Vol.2012-SLP-92 No.9
2012/7/20

F_0 パターン生成過程の確率モデルによる藤崎モデルパラメータの推定

吉里 幸太[1,a)]　亀岡 弘和[1,2,b)]　齋藤 大輔[1,c)]　嵯峨山 茂樹[1,d)]

概要：本稿では，統計的手法に基づく様々な音声アプリケーションに基本周波数 (F_0) パターンを組み入れるための強力な枠組みを確立するため，F_0 パターンの生成過程を表現したモデルである藤崎モデルの確率モデル化を行う．また，提案モデルを用いて観測 F_0 パターンから藤崎モデルパラメータを推定するためのアルゴリズムを導出する．実音声を用いた定量評価実験を通して，提案手法が既存手法を上回る推定性能を持っていることを明らかにした．

キーワード：F_0 パターン，確率モデル，藤崎モデル，隠れマルコフモデル，EM アルゴリズム

Estimation of Fujisaki Model Parameters from Speech Signals Using Probabilistic Model of Speech F_0 Contours

Kota Yoshizato[1,a)]　Hirokazu Kameoka[1,2,b)]　Daisuke Saito[1,c)]　Shigeki Sagayama[1,d)]

Abstract: This paper proposes a stochastic model of speech F_0 contours, based on the stochastic formulation of the Fujisaki model. It will allow open the door to incorporating the well-founded F_0 contour model into various statistical speech processing problems. We also propose a well-behaved algorithm for estimating the Fujisaki model parameters from a raw F_0 contour. We quantitatively evaluated the performance of our method in terms of an Fujisaki-model parameter estimation accuracy using real speech data. Experimental results revealed that our method was superior to a state-of-the-art Fujisaki model parameter extractor.

Keywords: speech F_0 contours, stochastic model, Fujisaki model, hidden Markov model, EM algorithm

1. はじめに

音声の抑揚を表す物理量である基本周波数（F_0）の時間変化パターン（以下 F_0 パターン）は，構文や意図の伝達に関する様々な種類の非言語情報を含んでおり，発話を用いたコミュニケーションにおいて重要な役割を果たしている．また近年，利用可能な音声データベースが増加してきたことに伴って，統計的手法を用いた音声信号処理に関する研究が盛んに行われ，多くの成果を挙げている．そこで本稿では，音声認識，話者認識，音声合成，言語識別といった統計的手法に基づく数々の音声アプリケーションに将来的に韻律モデルを組み込んでいくための強力な枠組みを作ることを目標に，F_0 パターンの確率モデルを提案する．

藤崎の F_0 パターン生成過程モデル（藤崎モデル）とは，甲状軟骨の運動に注目して F_0 パターンの生成過程を説明した，力学的モデルである [1]．藤崎モデルは，フレーズ指令，アクセント指令と呼ばれる物理学的，生理学的に意味のある少数のパラメータを用いて，観測 F_0 パターンを極めてよく近似するモデルとして広く知られている．また，このパラメータは発話の言語学的意図とも密接にかかわっ

[1] 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, Japan
[2] NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, NTT Corporation, Japan
[a)] yoshizato@hil.t.u-tokyo.ac.jp
[b)] kameoka@hil.t.u-tokyo.ac.jp
[c)] dsaito@hil.t.u-tokyo.ac.jp
[d)] sagayama@hil.t.u-tokyo.ac.jp

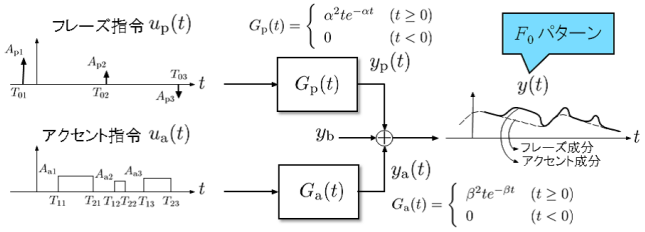


図 1 藤崎モデルのブロック線図
Fig. 1 Block diagram of Fujisaki model

ており，藤崎モデルのパラメータを操作することで発話意図を直接的に制御することもできる．こうした利点から，本研究で提案する F_0 パターンの確率モデルは，藤崎モデルをベースにする．

将来的に F_0 パターンの確率モデルを音声アプリケーション内で利用するにあたって，音声コーパスから藤崎モデルのパラメータを自動的に学習できれば非常に有用であると考えられる．観測 F_0 パターンから藤崎モデルのパラメータを推定する研究はこれまでにも行われてきたが，推定問題の解析的な複雑さから，限定的な成果しか挙げられていない [2], [3], [4]．そこで本研究のもう一つの目的は，提案した確率モデルを用いて観測 F_0 パターンから藤崎モデルのパラメータを推定する強力なアルゴリズムを提案することである．

我々は以前，本稿で提案するものとは別の，藤崎モデルをベースにした F_0 パターンの生成過程を表現する確率モデルと，藤崎モデルのパラメータ推定アルゴリズムを提案したことがある [5], [6]．その過去手法と比較しての利点を挙げながら提案モデルと提案アルゴリズムについて説明を行い，最後に定量評価実験を通して提案手法の有用性を示すのが本稿のおおまかな流れである．具体的な構成は，以下のようになっている．2 章では，提案モデルのベースになった藤崎モデルについて簡単に説明を行う．3 章では，本稿で提案する F_0 パターン生成過程の確率モデルについて解説する．4 章では，提案モデルを用いた藤崎モデルパラメータの推定アルゴリズムを定式化する．5 章では，提案手法を用いて実音声からの藤崎モデルパラメータ推定問題を解き，提案手法の有効性を確認する．最後に 6 章で，まとめと今後の展望を述べる．

2. 藤崎の F_0 パターン生成過程モデル

藤崎モデル [1] では，甲状軟骨の二つの独立な運動（平行移動運動と回転運動）にそれぞれ伴う声帯の伸びの合計が F_0 の時間的変化をもたらすと解釈され，声帯の伸びと F_0 パターンの対数値 $y(t)$ が比例関係にあるという仮定に基づいて F_0 パターンがモデル化される．甲状軟骨の平行移動運動によって生じる F_0 パターン $y_p(t)$ をフレーズ成分，回転運動によって生じる F_0 パターン $y_a(t)$ をアクセント成分と呼ぶ．藤崎モデルでは，音声の F_0 パターン $y(t)$ は，これらの成分に声帯の物理的制約によって決まるベースライン成分 y_b を足し合わせたものとして，

$$y(t) = y_p(t) + y_a(t) + y_b \tag{1}$$

と表現される．フレーズ成分は短区間の上昇のあと緩やかに下降していく成分であり，F_0 パターンの大域的な変化を表している．一方でアクセント成分は急激に上昇したあと急激に下降する成分であり，F_0 パターンの局所的な変化を表している．そのため，多くの言語に共通して，フレーズ成分は句単位の比較的穏やかな音調を，アクセント成分は語または音節単位での比較的急激な音調を実現する役割を担っていると考えられている．

これら二つの成分は二次の臨界制動系の出力であるとしてモデル化されており，

$$y_p(t) = G_p(t) * u_p(t), \tag{2}$$

$$G_p(t) = \begin{cases}\alpha^2 te^{-\alpha t} & (t\geq 0)\\ 0 & (t<0)\end{cases}, \tag{3}$$

$$y_a(t) = G_a(t) * u_a(t), \tag{4}$$

$$G_a(t) = \begin{cases}\beta^2 te^{-\beta t} & (t\geq 0)\\ 0 & (t<0)\end{cases}, \tag{5}$$

と表される（$*$ は時刻 t に関する畳み込み演算）．ここで $u_p(t)$ はフレーズ指令関数と呼ばれ，デルタ関数（フレーズ指令）の列からなり，$u_a(t)$ はアクセント指令関数と呼ばれ，矩形波（アクセント指令）の列からなる．また α と β はそれぞれフレーズ制御機構，アクセント制御機構の固有角周波数であり，話者や発話内容によらず，おおよそ $\alpha = 3$ rad/s，$\beta = 20$ rad/s 程度であることが経験的に知られている．以上をまとめたものを，図 1 に示した．

フレーズ指令とアクセント指令は発話の言語学的意図と密接に関わっており，観測 F_0 パターンからこれらの指令列を推定することは重要な課題である．しかしこれらの指令列には，発話の最初にはフレーズ指令が生起する，フレーズ指令は二連続で生起しない，異なる二つの指令は同時刻に生起しない，という制約条件があり，藤崎モデルパラメータの解析的な扱いを難しくしている要因となっている．この課題に対処し，藤崎モデルを確率モデル化する方法について，次章以降で説明していく．

3. F_0 パターン生成過程の確率モデル

本章では，藤崎モデルの離散時間表現に基づいて，音声の F_0 パターン生成過程を確率モデルとして表現する．

まずはフレーズ指令関数 $u_p[k]$ とアクセント指令関数 $u_a[k]$ を（k は離散時刻のインデックス）確率的に表現する方法を述べる．本研究では，指令関数に付随する各種制

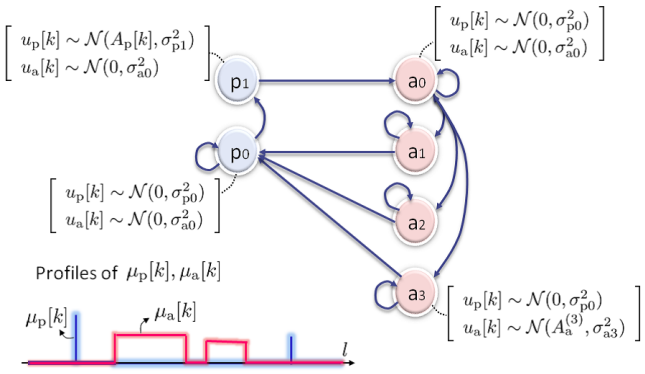


図 2 指令を出力する HMM
Fig. 2 Command function modeling with HMM

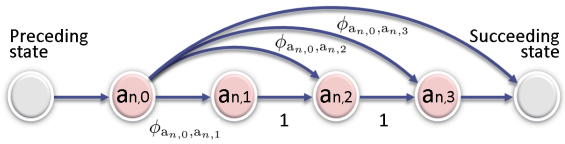


図 3 状態 a_n の 4 つの小状態への分割
Fig. 3 The splitting of state a_n into 4 substates

約をモデルに組み入れるために，フレーズ，アクセント指令のペア $\boldsymbol{o}[k] = (u_p[k], u_a[k])^\mathsf{T}$ を出力する隠れマルコフモデル（HMM）を用いる．具体的な HMM の構成は，図 2 に示した．この HMM では，出力される指令列 $\{\boldsymbol{o}[k]\}_{k=1}^K$ は，各時刻ごとにガウス分布に従い，

$$\boldsymbol{o}[k] \sim \mathcal{N}\left(\boldsymbol{o}[k]; \boldsymbol{\nu}[k], \boldsymbol{\Upsilon}[k]\right), \tag{6}$$

$$\boldsymbol{\nu}[k] = \begin{bmatrix} \mu_p[k] \\ \mu_a[k] \end{bmatrix}, \quad \boldsymbol{\Upsilon}[k] = \begin{bmatrix} v_p^2[k] & 0 \\ 0 & v_a^2[k] \end{bmatrix}, \tag{7}$$

と確率的に表現される．ここで平均ベクトル $\boldsymbol{\nu}[k]$ と分散共分散行列 $\boldsymbol{\Upsilon}[k]$ は，HMM の状態遷移の結果として定まる値である．

加えて，自己遷移の持続長をパラメータ化するために，それぞれの状態をいくつかの小状態に分割することを考える．なおこのとき，おのおのの小状態は全て同じ出力分布を持ち，小状態の数は十分大きな値となるようにしておく．図 3 に状態 a_n を分割した例を示した．例えばこの図のように全ての $m \neq 0$ に対して $a_{n,m}$ から $a_{n,m+1}$ への状態遷移確率を 1 に設定することで，$a_{n,0}$ から $a_{n,m}$ への遷移確率が状態 a_n が m ステップだけ持続する確率に対応し，アクセント指令の持続長を柔軟に制御できるようになる．同様に p_1 と p_0 と a_0 も小状態に分割することで，フレーズ指令の持続長と指令間の間隔の長さの分布をパラメータ化することが可能になる．こうした分割をふまえて，以後は改めて $p_0 = \{p_{0,0}, p_{0,1}, \ldots\}$，$a_0 = \{a_{0,0}, a_{0,1}, \ldots\}$，$a_n = \{a_{n,0}, a_{n,1}, \ldots\}$ と表記する．

提案する HMM の構成を定式化すると次のように書ける．

出力値系列: $\{\boldsymbol{o}[k]\}_{k=1}^K$
状態集合: $\mathcal{S} = \{p_0, p_1, a_0, \ldots, a_N\}$
状態系列: $\{s_k\}_{k=1}^K$
状態出力分布: $P(\boldsymbol{o}[k]|s_k) = \mathcal{N}(\boldsymbol{o}[k]; \boldsymbol{\nu}[k], \boldsymbol{\Upsilon}[k])$

$$\boldsymbol{\nu}[k] = \begin{cases} (0,0)^\mathsf{T} & (s_k \in p_0, a_0) \\ (A_p[k], 0)^\mathsf{T} & (s_k = p_1) \\ (0, A_a^{(n)})^\mathsf{T} & (s_k \in a_n) \end{cases}$$

$$\boldsymbol{\Upsilon}[k] = \begin{bmatrix} \sigma_{p,s_k}^2 & 0 \\ 0 & \sigma_{a,s_k}^2 \end{bmatrix}$$

状態遷移確率: $\phi_{i',i} = \log P(s_k = i | s_{k-1} = i')$

状態系列 $\boldsymbol{s} = \{s_k\}_{k=1}^K$ が与えられたとき，この HMM はフレーズ指令関数 $u_p[k]$ とアクセント指令関数 $u_a[k]$ のペアを出力する．式 (2) と式 (4) で示した通り，$u_p[k]$ と $u_a[k]$ はそれぞれ $G_p[k]$ と $G_a[k]$ というフィルタに畳み込まれてフレーズ成分 $x_p[k]$ とアクセント成分 $x_a[k]$ が出力される．これを式で表すと，

$$x_p[k] = u_p[k] * G_p[k], \tag{8}$$

$$x_a[k] = u_a[k] * G_a[k], \tag{9}$$

と書ける（$*$ は離散時刻 k に関する畳み込み演算）．このとき，F_0 パターン $x[k]$ は

$$x[k] = x_p[k] + x_a[k] + u_b, \tag{10}$$

と三種類の成分の重ね合わせで書ける．ただし u_b は時刻によらないベースライン成分である．

標準日本語のような非声調言語においては，フレーズ指令関数とアクセント指令関数は常に非負の値をとらなければならないという制約がある．我々が以前提案したモデル [5], [6] では，$u_p[k]$，$u_a[k]$，u_b を全て潜在変数（周辺化されたパラメータ）として扱っていたため，こうした非負制約を陽に導入することができなかった．一方で本モデルでは，これらの変数をモデルパラメータとして扱っているため，制約条件を陽に導入できるという利点がある．

実音声においては，常に信頼のできる F_0 の値が観測できるとは限らない．例えば音声データからピッチ抽出によって得られた F_0 の推定値は，無声区間においては全く信頼できない値である．藤崎モデルのパラメータ推定を行うにあたっては，信頼のおける観測区間の F_0 値のみを考慮に入れて，そうでない区間は無視することが望ましい．そこで，提案モデルに観測 F_0 値の時刻 k における不確かさの程度 $v_n^2[k]$ を導入する．具体的には，観測 F_0 値 $y[k]$ を，真の F_0 値 $x[k]$ とノイズ成分 $x_n[k] \sim \mathcal{N}\left(0, v_n^2[k]\right)$ との重ね合わせで

$$y[k] = x[k] + x_n[k] \tag{11}$$

と表現することで，信頼のおける区間かどうかに関わらず

全ての観測区間を統一的に扱えるようにした．

簡単のため，今後は $\phi_{i',i}$，u_b，$\sigma^2_{\mathrm{p},i}$，$\sigma^2_{\mathrm{a},i}$，$v^2_\mathrm{n}[k]$，α，β を定数とみなすことにすると，$x_\mathrm{n}[k]$ を周辺化することで，出力値系列 $\boldsymbol{o} = \{\boldsymbol{o}[k]\}_{k=1}^K$ が与えられたときの $\boldsymbol{y} = \{y[k]\}_{k=1}^K$ の確率密度関数

$$P(\boldsymbol{y}|\boldsymbol{o}) = \prod_{k=1}^K \mathcal{N}(y[k]; x[k], v^2_\mathrm{n}[k]),$$
$$x[k] = G_\mathrm{p}[k] * u_\mathrm{p}[k] + G_\mathrm{a}[k] * u_\mathrm{a}[k] + u_\mathrm{b}. \quad (12)$$

が得られる．状態系列 $\boldsymbol{s} = \{s_k\}_{k=1}^K$ と指令の振幅を表すパラメータ群 $\boldsymbol{\theta} = \{\{A_\mathrm{p}[k]\}_{k=1}^K, \{A_\mathrm{a}^{(n)}\}_{n=1}^N,\}$ が与えられたとき，式 (6) より出力値系列 $\boldsymbol{o}$ は

$$P(\boldsymbol{o}|\boldsymbol{s},\boldsymbol{\theta}) = \prod_{k=1}^K \mathcal{N}(\boldsymbol{o}[k]; \boldsymbol{\nu}[k], \boldsymbol{\Upsilon}[k]) \quad (13)$$

に従って生成される．また，$P(\boldsymbol{s})$ は状態遷移確率の積として

$$P(\boldsymbol{s}) = \phi_{s_1} \prod_{k=2}^K \phi_{s_k, s_{k-1}} \quad (14)$$

と書ける．なお本稿では，$\boldsymbol{\theta}$ は一様に分布すると仮定する．

4. パラメータ推定アルゴリズム

本章では，期待値最大化（EM）アルゴリズムを用いて，$\boldsymbol{y}$ が与えられたとき $P(\boldsymbol{o},\boldsymbol{\theta}|\boldsymbol{y})$ を局所最大化する $\boldsymbol{o}$ と $\boldsymbol{\theta}$ を求める最大事後確率（MAP）推定を行うための反復アルゴリズムについて説明する．提案アルゴリズムでは，$\boldsymbol{s}$ を潜在変数として扱い，$\boldsymbol{s}$ についての周辺化 $P(\boldsymbol{o},\boldsymbol{\theta},\boldsymbol{s}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{o})P(\boldsymbol{o}|\boldsymbol{s},\boldsymbol{\theta})P(\boldsymbol{s})$ を考える．このとき補助関数（Q 関数）は

$$Q(\boldsymbol{o},\boldsymbol{\theta},\boldsymbol{o}',\boldsymbol{\theta}') = \sum_{\boldsymbol{s}} P(\boldsymbol{s}|\boldsymbol{y},\boldsymbol{o}',\boldsymbol{\theta}') \log P(\boldsymbol{o},\boldsymbol{\theta},\boldsymbol{s}|\boldsymbol{y})$$
$$\stackrel{c}{=} \log P(\boldsymbol{y}|\boldsymbol{o}) + \sum_{\boldsymbol{s}} P(\boldsymbol{s}|\boldsymbol{y},\boldsymbol{o}',\boldsymbol{\theta}') \log P(\boldsymbol{o}|\boldsymbol{s},\boldsymbol{\theta})P(\boldsymbol{s}), \quad (15)$$

と書ける．ここで $\stackrel{c}{=}$ は定数部分を除いて一致することを意味する．

局所最適解を求めるための反復アルゴリズムは，次のステップを繰り返すことで得られる．まず E ステップでは Forward-Backward アルゴリズムを用いて $P(\boldsymbol{s}|\boldsymbol{y},\boldsymbol{o}',\boldsymbol{\theta}')$ を計算し，M ステップでは $Q(\boldsymbol{o},\boldsymbol{\theta},\boldsymbol{o}',\boldsymbol{\theta}')$ を増加させる $\boldsymbol{o}$ と $\boldsymbol{\theta}$ を求め，その $\boldsymbol{o}$ と $\boldsymbol{\theta}$ を改めて $\boldsymbol{o}'$ と $\boldsymbol{\theta}'$ に代入して次の反復に進む．

[7] のアイデアを用いることで，$Q(\boldsymbol{o},\boldsymbol{\theta},\boldsymbol{o}',\boldsymbol{\theta}')$ を増加させるステップで $\boldsymbol{o}$ に非負制約を導入することが可能である．ジェンセンの不等式を適用すると，

$$-\Big(\sum_{i\in\{\mathrm{p,a,b}\}}\sum_l G_i[k-l]u_i[l]\Big)^2 \geq -\sum_{i\in\{\mathrm{p,a,b}\}}\sum_l \frac{G_i^2[k-l]u_i^2[l]}{\lambda_{i,k,l}}, \quad (16)$$

という不等式が得られる．ただし $G_\mathrm{b}[k] = \delta[k]$（クロネッカーのデルタ）であり，また $\lambda_{i,k,l} \geq 0$ は補助変数であって $\sum_i \sum_l \lambda_{i,k,l} = 1$ を満たす．この不等式を $Q(\boldsymbol{o},\boldsymbol{\theta},\boldsymbol{o}',\boldsymbol{\theta}')$ の下限を求めるために適用することができ，するとその下限関数を最大化する $\boldsymbol{o}$（非負制約付き）と λ を簡単に解析的に求めることが可能になる．こうして求めた値によって $Q(\boldsymbol{o},\boldsymbol{\theta},\boldsymbol{o}',\boldsymbol{\theta}')$ の値は減少しないことが保証されている．

以上の反復アルゴリズムを収束するまで十分な回数繰り返した後，Viterbi アルゴリズムを用いることで，最適な状態系列 $\boldsymbol{s}$ を求めることができる．

5. 実験

本稿の研究の重要な成果は，藤崎モデルを確率モデルとして表現することに成功したことである．我々は，数多くの統計的手法に基づく音声アプリケーションに提案モデルを組み込むことによって，将来的には韻律を扱うする強力な手法が得られると考えている．そのためには，スペクトル特徴量と同じようにして，藤崎モデルのパラメータであるフレーズ，アクセント指令列が音声コーパスから自動的に学習できると非常に便利である．この点において，確率モデルとして定式化した提案モデルと提案アルゴリズムは，たとえば [4] のような統計的でない手法よりも優れていると言えるだろう．しかし，提案アルゴリズムを用いた実音声からの藤崎モデルパラメータの推定性能が，既存手法の性能を上回っているかどうかはまだ明らかでない．そこで本章では，提案手法のパラメータ推定性能を定量的に評価するための実験を行う．

詳しい実験条件を以下に記す．本実験で実音声データとして用いたのは，ATR 日本語音声データベースの B セット [8] である．これは 503 文の音素バランス文からなる音声データベースであり，我々はその中から一人の男性話者 (MHT) を選択した．また，その音声データベースに対して，ある韻律研究の専門家が手動で求めたフレーズ，アクセント指令列を正解データとして用いた．提案手法の入力として与える観測 F_0 パターンを音声データから抽出する手法には，我々が以前提案したアルゴリズムを用いた [9]．定数のパラメータは，それぞれ $N = 10$，離散時刻のサンプリング間隔 $t_0 = 8$ ms，$\alpha = 3.0$ rad/s，$\beta = 20.0$ rad/s，$v^2_\mathrm{p}[k] = 0.2^2$，$v^2_\mathrm{a}[k] = 0.1^2$，無声区間では $v^2_\mathrm{n}[k] = 10^{15}$，有声区間では $v^2_\mathrm{n}[k] = 0.2^2$，そして u_b は有声区間における $\log F_0$ の最小値に設定した．モデルパラメータの初期値は，非統計的な手法 [4] を用いて推定した値を用いた．EM アルゴリズムの反復回数は 20 回とした．HMM の小状態

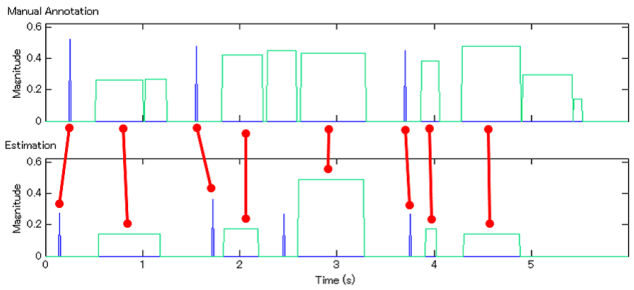


図 4 指令列のマッチング例
Fig. 4 An example of command sequence matching.

の個数や遷移確率 $\phi_{i',i}$ は，No.1 から No.200 までの 200 文の正解データから学習して決定した．パラメータ推定実験は，No.201 から No.503 までの 303 文を対象にして行った．

推定パラメータを評価する方法として，観測 F_0 パターンの再現性と，言語学的な妥当性の二つを考慮した．これらは一般にトレードオフの関係にある．例えば，短い区間に細かく大量の指令を立てれば観測 F_0 パターンを非常によく再現することができるが，そうして作った指令列は言語学的に妥当なものであるとは言えない．そこで本実験では，提案手法によって得られた推定パラメータが言語学的に十分妥当なものでありつつ，観測 F_0 パターンの再現性が非常に高いことを確認することを目的とする．

観測 F_0 パターンの再現性の評価基準には，観測 F_0 パターンと推定指令列から再構成された F_0 パターンとの平均二乗誤差（$\log F_0$[Hz] RMSE）を用い，この値が小さいほど再現性が高いとした．言語学的な妥当性の評価基準には，検出率という値を用い，これが大きいほど言語学的に妥当なパラメータであるとした．検出率は以下のように定義される．図 4 に例を示したように，推定パラメータ列と正解パラメータ列を比較して，指令単位でのマッチングをとる．指令と指令のマッチングがとれる条件は，二つの指令が同種の指令であること（フレーズ指令同士またはアクセント指令同士）と，二つの指令の時間のずれが $S = 0.3$ 秒以下であることとした．ただし，アクセント指令に関しては生起時刻と終了時刻の平均を基準にした．また，二つのマッチングは時刻に関して交差していてはならない．マッチングがとれた指令同士の距離を 1，そうでないときの距離を 0 として，これらの条件を満たしなおかつ距離最大になるようなマッチングは，動的計画法によって求めることができる．推定実験に用いた 303 文全てに対してこのマッチングをとったとき，マッチングの総数を N_{M} とする．また，推定パラメータ列における指令の総数を N_{E}，正解パラメータ列における指令の総数を N_{A} とおく．ここで，挿入エラー E_{I} を $(N_{\mathrm{E}} - N_{\mathrm{M}})/N_{\mathrm{A}}$ と，脱落エラー E_{D} を $(N_{\mathrm{A}} - N_{\mathrm{M}})/N_{\mathrm{A}}$ と定義し，最終的な検出率 D は $1 - E_{\mathrm{I}} - E_{\mathrm{D}}$ であると定義した．なお，この検出率の定義では指令の振幅を考慮に入れていない．これは，フレーズ，アクセント指令の振幅はベースライン成分の値に強く依存するが，このベースライン成分の値が提案手法と正解データで大きく異なるためである．具体的には，提案手法ではベースライン成分の値 u_{b} を有声区間における $\log F_0$ の最小値に設定しているが，正解データでは常に $\log 60$ Hz に固定しており，提案手法で u_{b} の値を固定すると推定性能が落ちることが確認されたためである．

表 1 検出率と $\log F_0$ RMSE
Table 1 Detection rates and $\log F_0$ RMSE

	検出率	$\log F_0$[Hz] RMSE
提案手法	0.695	0.0611
比較手法	0.688	0.1719

提案手法を用いたパラメータ推定結果と，比較手法として選んだ最新のパラメータ推定アルゴリズム（非統計的手法）[4] を用いた推定結果を表 1 にまとめた．この結果を見れば分かる通り，提案手法の検出率は比較手法と同程度である一方で，提案手法の $\log F_0$ RMSE の値は比較手法を大きく下回っている．つまり，提案手法を用いた実音声からのフレーズ，アクセント指令列の推定は，既存手法に匹敵する言語学的な妥当性を満たしつつ，観測 F_0 パターンの再現性では既存手法を上回る性能を持っていることが確認できた．

図 5 に，No.211 と No.353 を例に，比較手法 [4] と我々が以前提案した過去手法 [6] と提案手法のパラメータ推定結果を示した．この図からも，提案手法の観測 F_0 の再現性は，比較手法，過去手法のいずれと比較しても高いことが分かる．

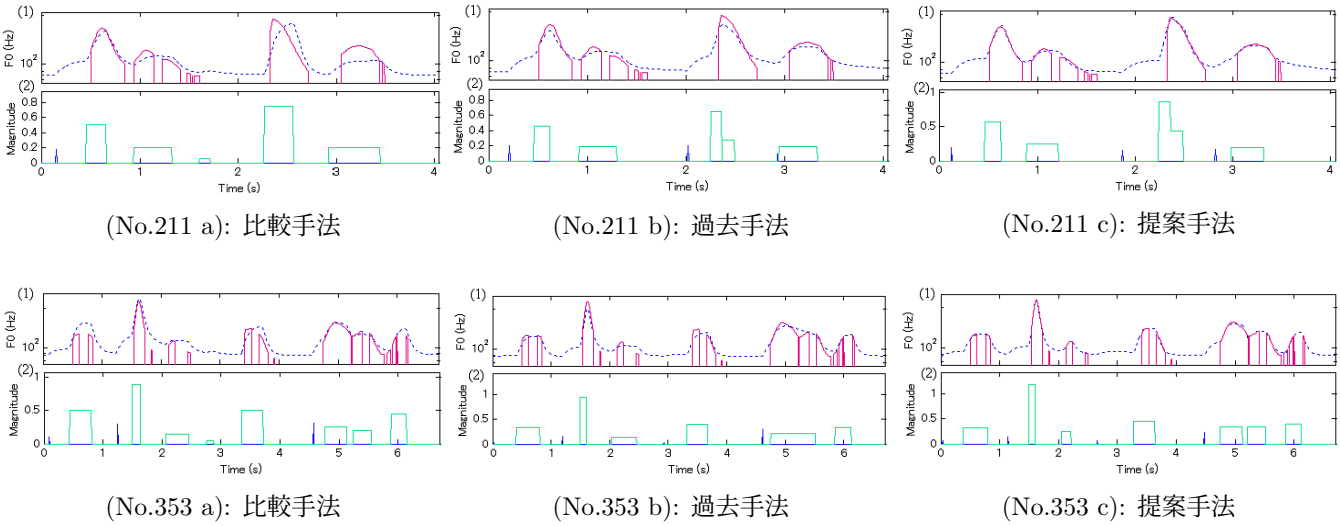
6. おわりに

本稿では，藤崎モデルをベースにした F_0 パターン生成過程の確率モデルを提案し，その提案モデルを用いて音声の F_0 パターンから藤崎モデルのパラメータを推定するアルゴリズムを提案した．また，ATR 日本語音声データベースを用いた定量評価実験を通して，提案手法が言語学的な妥当性を満たしつつ観測 F_0 パターンの再現性の高いフレーズ，アクセント指令列を推定できることを確認した．

今後は，統計的手法に基づく数々の音声アプリケーションに提案モデルを通して韻律モデルを組み込むことを目標に研究を進めていきたいと考えている．

7. 謝辞

本研究の実験では，東京大学の広瀬啓吉教授が ATR の音声データに手動で付与した藤崎モデルのパラメータを用いた．これを作成した同氏に，強い感謝の意を表する．



(No.211 a): 比較手法　(No.211 b): 過去手法　(No.211 c): 提案手法

(No.353 a): 比較手法　(No.353 b): 過去手法　(No.353 c): 提案手法

図 5 (1) 有声区間の観測 F_0 パターン（実線）と推定パラメータから再構築した F_0 パターン（点線）　(2) 推定フレーズ，アクセント指令列

Fig. 5 (1) An observed F_0 contour in voiced regions (in solid line) and the estimated F_0 contours (in dotted line) along with (2) the estimated phrase and accent commands

8. References

[1] H. Fujisaki, *In Vocal Physiology: Voice Production, Mechanisms and Functions*, Raven Press, 1988.

[2] H. Fujisaki and K. Hirose, "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] H. Mixdorf, "A novel approach to the fully automatic extraction of fujisaki model parameters," in *Proc. ICASSP*, 2000, vol. 3, pp. 1281–1284.

[4] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," in *Proc. ICASSP*, 2002, pp. 509–512.

[5] H. Kameoka, J. L. Roux, and Y. Ohishi, "A statistical model of speech F_0 contours," in *Proc. SAPA*, 2010, pp. 43–48.

[6] K. Yoshizato, H. Kameoka, D. Saito, and S. Sagayama, "Statistical approach to fujisaki-model parameter estimation from speech signals and its quantitative evaluation," in *Proc. Speech Prosody 2012*, 2012, pp. 175–178.

[7] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. ICASSP*, 2009, pp. 45–48.

[8] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[9] H. Kameoka, "Statistical speech spectrum model incorporating all-pole vocal tract model and F_0 contour generating process model," in *Tech. Rep. IEICE*, 2010, in Japanese.