2i^2+a_3i^3,\\ \overline{G_0(i)}=a_1+2a_2i+3a_3i^2,\\ \log_e F_0(j)=a_0+a_1j+a_2j^2+a_3j^3,\\ \overline{G_0(j)}=a_1+2a_2j+3a_3j^2,\end{cases} \quad (6)$$

を解くことによって得られる．ただし，本手法において，第 1 フレーズ指令の初期値の生起位置は，発話の

先頭部分付近で F_0 が抽出された位置から $1/\alpha$，つまり 0.33 s だけ戻った位置であるとしているため，それ以上の長さの無声区間に対しては，この補間処理を行わない．ここで，α は，式 (2) に示されているフレーズ制御機構の固有角周波数である．

5.1.4 全体の平滑化

ここまでの前処理により，発話の開始から終了までの時点に関して，句間休止を除くすべての区間において F_0 の値が求められている．これらの区間はすべて有声区間であると見なし，この全有声区間において，F_0 パターンを至るところ連続かつ微分可能な区分的 3 次曲線で近似するために，150 ms 区間ごとに F_0 パターンを式 (5) を用いて近似し，その係数を求める．ただし，それぞれの区間の接合部において，F_0 パターンの振幅値と 1 次微係数の連続性を保証するために，分析区間を 200 ms，近似区間を 150 ms，分析区間の移動幅を 150 ms として，以下に示す 2 つの段階に分けて曲線の係数を求める．ここで，近似区間を 150 ms とした根拠は，平均的な発話速度（1 s/7 モーラ $\simeq$ 0.143 s/モーラ）では，150 ms 程度のフレーム長なら，たかだか 1 個のアクセント指令の立上り位置（または立下り位置）が存在するだけなので，変曲点もたかだか 1 個であることによる．

(1) 発話の開始から 150 ms までの区間（初期分析区間）の平滑化

係数 $[a_0, a_1, a_2, a_3]$ を，最小二乗法により導かれた以下の連立方程式を用いて算出する：

$$\begin{cases} \displaystyle\sum_{i=1}^{N_1} \log_e F_0(i) = N_1 a_0 + \sum_{i=1}^{N_1} a_1 i + \sum_{i=1}^{N_1} a_2 i^2 + \sum_{i=1}^{N_1} a_3 i^3, \\ \displaystyle\sum_{i=1}^{N_1} \log_e F_0(i) i = \sum_{i=1}^{N_1} a_0 i + \sum_{i=1}^{N_1} a_1 i^2 + \sum_{i=1}^{N_1} a_2 i^3 + \sum_{i=1}^{N_1} a_3 i^4, \\ \displaystyle\sum_{i=1}^{N_1} \log_e F_0(i) i^2 = \sum_{i=1}^{N_1} a_0 i^2 + \sum_{i=1}^{N_1} a_1 i^3 + \sum_{i=1}^{N_1} a_2 i^4 + \sum_{i=1}^{N_1} a_3 i^5, \\ \displaystyle\sum_{i=1}^{N_1} \log_e F_0(i) i^3 = \sum_{i=1}^{N_1} a_0 i^3 + \sum_{i=1}^{N_1} a_1 i^4 + \sum_{i=1}^{N_1} a_2 i^5 + \sum_{i=1}^{N_1} a_3 i^6. \end{cases} \quad (7)$$

ここで，N_1 は発話の開始から 200 ms までの区間に含まれる F_0 の標本値の総数である．

(2) 初期分析区間以降の 150 ms 区間の平滑化

係数 $[a_0, a_1, a_2, a_3]$ を，初期分析区間の終わりの時点における F_0 値の対数，すなわち $\log_e F_0(j)$ とその傾き $G_0(j)$ および最小二乗法により導かれた以下の連立方程式を用いて算出する：

$$\begin{cases} \log_e F_0(j) = a_0 + a_1 j + a_2 j^2 + a_3 j^3, \\ G_0(j) = a_1 + 2a_2 j + 3a_3 j^2, \\ \displaystyle\sum_{j=1}^{N_2} \log_e F_0(j) = N_2 a_0 + \sum_{j=1}^{N_2} a_1 j + \sum_{j=1}^{N_2} a_2 j^2 + \sum_{j=1}^{N_2} a_3 j^3, \\ \displaystyle\sum_{j=1}^{N_2} \log_e F_0(j) j = \sum_{j=1}^{N_2} a_0 j + \sum_{j=1}^{N_2} a_1 j^2 + \sum_{j=1}^{N_2} a_2 j^3 + \sum_{j=1}^{N_2} a_3 j^4. \end{cases} \quad (8)$$

ここで，$j = 1$ は $i = 3N_1/4$ と同意であり，N_2 は初期分析区間以降の 200 ms 区間に含まれる F_0 の標本値の総数である．この処理を発話終了時点まで繰り返し行う．

以上の前処理により，実測の F_0 パターンは句間休止を除いて至るところ連続かつ微分可能な区分的 3 次曲線で近似される．

5.2 アクセント指令のパラメータの初期値の推定法

5.2.1 平滑化 F_0 パターンの 1 次微係数の計算

前述のように，F_0 パターンは有声区間全体にわたって平滑化 F_0 パターンで近似されているので，式 (5) を直接微分して 1 次微係数を算出する．

5.2.2 変曲点の検出

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．ただし，求めた極大・極小値の絶対値が閾値（本手法では，極大値が 0.6，極小値が 0.4）以下である場合には無視する．残った極値について，1 個または連続する複数個の同符号の極値を 1 つの集合とし，正の極値の集合の中の最大値および負の極値の集合の中の最小値を発話終了時点まで交互に選定する．

このようにして選定した 1 次微係数の正の極値の位置を平滑化 F_0 パターンの第 1 種変曲点，負の極値の位置を平滑化 F_0 パターンの第 2 種変曲点と定義する．

5.2.3 アクセント指令の生起位置の初期値の推定

隣接する第 1 種・第 2 種変曲点を 1 つの組として，第 1 種変曲点の位置よりも $1/\beta$ だけ先行する時点をアクセント指令の立上り位置の初期値とし，第 2 種変

曲点の位置よりも $1/\beta$ だけ先行する時点を指令の立下り位置の初期値とする．ただし，発話の先頭の韻律語のアクセントが頭高型の場合には，アクセント指令の立上りの位置は発話の開始に先行するため，第1種変曲点としては検出されず，まず第2種変曲点が検出される．このような場合には，その第2種変曲点の位置よりも平均1モーラ長（1s/7モーラ $\simeq$ 0.143s/モーラ）だけ先行する時点を仮想的に第1種変曲点の位置と見なす．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，句間休止が存在する場合は，第2種変曲点が検出されずに第1種変曲点だけが検出されるため，発話の終わりの位置を仮想的に第2種変曲点の位置と見なす．

5.2.4 アクセント指令の大きさの初期値の推定

j 番目のアクセント成分は式 (1) の右辺の第 3 項であり，式 (3) の条件から，第 1 種変曲点の位置 $t_{1j} = T_{1j} + 1/\beta$ では $G_a(t_{1j} - T_{2j}) = 0$ となる．つまり，第1種変曲点での1次微係数は式 (1) の右辺の第3項を1次微分した式にこれらの条件を代入した値であるが，この値は平滑化 F_0 パターンの1次微係数 $G_0(t_{1j})$ そのものである．したがって，第1種変曲点の位置におけるアクセント指令の大きさ A_{a1j} は，次式によって推定できる：

$$A_{a1j} = \frac{e}{\beta} G_0(t_{1j}). \tag{9}$$

ただし，e は自然対数の底である．

同様に，第2種変曲点の位置 $t_{2j} = T_{2j} + 1/\beta$ におけるアクセント指令の大きさ A_{a2j} は次式によって推定できる：

$$A_{a2j} = \frac{e}{\beta} |G_0(t_{2j})|. \tag{10}$$

式 (9)，(10) の A_{a1j} と A_{a2j} とは同一のアクセント指令の振幅を2つの時点で推定したものであるが，平滑化 F_0 パターンにはフレーズ成分の影響も含まれているため，この両者は必ずしも一致しない．したがって，A_{a1j} と A_{a2j} との平均値を求め，それを j 番目の指令の大きさの初期値とする．なお，5.2.3 項で述べたように，第1種または第2種変曲点のいずれか一方が存在しない場合は，その1次微係数を算出することができないため，存在するもう一方の変曲点の位置 t と勾配 $G_0(t)$ から算出した A_{a1j} または A_{a2j} を，便宜上，指令の大きさの初期値とする．

5.3 フレーズ指令のパラメータの初期値の推定法

各発話ごとに，以下の手順に従ってまず第1フレーズ指令のパラメータの初期値を推定し，続いて第2フレーズ指令以降のパラメータの初期値を推定する．なお，各発話ごとに，平滑化 F_0 パターンの開始時点を T_S，j 番目（$j = 1, 2, \cdots, J$）のアクセント指令の立上り位置を T_{1j}，立下り位置を T_{2j} とする．

5.3.1 第1フレーズ指令のパラメータの初期値の推定

(a) 5.2 節で求めたアクセント指令のパラメータの初期値からアクセント成分を計算し，それを平滑化 F_0 パターンから差し引いて，残差パターン $R_0(t)$ を求める．

(b) 区間 $[T_S, T_{21}]$ において，$R_0(t)$ の最大値 R_0max の位置 T_{R_0max} を検出する．

(c) T_{R_0max} より $1/\alpha$ だけ先行する位置を第1フレーズ指令の生起位置 T_{01}（第0近似）とする．

(d) T_{R_0max} の時点では，$R_0(t)$ が第1フレーズ成分と F_b のみから推定され，かつ T_{R_0max} での F_0 の推定値が R_0max と一致すると仮定して，式 (1) を A_{p1} について展開した次式

$$A_{p1} = \frac{e}{\alpha} \{\log_e R_0(T_{R_0max}) - \log_e F_b\} \tag{11}$$

から，第1フレーズ指令の大きさ A_{p1}（第0近似）を求める．ただしここで，F_b は，便宜上，microprosody を除去した F_0 パターンの最小値とする．

(e) 区間 $[T_S - 0.5, T_{21}]$ において，$R_0(t)$ に対するフレーズ成分の逐次近似を行い，第1フレーズ指令の生起位置 T_{01} および大きさ A_{p1}（第1近似）を決定し，これらを第1フレーズ指令のパラメータの初期値とする．

5.3.2 第2フレーズ指令以降のパラメータの初期値の推定

第2フレーズ指令以降のパラメータの初期値の推定は，以下の手順に従い逐次的に行う．ただし，これまでに推定された i 個のフレーズ指令のパラメータの初期値から計算されるフレーズ成分を，$R_0(t)$ から差し引いたものを残差パターン $R_i(t)$ $(i = 1, 2, \cdots)$ と定義する．

(f) 区間 $[T_{2j}, T_{2J}]$ にわたって $R_i(t)$ を計算する．

(g) $R_i(t)$ の積分値を求める．

(h) (g) の値が閾値 θ_1 を超えたら，$R_i(t)$ の積分値が増加し始める時点（(g) の値が閾値 θ_2 $(< \theta_1)$ に達した時点）を $i+1$ 番目のフレーズ指令の生起位置 T_{0i+1}（第0近似）とする．T_{0i+1} が $j+1$ 番目のアクセント指令の内部に存在する場合は，T_{2j} と T_{1j+1} の間で平滑化 F_0 パターンの1次微係数が負から正へと変化する時点を T_{0i+1} とする．この時点が存在しない場合は，便宜上，$(T_{2j} + T_{1j+1})/2$ を T_{0i+1} とする．

また，(g) の値が θ_1 を超えない場合は，もはやフレー

ズ指令は存在しないものとして推定処理を終了する．

(i) $T_{0i+1}+1/\alpha$ の時点での F_0 の推定値が R_imax と一致すると仮定して，(d) と同様に，次式

$$A_{pi+1}=\frac{e}{\alpha}\left\{\log_e R_imax-\log_e F_b\right\} \qquad (12)$$

から，第 $i+1$ フレーズ指令の大きさ A_{pi+1}（第 0 近似）を求める．

(j) 区間 $[T_{2j}, t]$ において，$R_i(t)$ に対するフレーズ成分の AbS を行い，第 $i+1$ フレーズ指令の生起位置 T_{0i+1} および大きさ A_{pi+1}（第 1 近似）を決定し，これらを第 2 フレーズ指令以降のパラメータの初期値とする．

(k) $R_i(t)$ の積分値をリセット（0 に戻す）し，i の値を 1 つ増やして (f) に戻る．

5.4 パラメータの最適化法（AbS 処理）

上述の 5.2 節および 5.3 節で求めたのは，それぞれアクセント指令およびフレーズ指令のパラメータの初期値であるので，逐次近似を行って，抽出したパラメータの最適化を図る必要がある．

具体的には，各分析区間において推定した指令から F_0 パターンを生成し，F_0 パターンの実測値と推定値との平均二乗誤差が最小となるようにパラメータを微小変化させる．ここで，最適化を行う分析区間の始点は指令の生起位置とし，終点は生成する F_0 パターンにパラメータの微小変化の与える影響が無視できる位置とする．また，フレーズ指令の大きさが 0.1 以下となった場合には，その大きさを 0 として指令を除去する．ここで，0.1 を閾値としたのは，文中の小さなフレーズ成分の立て直しに相当する指令の大きさが，0.15 以下のものはほとんどないことに基づいている[13]．

6. パラメータの自動抽出実験

6.1 音声資料とその処理

音声資料は，NHK ラジオ第 1 放送の朗読番組「私の本棚」で男性アナウンサ 1 名による朗読（1 回分，15 分間，全 85 文）の音声を録音したものを使用した．なお，話者の個人差の影響を調べるため，同じテキストを女性話者 1 名に朗読させた音声資料も使用した．

これらの音声を 10kHz，16bit で A/D 変換し，遅れ時間比例窓長の自己相関関数を用いた手法[12]により，10 ms 間隔で F_0 を抽出した．さらに，F_0 パターンに対して前記の前処理を施したのち，得られた平滑化 F_0 パターンからフレーズ指令とアクセント指令のパラメータの初期値を抽出し，AbS 処理を行いパラメータを最適化した．

6.2 実験結果とその評価

以下では，まず，上述の音声資料のうち，男性話者によるものについて論じる．

提案した手法を用いて分析した結果例を**図 3** に示す．分析に用いた音声資料は，“いくつかの音だけそこから借用して，ライブの雰囲気を損なわずに編集することができる.” という文である．

(1) の音声データから (2) の F_0 パターンを抽出した後，前処理として，gross error の修正を行った結果が (3) であり，それに加えて microprosody の除去を行った結果が (4) であり，さらに無声区間の補間および全体の平滑化を行った結果，(5) の平滑化 F_0 パターンを得る．次に (5) を微分して得られる (6) から (7) に示すアクセント指令の初期値を得る．さらに (7) のアクセント指令を用いて生成されるアクセント成分を (5) から除去した残差パターンが (8) であり，これまでに推定されたフレーズ成分をさらに (8) から差し引いたのちの残差の積分値 (9) から (10) に示すフレーズ指令の初期値を得る．この (9) と (10) の処理は反復的に行われるが，ここでは，便宜上，最終結果を 1 つの図として示した．また，推定した (7) と (10) の指令から (11) に示す F_0 パターンを得るが，(7) と (10) の指令を初期値として (5) に対する AbS 処理を行うことにより，(12) の F_0 パターンとそれを生成するフレーズ指令，アクセント指令 (13) を得る．ただし，(11) と (12) 中の点が示すのは (4) の F_0 パターンであり，生成した F_0 パターンがどれほど実測の F_0 パターン（変動を除去したもの）に類似した値であるかを判断するのに役立つものである．(11) の推定された F_0 パターンでは実測値からかなり外れる部分も見られるが，(12) の AbS 処理後の F_0 パターンは，(4) の実測の F_0 パターンから gross error を修正し，microprosody を除去したパターンにきわめて類似した形状である．これを定量的に評価するために，有声フレームに対して，(4) と (12) との間の，また (4) と手動推定によるパラメータから生成した F_0 パターンとの間の 1 標本値あたりの平均二乗誤差を計算すると，それぞれ 0.0016，0.0011 であった．これは，百分率誤差の実効値としては，それぞれ約 4.0%，約 3.3%の差に相当し，小さい量である．つまり，本手法を用いて生成した F_0 値は，全体にわたって実測値にきわめて近い値が得られているといえる．

以上の結果は，提案手法を用いることにより，音声波形から実測値にきわめて近い F_0 パターンを生成するパラメータを自動的に抽出できることを示している．

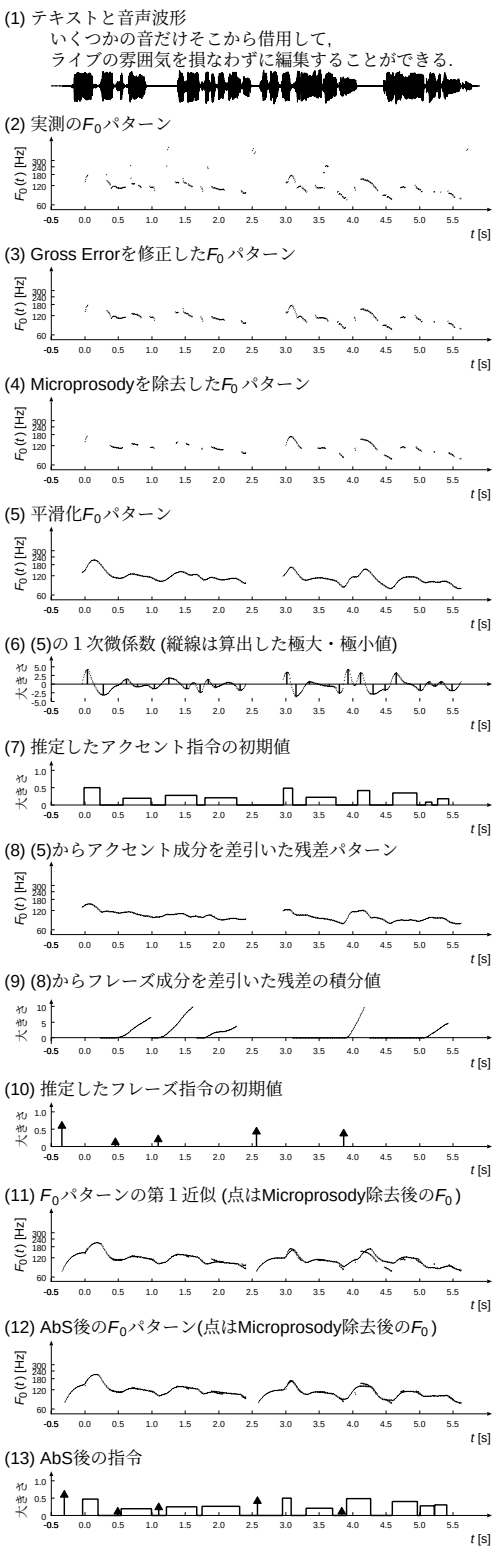

図 3 前処理およびパラメータ自動抽出の結果例
Fig. 3 An example of pre-processing and command estimation.

表 2 男性話者 1 名の音声資料に基づく提案手法と従来手法との比較
Table 2 Comparison of the proposed method and the previous methods based on the speech material by a male speaker.

(a) 従来手法の結果

	手法 A	手法 B
フレーズ指令		
自動抽出数	424	238
正解数	252	91
欠落数	144	305
挿入数	172	147
再現率 [%]	63.6	22.8
精度 [%]	59.4	38.4
アクセント指令		
自動抽出数	607	482
正解数	325	285
欠落数	296	336
挿入数	282	197
再現率 [%]	52.4	45.5
精度 [%]	53.5	59.1

(b) 提案した手法の前処理法を用いた場合の結果

	提案手法	手法 A	手法 B
フレーズ指令			
自動抽出数	418	425	267
正解数	332	280	99
欠落数	64	116	297
挿入数	86	145	168
再現率 [%]	83.8	70.7	24.7
精度 [%]	79.4	65.9	37.1
アクセント指令			
自動抽出数	603	623	469
正解数	492	417	298
欠落数	129	204	323
挿入数	111	206	171
再現率 [%]	79.2	67.1	47.2
精度 [%]	81.6	66.9	63.5

6.3 提案手法の評価および他の手法との比較

全 85 文の音声資料について，提案手法と他の手法[3),11)]を用いて分析した結果を手動分析による結果と比較したものを**表 2** に示す．以下，文献 3) の手法を手法 A，文献 11) の手法を手法 B と呼ぶ．また，表中の数字の単位は個数であるが，再現率（recall, R）と精度（precision, P）については，正解（correct, C），欠落（deletion, D），挿入（insertion, I）を用いた次式

$$R = \frac{C}{C+D} \times 100 \tag{13}$$

$$P = \frac{C}{C+I} \times 100 \tag{14}$$

から算出し，百分率で示す．ここで，$C+D$ は，韻律研究に深く携わっているエキスパートが F_0 パターンを手動で分析し，抽出した指令の数（フレーズ指令：396 個，アクセント指令：621 個）であり，これらがすべて正しく，また，これら以外には指令は存在しな

いものとする．一方，$C+I$ は，それぞれの手法を用いて自動的に抽出した指令の数である．

ただし，フレーズ指令に関しては，手動分析結果のフレーズ指令が存在した韻律節において推定されたものだけを正解とした．また，アクセント指令に関しては，音声資料を朗読した話者の話速における平均モーラ長を 1 モーラとして，推定したアクセント指令の立上り位置が，手動分析結果のアクセント指令の立上り位置を中心とした前後 0.5 モーラの範囲以内に存在し，かつ，推定したアクセント指令の立下り位置が，手動分析結果のアクセント指令の立下り位置を中心とした前後 0.5 モーラの範囲以内に存在するものだけを正解とした．なお，手動分析において振幅の値がわずかに異なる 2 個のアクセント指令が連続して抽出されたものを，自動分析では 1 個のアクセント指令として抽出した場合にも，それを正解と見なした．

ここで用いた再現率と精度は，元来情報検索の分野で用いられている概念であり，いずれも 100%に近いことが望ましいが，アクセント指令についていえば，欠落に関しては，振幅の小さい指令が欠落した場合，また挿入に関しても，1 つの長い指令が振幅のほぼ等しい 2 つの近接した指令で表現された場合には，実用上支障がないので，厳密には個々の欠落・挿入の効果まで考慮する必要がある．また，誤って挿入された指令を，AbS 処理による振幅の閾値処理を用いて除去することは可能であるが，抽出にもれた指令を AbS 処理により復活させることは不可能である．したがって，本研究の目的からは，欠落と挿入とは同等ではなく，欠落を可及的に少なくすることが望ましい．

なお，提案手法では，F_0 パターンの前処理は，パラメータ抽出の処理と一貫して行われるので，前処理の効果を分離して評価するのは困難であるが，手法 A と手法 B に関しては，提案した前処理の効果を調べた．まず表 2 の (a) には，手法 A と手法 B の分析結果を示し，次に，表 2 の (b) には，提案手法の前処理を行った場合の 3 者の分析結果を示す．表 2 の (a) では，アクセント指令の抽出精度を除いて，すべての指標で手法 A の方が良い．

この結果からも明らかなように，パラメータ抽出に先立ち前処理を行うことは重要である．手法 A と B でも gross error の修正や補間・平滑化に関して簡単な前処理は行っているが，細かい処理は手動で行っている．そのため，手法 A では，提案した前処理を施すことによって正解数が増すだけでなく，挿入数も大幅に減り，特にアクセント指令の抽出精度が飛躍的に向上する．手法 B では，前処理の有無に関係なくパラメータの抽出精度が低い☆が，提案した前処理を施すことによる抽出精度の向上はみられる．したがって，提案した前処理法の有効性が定量的に確認された．

また，前処理を行ったとしても，手法 A では 7 割程度の再現率でたかだか 7 割弱程度の精度（フレーズ指令とアクセント指令についての結果の平均）だが，提案手法を用いれば，8 割程度の再現率で 8 割強の精度を得ることが可能である．手法 A では，microprosody を積極的に除去しておらず，また，実測の F_0 パターンを変曲点を持たない 2 次曲線で近似したために，アクセント指令の抽出に失敗した部分が多数存在しており，その結果を利用してフレーズ指令を推定したためにフレーズ指令の推定精度も悪くなる．一方，提案手法は，3 次曲線による平滑化などを行っており，観測した F_0 パターンをパラメータ抽出に適した曲線に修正している．その結果，手法 A よりも良い結果が得られたといえる．

以上は，男性話者 1 名による大量の音声資料の分析結果について詳細に論じたものであるが，当然ながら，提案手法の性能は，F_0 の抽出精度や，話者，発話スタイルなどに依存すると考えられる．男性アナウンサの朗読音声は，“分析しやすい” と考えられる対象ではあるが，このような対象に対してある程度良好な性能が得られるということは，音声合成のための音声データベース構築といった用途を考えると意義のある結果である．なお，異なる話者の朗読音声でどのような性能が得られるかは，手法の評価には重要であるため，女性話者の音声資料から無作為に 10 文を選び，分析を行った．その結果を**表 3** に示す．この場合，エキスパートによる手動分析の結果は，フレーズ指令が 43 個，アクセント指令が 86 個である．この結果から，この女性話者に対しても，従来手法である手法 A に比べ，提案手法の方がパラメータの抽出精度が良いことが確認された．

ここで，表 3 の結果は，無作為に選んだ 10 文の音声資料に対するものであるが，表 2 に示したものとほぼ同等の結果が得られている．他の 10 文の選び方でもほぼ同等の結果が得られる．これから，表 2 中の標本データ量は，本手法の有効性を示すのに十分な量であるといえる．

以上の結果は，提案手法の有効性を実証するものである．

☆ ドイツ語音声を対象として最適化された手法であり，日本語に関して最適化されたものではない．

表 3 女性話者 1 名の音声資料に基づく提案手法と従来手法との比較
Table 3 Comparison of the proposed method and the previous method based on the speech material by a female speaker.

	提案手法	手法 A
フレーズ指令		
自動抽出数	53	57
正解数	36	29
欠落数	7	14
挿入数	17	28
再現率 [%]	83.7	67.4
精度 [%]	67.9	50.9
アクセント指令		
自動抽出数	92	90
正解数	70	46
欠落数	16	40
挿入数	22	44
再現率 [%]	81.4	53.5
精度 [%]	76.1	51.1

7. ま と め

観測された F_0 パターンから生成過程モデルのパラメータを抽出することは，パターン生成の逆問題であり，モデルが複雑なために，その解析的な解法は存在しない．本論文では，まず，実測の F_0 パターンに対して，gross error の修正，microprosody の除去，無音区間の補間，そして至るところ連続かつ微分可能な区分的 3 次曲線を用いた平滑化からなる前処理を施すことにより，この問題を，1 次方程式を解く問題に帰着させうることを示した．次に，この前処理によって得られた平滑化 F_0 パターンの変曲点の位置とそこでの 1 次微係数からアクセント指令のパラメータの初期値を求め，その結果を用いてフレーズ指令のパラメータの初期値を left-to-right 的に求める手順を示した．さらに，それらの初期値を用いた AbS 法による逐次近似を行って，抽出したパラメータの最適化を図り，言語的に意味のない振幅の小さな指令を排除する方法を示した．その結果，1 名の男性話者による 85 文の朗読音声資料に対し，エキスパートによる手動抽出の結果を基準として，再現率 82%，精度 80%が得られた．また，1 名の女性話者による 10 文の朗読音声資料に対しても，再現率 83%，精度 72%が得られた．これらの結果は，従来手法のものに比べても勝っており，提案手法の有効性を実証するものである．

なお，提案手法の有効性を聴覚的に確認するために，3 名の被験者に対して，予備的な聴取実験を行った[14]．その結果，韻律の自然性に問題があると判断されたのは，アクセント指令では 7.9%，フレーズ指令では 6.6%であった．しかし，この実験では被験者の数が少ないため，聴覚的評価に関しては，さらに多くの聴取実験を行い，稿を改めて報告したい．ただ，ここで興味あるのは，聴取実験の結果が 10%以上分析結果より良くなっている点である．これは，提案手法で作成した音声データベースを用いて音声合成をした場合，分析結果から予想されるよりも良好な結果が得られることを示唆するものともいえる．

本論文で提案した手法では，F_0 パターンのみから生成過程モデルのパラメータ値を推定している．しかしながら，音声合成，音声認識などへの応用を考えた場合，言語情報からみて意味のある指令の抽出が重要である．本研究では，指令の位置や大きさに閾値処理を施し，その結果，言語的に無意味な箇所に振幅の小さい指令を検出することを避けているが，この観点からは，発話の言語情報をさらに積極的に利用したパラメータ推定を行うことにも十分な意味がある．具体的には，言語情報を用いてパラメータ推定に制約を加えることにより，より良い推定結果を得ることが期待できる[15]．なお，本論文では，少数の話者による，日本語の朗読音声の分析結果について論じたが，今後は，さらにより多くの話者，多様な発話スタイルの音声資料を対象とし，また，日本語以外の外国語音声，たとえば英語音声も対象として提案手法を適用し，必要な改良を加えて，信頼性が高くかつ適用範囲の広い，パラメータ自動抽出の手法を確立する予定である．

参 考 文 献

1) Fujisaki, H. and Nagashima, S.: A model for synthesis of pitch contours of connected speech, *Annual Report of Engineering Research Institute, University of Tokyo*, Vol.28, pp.53–60 (1969).
2) Fujisaki, H. and Hirose, K.: Analysis of voice fundamental frequency contours for declarative sentences of Japanese, *J. Acoust. Soc. Jpn* (*E*), Vol.5, No.4, pp.233–242 (1984).
3) 広瀬啓吉，藤崎博也，山口幹雄，渡辺泰夫：基本周波数パタンの特徴パラメータの自動推定，日本音響学会昭和 57 年度春季研究発表会講演論文集，Vol.1, pp.93–94 (1983).
4) Hirst, D.: Automatic modelling of fundamental frequency using a quadratic spline function, *Travaux de l'Institut de Phonetique d'Aix*, Vol.15, pp.75–85 (1993).
5) Fujisaki, H.: Dynamic characteristics of voice fundamental frequency in speech and singing — Acoustical analysis and physiological interpretations, *Proc. 4th FASE Symposium*, pp.1–

14 (1981).
6) Geoffrois, E.: A pitch contour analysis guided by prosodic event detection, *Proc. Eurospeech '93*, Vol.2, pp.793–796 (1993).
7) 藤崎博也，大野澄雄，和田 豊：音声の基本周波数パタン生成過程モデルのパラメータ自動推定の一方法，日本音響学会平成6年度春季研究発表会講演論文集，Vol.1, pp.17–18 (1995).
8) Sakurai, A. and Hirose, K.: Detection of phrase boundaries in Japanese by low-pass filtering of fundamental frequency contours, *Proc. ICSLP '96*, Vol.2, pp.817–820 (1996).
9) Mersdorf, J., Rinscheid, A., Brüggen, M. and Schmidt, K.U.: Coding of large intonational units by linear prediction, *ESCA Workshop on Intonation: theory*, pp.18–20 (1997).
10) 平井俊男，樋口宣男：韻律ラベリングシステム J-ToBI のラベル情報を用いた重畳型基本周波数制御モデルパラメータの自動抽出，信学論，Vol.J81–D-II, No.6, pp.1058–1064 (1998).
11) Mixdorff, H.: A novel approach to the fully automatic extraction of Fujisaki model parameters, *Proc. ICASSP 2000*, Vol.3, pp.1281–1284 (2000).
12) Hirose, K., Fujisaki, H. and Seto, S.: A scheme for pitch extraction of speech using autocorrelation function with frame length proportional to the time lag, *Proc. ICASSP '92*, Vol.1, pp.149–152 (1992).
13) Hirose, K. and Fujisaki, H.: A system for the synthesis of high-quality speech from texts on general weather conditions, *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, Vol.E76-A, No.11, pp.1971–1980 (1993).
14) Fujisaki, H. and Narusawa, S.: Automatic extraction of model parameters from fundamental frequency contours of speech, *Proc. 2nd Plenary Meeting and Symposium on Prosody and Speech Processing*, pp.133–138 (2002).
15) Sakurai, A. and Hirose, K.: Designing a parameter-based prosodic speech database, *Proc. 1999 Oriental COCOSDA Workshop*, pp.5–8 (1999).

(平成13年11月21日受付)
(平成14年 4 月16日採録)

成澤 修一

1977年生．1999年東京理科大学基礎工学部電子応用工学科飛び級．2001年同大学大学院基礎工学研究科電子応用工学専攻修了．現在，東京大学大学院情報理工学系研究科電子情報学専攻在籍．音声の韻律に関する研究に従事．日本音響学会会員．

峯松 信明（正会員）

1966年生．1995年東京大学大学院工学系研究科電子工学専攻博士課程修了．博士（工学）．同年豊橋技術科学大学情報工学系助手．2000年東京大学大学院工学系研究科助教授，2001年同大学院情報理工学系研究科助教授．2002年瑞国KTH客員研究員．音声認識，音声分析，音声応用，音声知覚，および音声合成の研究に従事．電子情報通信学会，日本音響学会，日本音声学会，人工知能学会各会員．

広瀬 啓吉（正会員）

1949年生．1972年東京大学工学部電気工学科卒業．1977年同大学院博士課程修了．工学博士．同年東京大学工学部電気工学科講師．1994年同電子工学科教授．1996年東京大学大学院工学系研究科電子情報工学専攻教授．1999年4月より新領域創成科学研究科基盤情報学専攻教授．1987年米国MIT客員研究員．音声言語情報処理分野一般についての研究開発に従事，特に韻律に着目した研究．IEEE，米国音響学会，ISCA，日本音響学会，電子情報通信学会，人工知能学会，言語処理学会等各会員．

藤崎　博也（正会員）

1930 年生．1954 年東京大学工学部電気工学科卒業．1958 年 MIT 大学院留学．1959 年同電子工学研究所助手．1960 年スウェーデン王立工科大学客員研究員．1962 年東京大学講師．1963 年助教授．1973 年教授．1991 年名誉教授，東京理科大学教授．工学博士．専門は音声科学・音声言語処理・自然言語処理・人工知能等．著書「音声科学」，「Recent Research Toward Advanced Man-Machine Interface Through Spoken Language」等．日本音響学会名誉会員，電子情報通信学会およびアメリカ音響学会フェロー．電気通信学会および電気学会論文賞，電子通信学会業績賞，IEEE Third Millennium Medal 等受賞．東京都科学技術功労者表彰．