F_0 パターン生成過程の統計的モデルによる音声信号からのフレーズ・アクセント指令の推定*

☆吉里幸太 [1]，亀岡弘和 [1,2]，齋藤大輔 [1]，嵯峨山茂樹 [1]
([1] 東大院・情報理工, [2]NTT CS 研)

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．F_0 パターンは，甲状軟骨の運動によって生じる声帯の伸びに応じて決まる量である．ゆえに，音声の F_0 パターンにはある種の物理的制約が課せられる．音声アプリケーションで抑揚を扱う場合，この物理的制約を考慮に入れた F_0 パターンの生成モデルを考えることは非常に有用である．

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する，力学的モデルである [1]．藤崎モデルは，発話の言語学的意図と密接に関わる少数のパラメータを与えることで，実測の F_0 パターンに極めて近い F_0 パターンを表現することが可能である．

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処理に関して盛んに研究がおこなわれるようになり，多くの成果を挙げている．そうした成果に加えて，例えば，ある音声の集合から抑揚の付き方の傾向を学習でき，音声合成や音声認識に反映できるような統計的枠組が確立されれば極めて有用であろう．

そこで我々は，統計的モデルをベースにした音声認識，音声合成，言語識別，話者認識といった現代の多く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ことを目標にして，F_0 パターンを統計的に扱う枠組みを作る研究を進めている．以前我々は，隠れマルコフモデル (Hidden Markov Model; HMM) を用いて藤崎モデルのパラメータの生成過程を確率的に扱うためのモデルを提案した [2]．本稿ではまず，このモデルを実音声に適用する上で考慮すべき点について議論し，それに基づきモデルの改良を行う．そして，その提案モデルが藤崎モデルのパラメータを統計的に扱ううえで十分な性能を持っていることを確認するために，実音声の F_0 パターンを入力として藤崎モデルのパラメータを推定する問題を解いて定量評価を行う．

2 藤崎の F_0 パターン生成過程モデル

藤崎の F_0 パターン生成過程モデル (藤崎モデル) とは，甲状軟骨は平行移動と回転運動の二つの独立な運動を行い，それぞれの運動によって伸びた声帯の長さの和が $\log F_0$ の値に比例する，という仮定をもとに F_0 パターンの生成過程を表現した力学的モデルである [1]．藤崎モデルでは，甲状軟骨の平行移動によって生じる F_0 パターンをフレーズ成分 $y_{\mathrm{p}}(t)$，回転運動によって生じる F_0 パターンをアクセント成分 $y_{\mathrm{a}}(t)$ と呼び (t は時刻)，これらの和に声帯の物理的性質によって決まる定数値であるベースライン成分 y_{b} を加えたものが F_0 パターンの対数値 $y(t)$ であるとする．$y_{\mathrm{p}}(t)$ と $y_{\mathrm{a}}(t)$ は，それぞれフレーズ指令と呼ばれるパルス波の列 $u_{\mathrm{p}}(t)$ と，アクセント指令と呼ばれる矩形波の列 $u_{\mathrm{a}}(t)$ を入力とした臨界制動の二次線形系により表現される．任意の時刻においてフレーズ指令とアクセント指令が同時に正の値をとることはないという制約のもとで，これらの値の関係は

$$y(t) = y_{\mathrm{p}}(t) + y_{\mathrm{a}}(t) + y_b, \tag{1}$$

$$y_{\mathrm{p}}(t) = G_{\mathrm{p}}(t) * u_{\mathrm{p}}(t),\ y_{\mathrm{a}}(t) = G_{\mathrm{a}}(t) * u_{\mathrm{a}}(t), \tag{2}$$

$$G_{\mathrm{p}}(t) = \begin{cases} \alpha^2 t e^{-\alpha t} & (t \geq 0) \\ 0 & (t < 0) \end{cases}, \tag{3}$$

$$G_{\mathrm{a}}(t) = \begin{cases} \beta^2 t e^{-\beta t} & (t \geq 0) \\ 0 & (t < 0) \end{cases} \tag{4}$$

と書ける．ここで，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．

なお，フレーズ成分は急激な上昇のあと緩やかに下降していく成分で，アクセント成分は急激な上昇下降をなす成分であり，日本語を含む多くの言語では，前者が句単位での大局的な音調を，後者が語や音節単位での局所的な音調を表現する役割を担っていると考えられている．

3 統計的 F_0 パターン生成過程モデル

我々は以前，隠れマルコフモデル (Hidden Markov Model; HMM) を用いて藤崎モデルのパラメータの生成過程を確率的に扱うためのモデルを提案した [2]．本章では，このモデルを実音声に適用する上で考慮すべき点を 3 つ挙げ，これらを解決する改良モデルを定式化する．

3.1 藤崎モデルの離散時間表現

連続時間システムのフレーズ制御機構とアクセント制御機構に対して後退差分変換を施すことで，藤崎モデルの離散時間表現を得ることができる．k を離散時刻のインデックスとし，$y_{\mathrm{p}}[k]$, $u_{\mathrm{p}}[k]$, $y_{\mathrm{a}}[k]$, $u_{\mathrm{a}}[k]$ をそれぞれ $y_{\mathrm{p}}(t)$, $u_{\mathrm{p}}(t)$, $y_{\mathrm{a}}(t)$, $u_{\mathrm{a}}(t)$ の離散時間表現とすると，これらの値の関係は

$$u_{\mathrm{p}}[k] = a_0 y_{\mathrm{p}}[k] + a_1 y_{\mathrm{p}}[k-1] + a_2 y_{\mathrm{p}}[k-2], \tag{5}$$

$$a_2 = (\psi - 1)^2,\ a_1 = -2\psi(\psi - 1),\ a_0 = \psi^2, \tag{6}$$

$$u_{\mathrm{a}}[k] = b_0 y_{\mathrm{a}}[k] + b_1 y_{\mathrm{a}}[k-1] + b_2 y_{\mathrm{a}}[k-2], \tag{7}$$

$$b_2 = (\varphi - 1)^2,\ b_1 = -2\varphi(\varphi - 1),\ b_0 = \varphi^2 \tag{8}$$

と書ける．ただし $\psi = 1 + 1/(\alpha t_0)$,$\varphi = 1 + 1/(\beta t_0)$ であり，t_0 はサンプリング周期である．

3.2 藤崎モデルの統計モデル化

フレーズ指令 $u_{\mathrm{p}}[k]$ とアクセント指令 $u_{\mathrm{a}}[k]$ は同時に正の値をとらないという制約を満たしつつ，これ

---

*Estimation of phrase and accent commands from speech signals using statistical model of speech F_0 contours. by YOSHIZATO Kota, KAMEOKA Hirokazu, SAITO Daisuke, SAGAYAMA Shigeki (The University of Tokyo)

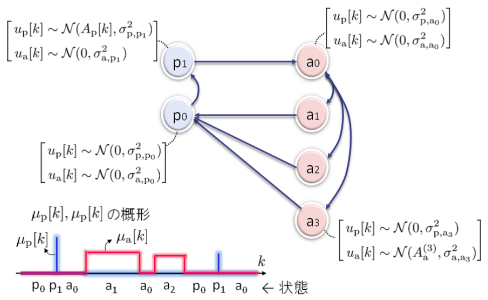


Fig. 1　HMM の構成

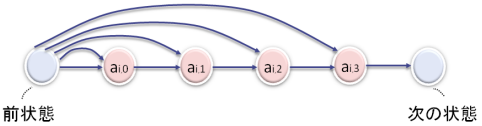


Fig. 2　HMM の状態の分割

らの指令を生成する確率モデルを得るため，藤崎モデルの指令列が，HMM の各状態からの出力 $\boldsymbol{o}[k] = (u_\mathrm{p}[k], u_\mathrm{a}[k])^\mathrm{T}$ として表せると仮定する．提案する統計的 F_0 パターン生成過程モデルの HMM の構成を，Fig. 1 に示す．出力 $\boldsymbol{o}[k]$ は正規分布に従い，

$$\boldsymbol{o}[k] \sim \mathcal{N}\left(\boldsymbol{\nu}[k], \boldsymbol{\Upsilon}[k]\right), \tag{9}$$

$$\boldsymbol{\nu}[k] = \begin{bmatrix}\mu_\mathrm{p}[k] \\ \mu_\mathrm{a}[k]\end{bmatrix}, \quad \boldsymbol{\Upsilon}[k] = \begin{bmatrix}\sigma_\mathrm{p}^2[k] & 0 \\ 0 & \sigma_\mathrm{a}^2[k]\end{bmatrix}, \tag{10}$$

と書けるとする．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決まる値である．

また，フレーズ指令とアクセント指令が生起するタイミングやアクセント指令の持続長には音声の重要な韻律特徴が含まれており，これらを統計的に学習できることが望ましい．そこで，HMM の各状態における停留時間長をより柔軟にパラメトライズできるようにするため，HMM の各々の状態を，全く同じ出力分布を持ついくつかの小状態に分割する．分割の例として，状態 a_i の分割を Fig. 2 に示す．任意の l に対して，小状態 $\mathrm{a}_{i,l}$ から小状態 $\mathrm{a}_{i,l+1}$ への状態遷移確率を 1 に設定すると，アクセント指令が持続長 j を持つ確率が，前状態から $\mathrm{a}_{i,j}$ への状態遷移確率で表現できる．a_i と同様の分割を，p_0, a_0 に対しても行う．なお，この状態分割は，[2] のモデルに新たに施された 1 つ目の改良点である．

提案モデルの HMM 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: $\mathcal{S} = \{\mathrm{p}_0, \mathrm{p}_1, \mathrm{a}_0, \cdots, \mathrm{a}_N\}$
状態系列: $\{s_k\}_{k=1}^K$
状態出力分布: $P(\boldsymbol{o}[k]|s_k = i) = \mathcal{N}(\boldsymbol{\nu}[k], \boldsymbol{\Upsilon}[k])$

$$\boldsymbol{\nu}[k] = \begin{cases} (0,0)^\mathrm{T} & (i = \mathrm{p}_0, \mathrm{a}_0) \\ (A_\mathrm{p}[k], 0)^\mathrm{T} & (i = \mathrm{p}_1) \\ (0, A_\mathrm{a}^{(n)})^\mathrm{T} & (i = \mathrm{a}_n) \end{cases}$$

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

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

また，ベースライン成分の離散時間表現も，$y_\mathrm{b}[k] \sim \mathcal{N}\left(\mu_\mathrm{b}, \sigma_\mathrm{b}^2\right)$ と正規分布に従うと仮定する．

さらに，実音声には F_0 の値が定義されない無声区間が存在するので，これに対処するために，観測 F_0 パターンの離散時刻 k における不確かさの程度 $\sigma_\mathrm{n}^2[k]$ という変数を導入する．$\sigma_\mathrm{n}^2[k]$ の値を無声区間では大きく，それ以外の区間では小さく設定することで，提案モデルで有声区間と無声区間とを統一的に扱うために，$y[k]$ を，真の F_0 の値と $y_\mathrm{n}[k] \sim \mathcal{N}\left(0, \sigma_\mathrm{n}^2[k]\right)$ の和 $y[k] = y_\mathrm{p}[k] + y_\mathrm{a}[k] + y_\mathrm{b}[k] + y_\mathrm{n}[k]$ であると考える．この不確かさの導入が，[2] のモデルに新たに施された 2 つ目の改良点である．

指令や成分の時系列データを，まとめて $\boldsymbol{u}_\mathrm{p} = (u_\mathrm{p}[1], \cdots, u_\mathrm{p}[K])^\mathrm{T}$, $\boldsymbol{u}_\mathrm{a} = (u_\mathrm{a}[1], \cdots, u_\mathrm{a}[K])^\mathrm{T}$, $\boldsymbol{\mu}_\mathrm{p} = (\mu_\mathrm{p}[1], \cdots, \mu_\mathrm{p}[K])^\mathrm{T}$, $\boldsymbol{\mu}_\mathrm{a} = (\mu_\mathrm{a}[1], \cdots, \mu_\mathrm{a}[K])^\mathrm{T}$, $\boldsymbol{y}_\mathrm{p} = (y_\mathrm{p}[1], \cdots, y_\mathrm{p}[K])^\mathrm{T}$, $\boldsymbol{y}_\mathrm{a} = (y_\mathrm{a}[1], \cdots, y_\mathrm{a}[K])^\mathrm{T}$, $\boldsymbol{y}_\mathrm{b} = (y_\mathrm{b}[1], \cdots, y_\mathrm{b}[K])^\mathrm{T}$, $\boldsymbol{y}_\mathrm{n} = (y_\mathrm{n}[1], \cdots, y_\mathrm{n}[K])^\mathrm{T}$, $\boldsymbol{y} = (y[1], \cdots, y[K])^\mathrm{T}$ と書くと，$\boldsymbol{u}_\mathrm{p}$ と $\boldsymbol{u}_\mathrm{a}$ は，

$$\boldsymbol{u}_\mathrm{p} = \boldsymbol{A}\boldsymbol{y}_\mathrm{p}, \ \boldsymbol{u}_\mathrm{a} = \boldsymbol{B}\boldsymbol{y}_\mathrm{a} \tag{11}$$

と表すことができる．ここで，

$$\boldsymbol{A} = \begin{bmatrix} a_0 & & & O \\ a_1 & a_0 & & \\ a_2 & a_1 & a_0 & \\ & \ddots & \ddots & \ddots \\ O & & a_2 & a_1 \ a_0 \end{bmatrix}, \boldsymbol{B} = \begin{bmatrix} b_0 & & & O \\ b_1 & b_0 & & \\ b_2 & b_1 & a_0 & \\ & \ddots & \ddots & \ddots \\ O & & b_2 & b_1 \ b_0 \end{bmatrix} \tag{12}$$

である．

簡単のため $\phi_{i',i}$, μ_b $\sigma_{\mathrm{p},i}^2$, $\sigma_{\mathrm{a},i}^2$, σ_b^2, $\sigma_\mathrm{n}^2[k]$, α, β を全て，観測 F_0 パターンに応じて決まる定数とすると，藤崎モデルのパラメータ $\Theta = \{\{A_\mathrm{p}[k], s[k]\}_{k=1}^K, \{A_\mathrm{a}^{(n)}\}_{n=1}^N\}$ が与えられたとき $\boldsymbol{y} = (y[1], \cdots, y[K])^\mathrm{T}$ を出力する条件付確率 $P(\boldsymbol{y}|\Theta)$ は

$$P(\boldsymbol{y}|\Theta) = \frac{|\boldsymbol{\Sigma}^{-1}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\mathrm{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\},$$

$$\boldsymbol{\mu} = \boldsymbol{A}^{-1}\boldsymbol{\mu}_\mathrm{p} + \boldsymbol{B}^{-1}\boldsymbol{\mu}_\mathrm{a} + \mu_\mathrm{b}\boldsymbol{1}, \tag{13}$$

$$\boldsymbol{\Sigma} = \boldsymbol{A}^{-1}\boldsymbol{\Sigma}_\mathrm{p}\left(\boldsymbol{A}^\mathrm{T}\right)^{-1} + \boldsymbol{B}^{-1}\boldsymbol{\Sigma}_\mathrm{a}\left(\boldsymbol{B}^\mathrm{T}\right)^{-1} + \boldsymbol{\Sigma}_\mathrm{b} + \boldsymbol{\Sigma}_\mathrm{n}$$

と書ける．ここで，

$$\boldsymbol{\Sigma}_\mathrm{p} = \begin{bmatrix}\sigma_\mathrm{p}^2[1] & & O \\ & \ddots & \\ O & & \sigma_\mathrm{p}^2[K]\end{bmatrix}, \boldsymbol{\Sigma}_\mathrm{a} = \begin{bmatrix}\sigma_\mathrm{a}^2[1] & & O \\ & \ddots & \\ O & & \sigma_\mathrm{a}^2[K]\end{bmatrix},$$

$$\boldsymbol{\Sigma}_\mathrm{b} = \begin{bmatrix}\sigma_\mathrm{b}^2 & & O \\ & \ddots & \\ O & & \sigma_\mathrm{b}^2\end{bmatrix}, \boldsymbol{\Sigma}_\mathrm{n} = \begin{bmatrix}\sigma_\mathrm{n}^2[1] & & O \\ & \ddots & \\ O & & \sigma_\mathrm{n}^2[K]\end{bmatrix} \tag{14}$$

である．

4　パラメータ推定アルゴリズム

本章では，3 章で提案した統計的 F_0 パターン生成過程モデルを用いて，観測 F_0 パターン $\boldsymbol{y}$ が与えられたときに $P(\boldsymbol{y}|\Theta)$ を最大化する Θ を求める問題を解くことを考える．この問題の大域的最適解を求めるのは難しいが，$\boldsymbol{x} = (\boldsymbol{y}_\mathrm{p}^\mathrm{T}, \boldsymbol{y}_\mathrm{a}^\mathrm{T}, \boldsymbol{y}_\mathrm{b}^\mathrm{T}, \boldsymbol{y}_\mathrm{n}^\mathrm{T})^\mathrm{T}$ を完全データと見なして EM アルゴリズムによる不完全データ問題に帰着することで，局所最適解を求めることはできる．このとき，$\boldsymbol{x}$ が与えられたもとでの Θ の対数尤度関数は

$$\log P(\boldsymbol{x}|\Theta) \overset{\mathrm{c}}{=} \frac{1}{2}\log|\boldsymbol{\Lambda}^{-1}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^\mathrm{T}\boldsymbol{\Lambda}^{-1}(\boldsymbol{x}-\boldsymbol{m}), \tag{15}$$

と書ける．ただし，

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{y}_{\mathrm{p}} \\ \boldsymbol{y}_{\mathrm{a}} \\ \boldsymbol{y}_{\mathrm{b}} \\ \boldsymbol{y}_{\mathrm{n}} \end{bmatrix},\ \boldsymbol{m} = \begin{bmatrix} \boldsymbol{A}^{-1}\boldsymbol{\mu}_{\mathrm{p}} \\ \boldsymbol{B}^{-1}\boldsymbol{\mu}_{\mathrm{a}} \\ \mu_{\mathrm{b}}\mathbf{1} \\ \mathbf{0} \end{bmatrix}, \tag{16}$$

$$\boldsymbol{\Lambda}^{-1} = \begin{bmatrix} \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{A} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{a}}^{-1}\boldsymbol{B} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{\Sigma}_{\mathrm{b}}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{\Sigma}_{\mathrm{n}}^{-1} \end{bmatrix}$$

である．また，Q 関数 $Q(\Theta, \Theta')$ は，

$$\begin{aligned} Q(\Theta, \Theta') \stackrel{\mathrm{c}}{=} \frac{1}{2}\Big[&\log|\boldsymbol{\Lambda}^{-1}| - \mathrm{tr}(\boldsymbol{\Lambda}^{-1}\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}|\boldsymbol{y};\Theta']) \\ &+ 2\boldsymbol{m}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta'] - \boldsymbol{m}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{m}\Big] + \log P(\Theta) \end{aligned} \tag{17}$$

と書ける．ここで，$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta]$ と $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}|\boldsymbol{y};\Theta]$ は，$\boldsymbol{H} = [\boldsymbol{I}, \boldsymbol{I}, \boldsymbol{I}, \boldsymbol{I}]$ を用いると，$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}$ より，

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta] = \boldsymbol{m} + \boldsymbol{\Lambda}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^{\mathrm{T}})^{-1}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{m}), \tag{18}$$

$$\begin{aligned} \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}|\boldsymbol{y};\Theta] = \boldsymbol{\Lambda} - \boldsymbol{\Lambda}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^{\mathrm{T}})^{-1}\boldsymbol{H}\boldsymbol{\Lambda} \\ + \mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta]\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta]^{\mathrm{T}} \end{aligned} \tag{19}$$

と書ける．これらが E ステップで更新すべき値である．

$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta]$ を 4 つの $K \times 1$ 行列に分割して，各小行列を $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta] = (\bar{\boldsymbol{x}}_{\mathrm{p}}^{\mathrm{T}}, \bar{\boldsymbol{x}}_{\mathrm{a}}^{\mathrm{T}}, \bar{\boldsymbol{x}}_{\mathrm{b}}^{\mathrm{T}}, \bar{\boldsymbol{x}}_{\mathrm{n}}^{\mathrm{T}})^{\mathrm{T}}$ とすると，M ステップの更新式は次のように書ける．

1) 状態系列: Q 関数の中から $s := \{s_k\}_{k=1}^{K}$ に関係する項だけを抜き出すと，

$$\begin{aligned} \mathcal{I}(s) := -\frac{1}{2}\sum_{k=1}^{K}(\boldsymbol{o}[k] - \boldsymbol{\nu}[k])^{\mathrm{T}}\boldsymbol{\Upsilon}^{-1}(\boldsymbol{o}[k] - \boldsymbol{\nu}[k]) \\ + \log P(s_1) + \sum_{k=2}^{K}\log P(s_k|s_{k-1}) \end{aligned} \tag{20}$$

となる．ここで，$\boldsymbol{o}[k] := ([\boldsymbol{A}\bar{\boldsymbol{x}}_{\mathrm{p}}]_k, [\boldsymbol{B}\bar{\boldsymbol{x}}_{\mathrm{a}}]_k)^{\mathrm{T}}$ である．ただし $[\cdot]_k$ はベクトルの k 番目の要素を表す．$\mathcal{I}(s)$ を最大化する状態系列 $\{s_k\}_{k=1}^{K}$ は，動的計画法を用いて効率的に求めることができる．

2) フレーズ・アクセント指令の大きさ: Q 関数の値を最大化する $A_{\mathrm{p}}[k]$ と $A_{\mathrm{a}}^{(n)}$ の値は，

$$A_{\mathrm{p}}[k] = [\boldsymbol{A}\bar{\boldsymbol{x}}_{\mathrm{p}}]_k\ (k \in \mathcal{T}_{\mathrm{p}_1}), \mathcal{T}_{\mathrm{p}_1} = \{k|s_k = \mathrm{p}_1\}, \tag{21}$$

$$A_{\mathrm{a}}^{(n)} = \frac{1}{|\mathcal{T}_{\mathrm{a}_n}|}\sum_{k\in\mathcal{T}_{\mathrm{a}_n}}[\boldsymbol{B}\bar{\boldsymbol{x}}_{\mathrm{a}}]_k, \mathcal{T}_{\mathrm{a}_n} = \{k|s_k = \mathrm{a}_n\} \tag{22}$$

と書ける．

式 (20) から (22) の中には，$\boldsymbol{A}\bar{\boldsymbol{x}}_{\mathrm{p}}$ と $\boldsymbol{B}\bar{\boldsymbol{x}}_{\mathrm{a}}$ の計算が含まれている．この計算は，フレーズ成分 $\bar{\boldsymbol{x}}_{\mathrm{p}}$ とアクセント成分 $\bar{\boldsymbol{x}}_{\mathrm{a}}$ から，それを生成するフレーズ指令 $\bar{\boldsymbol{u}}_{\mathrm{p}}$ とアクセント指令 $\bar{\boldsymbol{u}}_{\mathrm{a}}$ を求めることに相当する．しかし，例えば日本語のように，指令の大きさは必ず 0 以上でなくてはならないという制約を持つ言語も存在する．こうした言語を扱う際に，単純に $\boldsymbol{A}\bar{\boldsymbol{x}}_{\mathrm{p}}$ や $\boldsymbol{B}\bar{\boldsymbol{x}}_{\mathrm{a}}$ を計算するだけでは，計算結果として得られるベクトルは非負の要素を含みうる．これを避けるために，$\boldsymbol{A}\bar{\boldsymbol{x}}_{\mathrm{p}}$ と $\boldsymbol{B}\bar{\boldsymbol{x}}_{\mathrm{a}}$ を直接計算する代わりに，$\|\boldsymbol{A}^{-1}\bar{\boldsymbol{u}}_{\mathrm{p}} - \bar{\boldsymbol{x}}_{\mathrm{p}}\|_2^2$ と $\|\boldsymbol{B}^{-1}\bar{\boldsymbol{u}}_{\mathrm{a}} - \bar{\boldsymbol{x}}_{\mathrm{a}}\|_2^2$ を最小化する $\bar{\boldsymbol{u}}_{\mathrm{p}}$ と $\bar{\boldsymbol{u}}_{\mathrm{a}}$ の値を非負制約のもとで求める問題を考える．これらは，$G_{\mathrm{p}}[k]$ と $G_{\mathrm{a}}[k]$ をフレーズ制御機構とアクセント制御機構のインパルス応答の離散時間表現とすると，$\sum_k |G_{\mathrm{p}}[k] * \bar{u}_{\mathrm{p}}[k] - \bar{x}_{\mathrm{p}}[k]|^2$ と $\sum_k |G_{\mathrm{a}}[k] * \bar{u}_{\mathrm{a}}[k] - \bar{x}_{\mathrm{a}}[k]|^2$ と書ける．この非負制約つき逆畳み込み問題は，[3] の方法で非常に効率的に解くことができる．なお，こうして指令の大きさに非負制約を導入する方法は，[2] に新たに追加された 3 つ目の改良点である．

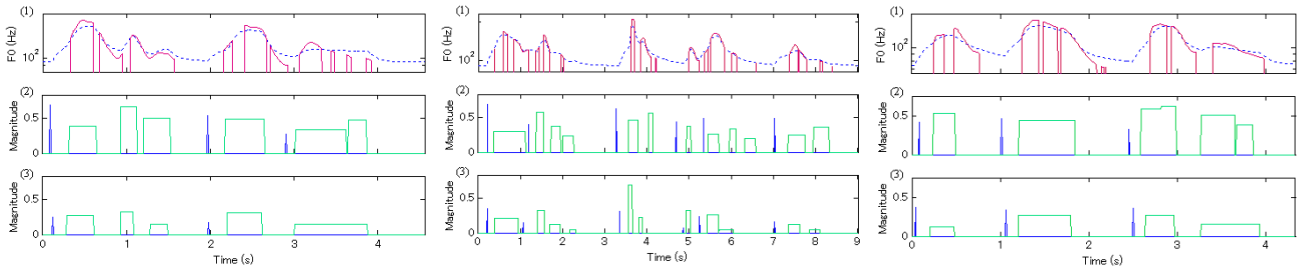
5 実験

統計的 F_0 パターン生成過程モデルが，藤崎モデルのパラメータを扱う上で十分な性能を持っていることを確かめるため，4 章で述べたアルゴリズムを使って，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問題を解き，その結果を定量評価する．実験には，音素のバランスがとれた 503 文を集めたデータベースである，ATR 日本語発話データベースの B セットを用いる [4]．話者は MHT(男性話者) を選択した．また，[?] の方法を用いて入力音声から F_0 パターンを推定した．加えて，推定パラメータの妥当性を評価する目的といくつかのパラメータを学習する目的で，韻律研究の専門家の手によって発話音声に付与された藤崎モデルのパラメータを用いた．なお，付与されたこのパラメータのベースライン成分は，常に log(60 Hz) である．

この実験において，定数値のパラメータは，$N = 20$, $t_0 = 8$ ms, $\alpha = 3.0$, $\beta = 20.0$, $\sigma_{\mathrm{p}}^2[k] = 0.2^2$, $\sigma_{\mathrm{a}}^2[k] = 0.02^2$, $\sigma_{\mathrm{b}}^2 = 0.001^2$, 無声区間では $\sigma_{\mathrm{n}}^2[k] = 10^{15}$，それ以外の区間では $\sigma_{\mathrm{n}}^2[k] = 10^{-15}$ に設定し，μ_{b} は有声区間の $\log F_0$ の最小値とした．Θ の初期値は [?] の方法で決定し，EM アルゴリズムの反復回数は 20 回とした．HMM の小状態数と状態遷移確率 $\phi_{i',i}$ は，ATR の No.1 から No.200 までの 200 文に専門家が手動で付与した藤崎モデルのパラメータから学習した．定量評価は，残りの 303 文を用いて行った．

手動で決定された指令列と，提案手法により推定された指令列を，3 つの例について比較した図が Fig. 3 である．より詳しく，推定性能を定量評価するため，次のようにした．まず，Fig. 4 のように，S 秒以下のずれしかないフレーズ指令同士，アクセント指令同士がマッチングされる可能性があるとみなして，手動で決定された指令列と推定された指令列との間のマッチング数が最大になるようにマッチングをとった．ここで，アクセント指令の時間のずれとは，アクセント指令の立ち上がり時刻のずれと立ち下がり時刻のずれの平均であるとした．N_{E} を推定された指令列の指令数，N_{A} を手動で決定された指令列の指令数，N_{M} をマッチング数，N_{Esum}, N_{Asum}, N_{Msum} をそれぞれ N_{E}, N_{A}, N_{M} を 303 文について足し合わせた値としたとき，挿入エラーを $E_{\mathrm{I}} = (N_{\mathrm{Esum}} - N_{\mathrm{Msum}})/N_{\mathrm{Asum}}$，脱落エラーを $E_{\mathrm{O}} = (N_{\mathrm{Asum}} - N_{\mathrm{Msum}})/N_{\mathrm{Asum}}$ と定義した．また，正解率を $A = 1 - E_{\mathrm{I}} - E_{\mathrm{O}}$ と定義し，これを定量評価の基準とした．なお，この実験においては，ベースライン成分の値のずれを考慮して，フレーズ・アクセント指令の大きさの違いは評価基準に含めなかった．

$S = 0.3$ 秒としたときの実験結果を Table 1 に記した．これらの表は，フレーズ・アクセント両指令，フレーズ指令のみ，アクセント指令のみに注目したときの，正解率 A，挿入エラー E_{I}，脱落エラー E_{O} を，



(a): ATR 発話文 No. 347　(b): ATR 発話文 No. 400　(c): ATR 発話文 No. 435

Fig. 3　パラメータ推定結果の例: (1) 実線は有声区間の観測 F_0 パターン，点線は推定された $\boldsymbol{y}_{\mathrm{p}} + \boldsymbol{y}_{\mathrm{a}} + \boldsymbol{y}_{\mathrm{b}}$ の値．(2) 手動で付与されたフレーズ・アクセント指令 (3) 提案手法で推定されたフレーズ・アクセント指令

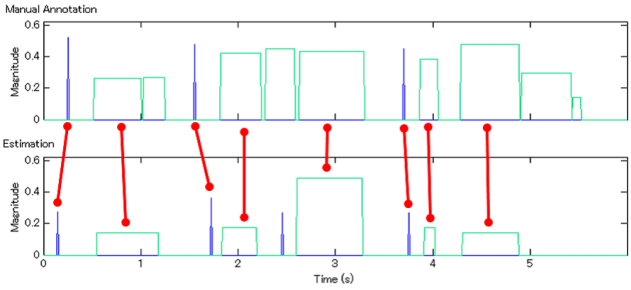


Fig. 4　指令列のマッチングの例

Table 1　定量評価の結果 (S=0.3 秒).

(a) 両指令

	A	E_{I}	E_{O}
初期状態	0.688	0.088	0.224
反復後	0.697	0.127	0.177

(b) フレーズ指令

	A	E_{I}	E_{O}
初期状態	0.647	0.109	0.244
反復後	0.680	0.207	0.112

(c) アクセント指令

	A	E_{I}	E_{O}
初期状態	0.711	0.076	0.213
反復後	0.708	0.083	0.207

それぞれ初期状態と EM アルゴリズムの反復後に計算した結果である．この表より，初期状態の正解率と反復後の正解率には，大きな差がないことが分かる．パラメータの初期値を定めるのに使った [**?**] の方法は，F_0 パターンに関する様々な知見を利用して観測 F_0 パターンから藤崎モデルのパラメータを推定するアルゴリズムであり，この実験結果は，提案した統計的手法によるパラメータ推定がそれと同程度の性能を持っていることを示している．ゆえに我々は，提案した統計的 F_0 パターン生成過程モデルは，藤崎モデルのパラメータを統計的に扱ううえで十分な性能を持っていると結論づけた．

6　おわりに

本研究では，統計的モデルをベースにした様々な音声アプリケーションに F_0 パターンを組み入れるために統計的 F_0 パターン生成過程モデルを提案し，そのモデルが藤崎モデルのパラメータを統計的に扱うことができることを実験によって確認した．今後は，統計的 F_0 パターン生成過程モデルを用いて大量の音声データからフレーズ・アクセント指令の傾向を学習することを通して，F_0 パターンを適切に制御した表情豊かな音声合成や，抑揚の特徴をふまえた話者認識など，実際に様々な音声アプリケーションに韻律モデルを組み込んでいきたい．

謝辞　本研究の一部は，文部科学省科学研究費補助金 (23240021) の助成を受けて行われた．また，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．

参考文献

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

[2] H. Kameoka *et al.*, “A statistical model of speech F0 contours,” *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition (SAPA)*, pp. 43-48, 2010.

[3] H. Kameoka *et al.*, “Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms,” *In Proc. 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2009)*, pp. 45-48, 2009.

[4] A. Kurematsu *et al.*, “ATR japanese speech database as a tool of speech recognition and synthesis,” *Speech Communication*, vol. 27, pp. 187-207, 1999.

[5] 亀岡, “全極型声道モデルと F0 パターン生成過程モデルを内部にもつ統一的音声生成モデル,” 日本音響学会 2010 年秋季研究発表会, pp. 211-214,2010.

[6] 成澤他, “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の評価,” 情報処理学会音声言語情報処理研究会, pp. 1-6, 2003.