音声 F_0 パターン生成過程の確率モデルによるテキストからの韻律生成とその評価*

☆門脇健人 [1]，亀岡弘和 [1,2]
([1] 東大院・情報理工, [2]NTT CS 研)

1 はじめに

本研究では，テキスト音声合成を目的としてテキストから F_0 パターンを生成する問題を扱う．音声基本周波数（F_0）パターンは，音声のイントネーションを表す特徴量であり，テキスト音声合成において高品質な F_0 パターンをいかに生成するかは重要課題の一つである．

テキスト音声合成において，隠れマルコフモデル (Hidden Markov Model; HMM）に基づく統計的アプローチ [1] が成功を収めている．HMM 音声合成 [1] では，各フレームの音韻的特徴量とともに F_0，及びそれらの 1 階差分，2 階差分を組にしたベクトルが特徴量として扱われ，学習データから HMM のパラメータを学習することで，学習した HMM を用いてテキストから音韻的特徴量系列と F_0 パターンを同時生成することが可能である．

音声合成において，自然なイントネーションをもつ合成音声を実現するためには，言語的に妥当でありつつ発声器官による音声の物理的な生成プロセスに即した F_0 パターンを適切に生成することが重要である．F_0 パターンの物理的な生成過程を模したモデルとして，藤崎らのモデル [2]（以後，藤崎モデル）が有名である．藤崎モデルは，生理学的・言語学的に意味のある少数のパラメータを用いて実測の F_0 パターンに非常によく近似できることが知られており，音声の F_0 パターンを表現するモデルとしては秀逸である．ただし，藤崎モデルはいわゆる trainable なモデルの形態をなしておらず，統計的アプローチとの親和性が必ずしも高いとは言えなかった．我々はこれまで，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ており，統計的手法に基づき観測 F_0 パターンから藤崎モデルのパラメータを推定するための基本アルゴリズムを導出するのに成功している [4, 5]．本手法の中心的なアイデアは，フレーズ・アクセント指令列の生成プロセスを HMM により表現した点にあり，音声コーパスを用いた統計学習を通して HMM の各状態をコンテキスト情報に対応付けることにより任意テキストから F_0 パターンを生成することが可能である [3]．以前の報告 [3] では，フレーズ指令とアクセント指令の開始・終了時刻はコンテキストラベルから決定し，学習においては固定としていた（各指令強度のみを学習していた）が，その後の検証により，コンテキストラベルから予測されるフレーズ・アクセント指令の生起時刻と観測 F_0 パターンのフレーズ・アクセント指令の生起時刻とが必ずしも一致せず，このことが原因で各指令の強度が適切に学習されていないことが明らかとなった．そこで本研究では，コンテキストラベルをもとに Left-to-Right 型の HMM を設計し，パラメータ学習において，各状態の出力分布の平均（各指令の強度に相当）だけでなく状態系列（各指令の生起時刻に相当）も未知パラメータとして観測 F_0 パターンから推定することを試みた．

2 音声 F_0 パターンの確率モデル

2.1 藤崎モデル

藤崎モデル [2] とは，甲状軟骨の二つの独立な運動（平行移動運動と回転運動）に伴う声帯の伸びの長さの和が声帯の固有振動数の対数（$\log F_0$）に比例する，という仮定をもとに，甲状軟骨の運動方程式を通して F_0 パターンの生成過程を表現したモデルである．甲状軟骨の平行移動運動に関係する F_0 パターンの成分をフレーズ成分 $y_{\mathrm{p}}(t)$，回転運動に関係する F_0 パターンの成分をアクセント成分 $y_{\mathrm{a}}(t)$ と呼び（t は時刻），対数 F_0 軌跡 $y(t)$（以後，F_0 パターン）はこれらの成分と声帯の物理的性質によって決まるベースライン成分と呼ぶ定数 y_{b} を加えたものとして表される．$y_{\mathrm{p}}(t)$ 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 $u_{\mathrm{a}}(t)$（ただしフレーズ指令とアクセント指令は同時に生起しない）を入力とした臨界制動の二次線形系により表現され，これらの値の関係は次のように書ける．

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

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

$$G_{\mathrm{p}}(t) = \alpha^2 t e^{-\alpha t}(t \geq 0), G_{\mathrm{a}}(t) = \beta^2 t e^{-\beta t}(t \geq 0). \tag{3}$$

ここで，$*$ は畳み込みを表す．また，α, β はそれぞれの制御機構の固有角周波数を表し，話者の個人差や言語によらずおおよそ $\alpha = 3, \beta = 20$[rad/s] 程度であることが経験的に知られている．日本語においては，藤崎モデルのフレーズ成分が F_0 パターン全体における緩やかな下降に相当し，フレーズ指令は主に息継ぎ，つまり呼気段落毎に生起する事がよく知られている．また，アクセント成分は主にアクセント句単位の急激な上がり下がりに対応しており，アクセント句毎のアクセント型によって決まるアクセント核においてアクセント指令が生起することが知られている．

2.2 藤崎モデルの確率モデル化

ここでは，今までに我々が開発してきた，藤崎モデルをベースにした F_0 パターンの生成過程の確率モデル [4, 5] の概説を行なう．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 $y_{\mathrm{p}}(t)$, $u_{\mathrm{p}}(t)$, $y_{\mathrm{a}}(t)$, $u_{\mathrm{a}}(t)$ の離散時間表現として，観測 F_0 パターンの対数値 $y[k]$ を次のように表現する．

$$y[k] \mid u_{\mathrm{p}}[k], u_{\mathrm{a}}[k] \sim \mathcal{N}(x[k], v_{\mathrm{n}}^2[k]), \tag{4}$$

$$x[k] = G_{\mathrm{p}}[k] * u_{\mathrm{p}}[k] + G_{\mathrm{a}}[k] * u_{\mathrm{a}}[k] + u_{\mathrm{b}}. \tag{5}$$

ここで $v_{\mathrm{n}}^2[k]$ は時刻 k における観測 F_0 パターンの“ 不確かさ ”を表すために導入した変数であり，これにより全時刻で正しい F_0 の値が観測できるとは限らないという問題をノイズとして統一的に扱うことを可能にした．

次に，$u_{\mathrm{p}}[k]$ と $u_{\mathrm{a}}[k]$ は，それぞれインパルス列状および矩形パルス列状の指令列関数であり，各パルスが同時刻には生起しない，という制約を満たす必要がある．[4, 5] では，両指令列関数のペア $\boldsymbol{o}[k] = (u_{\mathrm{p}}[k], u_{\mathrm{a}}[k])^{\mathsf{T}}$ を以下に示す HMM（以後，指令列生成 HMM）の出力系列と見なそうというアイディアにより，上述の制約を満たした指令列関数の確率モデルが提案されている．

---

* Text-to-speech prosody synthesis based on probabilistic model of F_0 contour and its evaluation by KADOWAKI Kento, KAMEOKA Hirokazu (The University of Tokyo/Nippon Telegraph and Telephone Corporation)

3　コンテキスト依存型 F_0 パターン生成過程モデル

3.1　コンテキスト依存フレーズ・アクセント指令列

2.1 節で述べたように，藤崎モデルにおけるフレーズ・アクセント指令列は言語情報と深く関連しており，任意のテキストが与えられた時に呼気段落の先頭にフレーズ指令を，対応するアクセント核の位置にアクセント指令を立てる事で自然な F_0 パターンが得られると仮定できる．この時，対応するフレーズ指令，アクセント指令の強度をコンテキストラベルからいかにして予測するかが問題となるが，各指令の強度をコンテキストに依存するパラメータとして扱うことで，コンテキストラベルを用いた統計学習により指令強度とコンテキスト情報を紐付けることができる．F_0 パターンは音声イントネーションの大域的特徴を表していることから，以上で述べたフレーズ指令やアクセント指令は前後の呼気段落及びアクセント句の環境に依存していると考える事ができる．以上のような考えから，フレーズ指令，アクセント指令のパラメータがそれぞれ前後の呼気段落及びアクセント句環境（以下，コンテキストと呼ぶ）に基づいて決定できるという可能性が示唆される．この様なコンテキストに基づいて，同じ文脈情報を持つ指令列を同じクラスタと仮定し，決定木に基づくコンテキストクラスタリングを行って各パラメータの強度とその決定木を学習するアルゴリズムを提案する．コンテキストには様々な要素が考えられるが，本研究で考慮した要素を以下に挙げる．

- フレーズ成分に関する言語情報
  - {先行，当該，後続} アクセント句モーラ数
  - {先行，当該，後続} アクセント型
  - 当該アクセント句の文における位置
  - 当該アクセント句の呼気段落における位置
  - {先行，当該，後続} 呼気段落モーラ数
  - 当該アクセント句のある呼気段落モーラ数
  - 当該アクセント句のある呼気段落の文における位置
  - 文のモーラ数
- アクセント成分に関する言語情報
  - {先行，当該，後続} 呼気段落モーラ数
  - {先行，当該，後続} 呼気段落内のアクセント句数
  - 当該呼気段落の文における位置
  - 文のモーラ数

ここで，フレーズ成分は呼気段落，アクセント成分はアクセント句と 1 対 1 に対応した状態である．

3.2　コンテキスト依存型指令列生成 HMM を内包する F_0 パターンの確率モデル

本節では 2.2 節で述べた確率モデルを，コンテキスト依存型のモデルとするため Fig. 1 に示すような HMM の状態遷移を考え，フレーズ指令，アクセント指令が言語情報毎にそれぞれ M 種類，N 種類に分類されるようなモデルを考える．この様なモデルは新たに以下の HMM で表現できる．

出力系列: $\boldsymbol{o}[k] = (u_{\mathrm{p}}[k], u_{\mathrm{a}}[k])^{\mathsf{T}}\ (k = 1, \ldots, K)$
状態集合: $\mathcal{S} = \{\mathrm{p}_0, \cdots, \mathrm{p}_M, \mathrm{a}_0, \cdots, \mathrm{a}_N\}$
状態系列: $\boldsymbol{s} = \{s_k \in \mathcal{S} | k = 1, \ldots, K\}$
出力分布: $P(\boldsymbol{o}[k]|s_k = i) = \mathcal{N}(\boldsymbol{c}_i[k], \boldsymbol{\Upsilon})$

$$\boldsymbol{c}_i[k] = \begin{cases} (0,0)^{\mathsf{T}} & (i \in \mathrm{p}_0, \mathrm{a}_0) \\ (A_{\mathrm{p}}^{(m)}, 0)^{\mathsf{T}} & (i \in \mathrm{p}_m) \\ (0, A_{\mathrm{a}}^{(n)})^{\mathsf{T}} & (i \in \mathrm{a}_n) \end{cases} \quad \boldsymbol{\Upsilon} = \begin{bmatrix} v_{\mathrm{p},i}^2 & 0 \\ 0 & v_{\mathrm{a},i}^2 \end{bmatrix}$$

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

以上の指令列生成 HMM において，フレーズ指令，アクセント指令の生起時刻がそれぞれ呼気段落とアクセント句に対応する様な制約を加えなければならないが，[3] では状態系列 $\boldsymbol{s}$（各指令の開始・終了時刻に相当）をコンテキストラベルから決定し，学習においては固定としていた．しかし，学習の際，コンテキストラベルから予測されるフレーズ・アクセント指令の生起時刻と観測 F_0 パターンのフレーズ・アクセント指令の生起時刻とが必ずしも一致せず，このことが原因で Fig. 3 と Fig. 4 の上段に示すように各指令の強度が適切に学習されないことが実験を通して明らかとなった．そこで，Fig.2 のような Left-to-Right 型の HMM をコンテキストラベルをもとに設計し，パラメータ学習において状態系列も未知パラメータとして観測 F_0 パターンから推定することを考える．

状態系列 $\boldsymbol{s}$ が決定されれば $\{(u_{\mathrm{p}}[k], u_{\mathrm{a}}[k])^{\mathsf{T}}\}_{k=1}^{K}$ が決定される．更に，状態系列 $\{s_k\}_{k=1}^{K}$ が決まればフレーズ・アクセント指令関数の平均系列 $\mu_{\mathrm{p}}[k], \mu_{\mathrm{a}}[k]\ (k = 1, \ldots, K)$ は，$(\mu_{\mathrm{p}}[k], \mu_{\mathrm{a}}[k])^{\mathsf{T}} = \boldsymbol{c}_{s_k}$ で与えられる．ここで，以下の文字をまとめて，

$$\boldsymbol{y} = \{y[k]\}_{k=1}^{K}, \boldsymbol{s} = \{s_k\}_{k=1}^{K},$$
$$\boldsymbol{o} = \{(u_{\mathrm{p}}[k], u_{\mathrm{a}}[k])^{\mathsf{T}}\}_{k=1}^{K},$$
$$\boldsymbol{\theta} = \{\{A_{\mathrm{p}}^{(m)}\}_{m=1}^{M}, \{A_{\mathrm{a}}^{(n)}\}_{n=1}^{N}\},$$

及び，

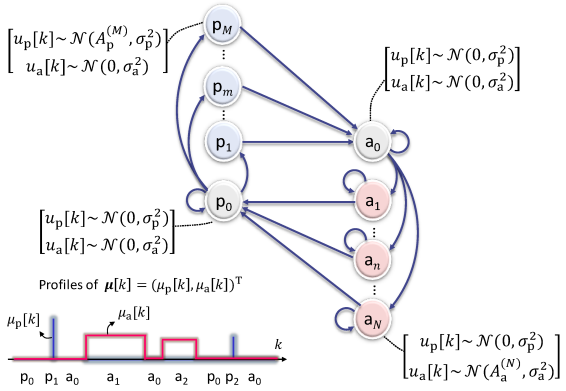

Fig. 1　提案する指令列生成 HMM の状態遷移モデル．従来の指令列生成 HMM[4, 5] とは異なり，フレーズ指令，アクセント指令パラメータが言語情報によってそれぞれフレーズ指令は M 種類，アクセント指令は N 種類に分類される HMM から出力されるモデルになっている．

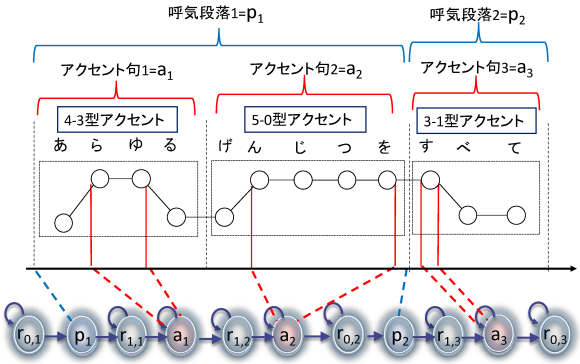

Fig. 2　コンテキストラベルをもとに設計される Left-to-Right 型の HMM.

$\boldsymbol{\mu}_{\rm p}$=$(\mu_{\rm p}[1],\ldots,\mu_{\rm p}[K])^{\mathsf{T}}$, $\boldsymbol{\mu}_{\rm a}$=$(\mu_{\rm a}[1],\ldots,\mu_{\rm a}[K])^{\mathsf{T}}$,
$\boldsymbol{y}$=$(y[1],\ldots,y[K])^{\mathsf{T}}$,

と表記する．簡単のため $\phi_{i',i}$, $\mu_{\rm b}$, $v^2_{{\rm p},i}$, $v^2_{{\rm a},i}$, $v^2_{\rm b}$, $v^2_{\rm n}[k]$, α, β は定数と仮定すると，指令列生成 HMM の状態系列 $\boldsymbol{s}$ と状態出力分布パラメータ $\boldsymbol{\theta}$ が与えられた下で F_0 パターン $\boldsymbol{y}$ が生成される確率（$\boldsymbol{s}$ および $\boldsymbol{\theta}$ の尤度関数）は，

$$P(\boldsymbol{y}|\theta,\boldsymbol{s}) = \frac{|\boldsymbol{\Sigma}^{-1}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\}, \tag{6}$$

$$\boldsymbol{\mu} = \boldsymbol{A}^{-1}\boldsymbol{\mu}_{\rm p} + \boldsymbol{B}^{-1}\boldsymbol{\mu}_{\rm a} + \mu_{\rm b}\boldsymbol{1},$$
$$\boldsymbol{\Sigma} = \boldsymbol{A}^{-1}\boldsymbol{\Sigma}_{\rm p}\big(\boldsymbol{A}^{\mathsf{T}}\big)^{-1} + \boldsymbol{B}^{-1}\boldsymbol{\Sigma}_{\rm a}\big(\boldsymbol{B}^{\mathsf{T}}\big)^{-1} + \boldsymbol{\Sigma}_{\rm b}.$$

によって与えられる．ただし，$\boldsymbol{A}$ と $\boldsymbol{B}$ は，

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

であり，a_2, a_1, a_0 及び b_2, b_1, b_0 は

$$a_2 = (\psi-1)^2, a_1 = -2\psi(\psi-1), a_0 = \psi^2,$$
$$b_2 = (\varphi-1)^2, b_1 = -2\varphi(\varphi-1), b_0 = \varphi^2,$$
$$\psi = 1 + \frac{1}{\alpha t_0}, \varphi = 1 + \frac{1}{\beta t_0},$$

である．なお，詳しい導出は [4] を参照されたい．

4 パラメータ学習と F_0 パターン生成

4.1 コンテキストクラスタリング

本章では，豊富なコンテキスト情報を用いて藤崎モデル指令列のパラメータ $\boldsymbol{\theta} = \{\{A_{\rm p}^{(m)}\}_{m=1}^{M}, \{A_{\rm a}^{(n)}\}_{n=1}^{N}\}$ を決定木によるコンテキストクラスタリング [9] に基づき学習するアルゴリズムを提案する．これによって，学習データのあらゆる指令列パラメータを用いて統計的にモデルを学習し，未知入力データに対してもコンテキストラベルによって指令列の強度を決定することが可能になる．本手法ではノード分割の規準に対して最小記述長（Minimum Description Length; MDL）規準を採用する．また，MDL 規準における尤度は藤崎モデルパラメータ $\boldsymbol{\theta}$ および状態系列 $\boldsymbol{s}$ が与えられた下での F_0 パターンの確率密度関数を採用する．この時，決定木の葉ノードは各指令列パラメータの自由度 M, N と一致しており，決定木が深くなるほど指令列パラメータの自由度が増える構造になっている．具体的な MDL 規準の式は，パラメータ $\boldsymbol{s}, \boldsymbol{\theta}$，学習データのインデックスを $d = 1,\ldots,D$，データ d における観測 F_0 パターン $\boldsymbol{y}^{(d)} = \{y^{(d)}[k]\}_{k=1}^{K^{(d)}}$ とすると学習データにおける対数尤度関数 $L(\boldsymbol{\theta})$ を用いて，

$$MDL = -L(\boldsymbol{\theta}) + c(N+M)\log W + C,$$
$$L(\boldsymbol{\theta}) = \sum_{d=1}^{D}\Big\{\frac{1}{2}\log|\boldsymbol{\Sigma}^{-1}| - \frac{K^{(d)}}{2}\log 2\pi - \frac{1}{2}(\boldsymbol{y}^{(d)}-\boldsymbol{\mu}^{(d)})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}^{(d)}-\boldsymbol{\mu}^{(d)})\Big\}, \tag{7}$$
$$\boldsymbol{\mu}^{(d)} = \boldsymbol{A}^{-1}\boldsymbol{\mu}_{\rm p}^{(d)} + \boldsymbol{B}^{-1}\boldsymbol{\mu}_{\rm a}^{(d)} + \mu_{\rm b}\boldsymbol{1},$$
$$\boldsymbol{\Sigma} = \boldsymbol{A}^{-1}\boldsymbol{\Sigma}_{\rm p}\big(\boldsymbol{A}^{\mathsf{T}}\big)^{-1} + \boldsymbol{B}^{-1}\boldsymbol{\Sigma}_{\rm a}\big(\boldsymbol{B}^{\mathsf{T}}\big)^{-1} + \boldsymbol{\Sigma}_{\rm b},$$

で与えられる．なお式 (7) におけるパラメータ c はモデルの大きさを調整する為の重みパラメータであり，小さいほど決定木が深くなるように調節できる．また，C はモデルを決める際に必要な符号長であり，ここでは常に定数である．ここで，ノードが増える度に，指令列パラメータ $\boldsymbol{\theta}$ を再推定する必要がある．各学習データに対して $\boldsymbol{\theta}$ を推定するアルゴリズムは [4] において提案されており，本手法ではそれをもとにして Fig. 2 の様な HMM を用いて推定を行った．

4.2 パラメータ学習アルゴリズム

本節では，コンテキストに依存する藤崎モデル指令列パラメータ $\boldsymbol{\theta}$ を反復計算し，決定木におけるモデルパラメータ $\boldsymbol{\theta}$ を学習するアルゴリズムについて説明する．これは，[4] で示されたように，学習データ d における観測 F_0 パターン $\boldsymbol{y}^{(d)}$ が与えられたとき $P(\boldsymbol{\theta}|\boldsymbol{y}^{(d)})$ をパラメータ $\boldsymbol{\theta}$ に関して最大化する問題として定式化できる．これにより学習データの F_0 パターンに最もフィットする様にモデルパラメータ $\boldsymbol{\theta}$ が再推定される．ここで $P(\boldsymbol{\theta}|\boldsymbol{y}^{(d)})$ を最大化する問題を解析的に解くのは難しいが，[4] で示されるように $\boldsymbol{x}^{(d)} = (\boldsymbol{y}_{\rm p}^{(d)\mathsf{T}}, \boldsymbol{y}_{\rm a}^{(d)\mathsf{T}}, \boldsymbol{y}_{\rm b}^{(d)\mathsf{T}})^{\mathsf{T}}$ を完全データとみなすことで EM アルゴリズムによる不完全データ問題に帰着し，局所最適解を得ることができる．この時，本モデルにおける Q 関数は，

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}') \stackrel{\rm c}{=} \frac{1}{2}\Big[\log|\boldsymbol{\Lambda}^{-1}| - {\rm tr}(\boldsymbol{\Lambda}^{-1}\mathbb{E}[\boldsymbol{x}^{(d)}\boldsymbol{x}^{(d)\mathsf{T}}|\boldsymbol{y}^{(d)};\boldsymbol{\theta}']) + 2\boldsymbol{m}^{(d)\mathsf{T}}\boldsymbol{\Lambda}^{-1}\mathbb{E}[\boldsymbol{x}^{(d)}|\boldsymbol{y}^{(d)};\boldsymbol{\theta}'] - \boldsymbol{m}^{(d)\mathsf{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{m}^{(d)}\Big] + \log P(\boldsymbol{\theta}), \tag{8}$$

と書ける．ただし，$\stackrel{\rm c}{=}$ は定数部分を除いて一致する事を意味する．また，

$$\boldsymbol{x}^{(d)} = \begin{bmatrix} \boldsymbol{y}_{\rm p}^{(d)} \\ \boldsymbol{y}_{\rm a}^{(d)} \\ \boldsymbol{y}_{\rm b}^{(d)} \end{bmatrix}, \quad \boldsymbol{m}^{(d)} = \begin{bmatrix} \boldsymbol{A}^{-1}\boldsymbol{\mu}_{\rm p}^{(d)} \\ \boldsymbol{B}^{-1}\boldsymbol{\mu}_{\rm a}^{(d)} \\ \mu_{\rm b}\boldsymbol{1} \end{bmatrix},$$
$$\boldsymbol{\Lambda}^{-1} = \begin{bmatrix} \boldsymbol{A}^{\mathsf{T}}\boldsymbol{\Sigma}_{\rm p}^{-1}\boldsymbol{A} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{B}^{\mathsf{T}}\boldsymbol{\Sigma}_{\rm a}^{-1}\boldsymbol{B} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{\Sigma}_{\rm b}^{-1} \end{bmatrix}.$$

である．

E ステップでは直前のステップで更新されたモデルパラメータを $\boldsymbol{\theta}'$ に代入し Q 関数を更新する．紙面の都合上詳細は省くが詳しくは [4] を参照されたい．M ステップでは，E ステップの Q 関数を最大化させるようにに各パラメータを更新する．具体的に更新すべきパラメータは，状態系列 $\boldsymbol{s}$，及びフレーズ指令の振幅平均 $A_{\rm p}^{(m)}$ とアクセント指令の振幅平均 $A_{\rm a}^{(n)}$ である．従って，M ステップは，Q 関数を状態系列に関して最大化するステップ，及びフレーズ指令の振幅平均 $A_{\rm p}^{(m)}$ とアクセント指令の振幅平均 $A_{\rm a}^{(n)}$ に関して最大化するステップとなる．状態系列に関して最大化するステップは，Viterbi アルゴリズムを用いて解くことができるが，詳細については [4] を参照されたい．

各指令の振幅平均に関して Q 関数を最大化する更新則はそれぞれ，

$$A_{\rm p}^{(m)} = \frac{1}{|\mathcal{T}_{{\rm p}_m}|}\sum_{k\in\mathcal{T}_{{\rm p}m}}[\boldsymbol{A}\bar{\boldsymbol{x}}_{\rm p}^{(d)}]_k, \quad \mathcal{T}_{{\rm p}_m} = \{k|s_k = {\rm p}_m\},$$

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

で与えられる．E ステップと M ステップの反復計算により，$P(\boldsymbol{\theta}|\boldsymbol{y}^{(d)})$ を局所最大化する $\boldsymbol{\theta}$ を得る事ができる．

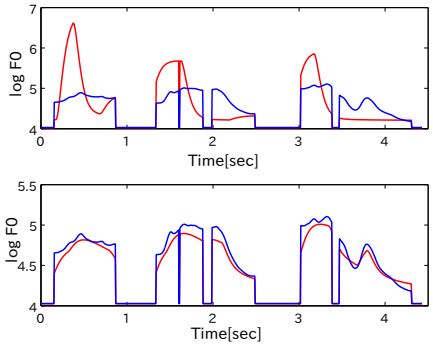

Fig. 3　従来法と提案法の推定結果の比較図．男性話者による『叔父さんは岬の一軒家に，一人ぼっちで住んでいた.』という発話の音声の F_0 パターン（青線）と従来法による推定 F_0 パターン（上段赤線），提案法による推定 F_0 パターン（下段赤線）．

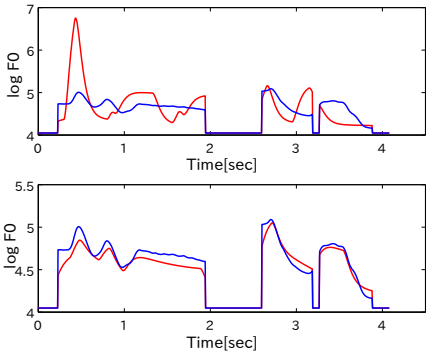

Fig. 4　従来法と提案法の推定結果の比較図．男性話者による『着用中に，ダウンやフェザーが飛び出す原因ともなります.』という発話の音声の F_0 パターン（青線）と従来法による推定 F_0 パターン（上段赤線），提案法による推定 F_0 パターン（下段赤線）．

4.3　テキストからの F_0 パターン生成

ここでは，入力テキストが与えられた時に対応する F_0 パターンを生成する手順について説明する．まず入力テキストが与えられた時に，コンテキストラベルを保持した呼気段落およびアクセント句を抽出する．次に，それぞれの呼気段落及びアクセント句に対して，コンテキストラベルを基に学習された決定木をたどっていき，対応する葉ノードの指令列パラメータを呼気段落の先頭，及び各アクセント句のアクセント核に立て，$\bar{\boldsymbol{o}} = \{(\bar{u}_{\rm p}[k], \bar{u}_{\rm a}[k])^{\mathsf{T}}\}_{k=1}^{K}$ を求める．後は式 (4),(5) に従って F_0 パターンを生成すればよい．

5　実験

今回の実験では，Fig. 2 のような HMM を用いてパラメータ学習を行い，状態系列を固定して学習を行っていた従来法 [3] と比較し，学習に用いた F_0 パターンに対してより高い精度でパラメータ学習が可能であることを確認する為の実験を行った．本実験における確率モデルの定数パラメータは以下のようにセットした．$t_0 = 5$ ms，$\alpha = 3.0$ rad/s，$\beta = 20.0$ rad/s，$\upsilon_{\rm p}^2[k] = 3^2$，$\upsilon_{\rm a}^2[k] = 0.03^2$，$\upsilon_{\rm b}^2 = 10^{-8}$，有声区間において $\upsilon_{\rm n}^2[k] = 10^{15}$，無声区間において $\upsilon_{\rm n}^2[k] = 0.1^2$．$\mu_{\rm b}$ は全 log F_0 の有声区間の値の最低値にセットし，EM アルゴリズムの反復回数は 100 回とした．また，各指令列の状態を小状態に分割し，持続長の確率分布をガウス分布として設定した．ガウス分布の平均値はコンテキストラベルから得られる値とし，分散は 2500 とした．学習用データには HTS2.1 のデモスクリプト [7] に同梱された男性話者の音声のうち，最初の 50 文を用いた．本手法によって推定された F_0 パターン及び [3] によって得られた F_0 パターンをそれぞれ学習用の F_0 パターンと比較した例を Fig. 3, Fig. 4 に示す．

Fig. 3, Fig. 4 に示されるように，状態系列を固定した場合 [3] に比べて，より観測 F_0 パターンにフィットした推定結果が得られるようになったことを確認した．

6　おわりに

以前の報告 [3] で提案したテキスト韻律合成手法では，生起時刻をコンテキストラベルから決定し，固定値として扱っていたが，コンテキストラベルから予測されるフレーズ・アクセント指令の生起時刻と観測 F_0 パターンにおけるフレーズ・アクセント指令の生起時刻とは必ずしも一致せず，このことが原因で各指令の強度を適切に学習することができていなかった．そこで本研究では，この問題を解決するため，コンテキストラベルをもとに Left-to-Right 型の HMM を設計し，パラメータ学習において，各状態の出力分布の平均（各指令の強度に相当）だけでなく状態系列（各指令の生起時刻に相当）も未知パラメータとして観測 F_0 パターンから推定する方法を検討した．実験結果に示されるように，[3] の方法に比べて，本手法によって学習 F_0 パターンに対してよりフィットするような指令列パラメータを学習可能であることを確認した．今後は，本手法によって学習されたパラメータを用いて音声合成を行い，その評価を行う予定である．

謝辞　本研究は JSPS 科研費 26280060 の助成を受けたものです．

参考文献

[1] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in Proc. ICASSP, vol. 3, pp. 1315–1318, 2000.

[2] H. Fujisaki, “In Vocal Physiology: Voice Production, Mechanisms and Functions,” Raven Press, 1988.

[3] 門脇, 石原, 北条, 亀岡, “音声 F0 パターン生成過程の確率モデルに基づくテキストからの韻律生成,” 日本音響学会春季研究発表会講演集, 3-6-17, pp. 361—364, Mar. 2014.

[4] H. Kameoka, J. Le Roux, and Y. Ohishi, “A statistical model of speech F_0 contours,” in Proc. SAPA, pp. 43–48, 2010.

[5] K. Yoshizato, H. Kameoka, D. Saito, and S. Sagayama, “Hidden Markov convolutive mixture model for pitch contour analysis of speech,” in Proc. Interspeech, 2012.

[6] T. Mausko, *et al*, “Multi-Space Probability Distribution HMM,” IEIC Technical Report, vol. 101, no. 323, pp. 41–42, 2001.

[7] “HMM-based Speech Synthesis System (HTS),” `http://hts.sp.nitech.ac.jp/`

[8] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, vol. 27, no. 3, pp. 187–207, 1999.

[9] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura,“ Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” Proc. of Eurospeech, pp. 2347–2350, 1999.