複合ウェーブレットモデルと F_0 パターン生成過程の確率モデルを用いたテキスト音声合成*

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

1　はじめに

本稿ではこれまで我々が提案してきたスペクトルパラメータ生成法と F_0 パターン生成法を併用したパラメトリックテキスト音声合成方式を提案する．

HMM(Hidden Markov Model) 音声合成方式 [1] では通常，メルケプストラム係数とそのデルタ量を特徴量とし，それらの系列を生成する HMM を仮定する。状態出力分布に正規分布または GMM(Gaussian Mixture Model) を仮定した場合，学習において状態出力分布の平均は同一状態に割り当てられたすべての時刻の観測特徴量の平均となる．メルケプストラムの平均をとる操作は対数スペクトルの平均をとる操作と等価であるため，各々が山と谷がはっきりしたスペクトルであってもそれらの平均をとると山と谷が平滑化されたスペクトルになってしまう．平滑化されたスペクトル包絡による合成音声は buzzy でこもった音になることが知られており，この問題の解決はパラメトリック音声合成における重要課題の一つである．[2] では，この問題に対し，従来のメルケプストラム系列を生成する GMM-HMM 系から脱却し，スペクトル系列を生成する複合ウェーブレットモデル (Composite Wavelet Model; CWM)[6, 7] と HMM の統合系を提案した．CWM はスペクトルのピークとパワーの両方に相当するパラメータを有していることが特色で，CWM-HMM 系の学習では各状態出力分布の平均はスペクトルのピークの周波数およびパワー方向に平均をとったものとなるため，従来のメルケプストラム GMM-HMM 系に比べ合成音声におけるスペクトル包絡の過剰平滑化が起きにくいのが特長である．

また，音声において音韻とともに韻律が果たす役割は重要で，HMM 音声合成方式では基本周波数 (F_0) パターンをいかに言語的に自然でかつ肉声らしいものにできるかも重要課題の一つである．従来の方式では，学習データの量によっては過学習を起こし不自然な F_0 パターンを生成することがあったが，言語的および物理的に不自然な F_0 パターンを生成することをできるだけ防ぎたいという動機から [3] では F_0 パターンの物理的生成過程を模した藤崎モデル [4] の確率モデル版 [8] を用いた韻律生成法を提案した．この方式では言語的かつ物理的に意味のあるパラメータが生成されるため，合成音声のイントネーションの話者性や発話スタイルを柔軟に加工できるという利点もある．

本稿では，当研究室で開発した上述の二つのモデルを併用したパラメトリック音声合成方式を提案し，それぞれの効果を検証する．

2　CWM-HMM 統合系

本章では，[2] で提案したスペクトル包絡系列の生成モデルについて概説する．

スペクトル包絡ピークの周波数とパワーをパラメータにもつ CWM により, その確率モデル化を行うことが可能である．CWM は,GMM によりスペクトル包絡を近似的に表現するモデルである.CWM によるスペクトル包絡 $f_{\omega,l}$ は,Fig. 1 のように,

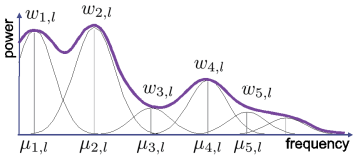

Fig. 1　CWM によるスペクトル表現

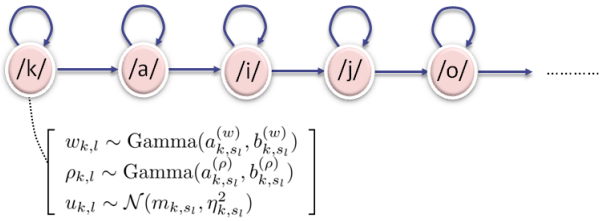

Fig. 2　CWM パラメータを出力する HMM の構成

$$f_{\omega,l} = \sum_{k=1}^{K} \frac{w_{k,l}}{\sqrt{2\pi}\sigma_{k,l}} \exp\left(-\frac{(\omega - \mu_{k,l})^2}{2\sigma^2_{k,l}} \right) \qquad (1)$$

と表される．ただし，ω, l は周波数と時刻のインデックス，k は Gauss 関数のインデックスであり，K は GMM の混合数である．また，$\mu_{k,l}, \sigma^2_{k,l}, w_{k,l}$ は，それぞれ Gauss 分布関数を統計分布と見なした際の平均・分散・重みに対応し，各スペクトル包絡ピークの周波数・鋭さ・強度に対応するパラメータである．

以上の準備のもと，観測スペクトル系列が生成される過程を確率モデル化する．図 2 のような，各離散時刻 l ごとに平均 $\mu_{k,l}$，分散の逆数 $\rho_{k,l} := 1/\sigma^2_{k,l}$，重み $w_{k,l}$ の CWM パラメータを出力する HMM を考える．HMM の各状態は，言語ラベルの一状態を表しており，例えば Fig. 2 のようにそれぞれ一つの音素に対応させることもできるが，HTS[1] などの手法と同様に，音素状態に加え，前後の音素のやアクセント位置などの情報を用いたコンテキストラベルの一状態を対応させることも可能である．時刻 l における状態番号を s_l とし，本稿では状態出力分布（各状態における CWM パラメータの生成確率）を，後述のパラメータ推定アルゴリズムの導出の便宜上の都合により，

$$\mathrm{P}(\mu_{k,l}|s_l) = \mathcal{N}(\mu_{k,l}; m_{k,s_l}, \eta^2_{k,s_l}) \qquad (2)$$

$$\mathrm{P}(\rho_{k,l}|s_l) = \mathrm{Gamma}(\rho_{k,l}; a^{(\rho)}_{k,s_l}, b^{(\rho)}_{k,s_l}) \qquad (3)$$

$$\mathrm{P}(w_{k,l}|s_l) = \mathrm{Gamma}(w_{k,l}; a^{(w)}_{k,s_l}, b^{(w)}_{k,s_l}) \qquad (4)$$

と仮定した．ここで，$\mathcal{N}(x; m, \eta^2)$ は正規分布，$\mathrm{Gamma}(x; a, b)$ はガンマ分布

$$\mathrm{Gamma}(x; a, b) = x^{a-1} \frac{\exp(-x/b)}{\Gamma(a)\ b^a} \qquad (5)$$

---

* Text-to-speech synthesis combining CWM-HMM and probabilistic Fujisaki model. by KADOWAKI Kento (The Univ. of Tokyo), Hojo Nobukatsu (NTT), KAMEOKA Hirokazu (The Univ. of Tokyo/NTT)

である．また，上述の HMM により生成された CWM パラメータの系列 $\boldsymbol{\mu} = \{\mu_k\}_{k,l}, \boldsymbol{\rho} = \{\rho_k\}_{k,l}, \boldsymbol{w} = \{w_k\}_{k,l}$ が与えられたとき，時刻 l において，観測スペクトル $y_{w,l}$ が生成される確率分布についても後述のパラメータ推定アルゴリズムの導出の便宜上の都合により，

$$\mathrm{P}(y_{w,l}|\boldsymbol{\mu},\boldsymbol{\rho},\boldsymbol{w}) = \mathrm{Poisson}(y_{w,l}|f_{w,l}) \tag{6}$$

と定めた．ここで，$f_{w,l}$ は，cwm パラメータ系列 $\boldsymbol{\mu},\boldsymbol{\rho},\boldsymbol{w}$ が与えられたとき，時刻 l の cwm パラメータを用いて (1) で表されるスペクトル包絡モデルであり，$\mathrm{Poisson}(x;\lambda)$ はポアソン分布

$$\mathrm{Poisson}(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}. \tag{7}$$

である．上記の生成モデルを定めることにより，次節のパラメータ推定アルゴリズムを導出する．

スペクトル包絡系列生成モデルパラメータの推定は，スペクトル包絡生成モデルのいわば逆問題である．これは，観測スペクトル系列 $Y = \{y_{w,l}\}_{w,l}$ が与えられたときに，スペクトル系列生成モデルパラメータ Θ の事後確率 $\mathrm{P}(\Theta|Y) \propto P(Y|\Theta)P(\Theta)$ を最大化する問題として定式化される．推定すべきパラメータ Θ は，HMM の出力状態列 ($\boldsymbol{s} = \{s_l\}_l$)，HMM の各状態 i の出力分布パラメータ ($\boldsymbol{\theta} = \{m_{k,i}, \eta_{k,i}, a^{(\sigma)}_{k,i}, b^{(\sigma)}_{k,i}, a^{(w)}_{k,i}, b^{(w)}_{k,i}\}_{k,i}$) と，CWM パラメータ系列 $(\boldsymbol{\mu},\boldsymbol{\rho},\boldsymbol{w})$ である．

Θ の事後確率 $\mathrm{P}(\Theta|Y)$ を最大化する Θ を求めることは難しいが，各変数について局所最適化を繰り返すことは可能である．このとき $\mathrm{P}(\Theta|Y)$ は，

$$\log P(\Theta|\boldsymbol{Y}) \overset{c}{=} \log P(\boldsymbol{Y}|\Theta) + \alpha \log P(\Theta), \tag{8}$$

$$\log P(\Theta) \overset{c}{=} \log P(\boldsymbol{s}) + \log P(\boldsymbol{\rho}|\boldsymbol{s},\boldsymbol{\theta}) + \log P(\boldsymbol{w}|\boldsymbol{s},\boldsymbol{\theta}) + \log P(\boldsymbol{\mu}|\boldsymbol{s},\boldsymbol{u},\boldsymbol{\theta}) \tag{9}$$

と書ける．ここで，また，$\overset{c}{=}$ は定数部分を除いて一致することを意味する．$P(\Theta|\boldsymbol{y})$ を最大化（または $-P(\Theta|\boldsymbol{y})$ を最小化）する Θ を解析的に求めることは難しいが，補助関数法に基づいて局所最適化アルゴリズムを導くことができる．紙面の都合上詳細は省略するが，詳しくは [2] を参照されたい．

3 F_0 パターン生成過程の確率モデル

本章では，F_0 パターン生成過程の確率モデルに基づく韻律合成手法 [3] の概要を述べる．

藤崎モデル [4] 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．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 [8, 9] は以下で定式化される．k を離散時刻のインデックスとし，$y_\mathrm{p}[k]$, $u_\mathrm{p}[k]$, $y_\mathrm{a}[k]$, $u_\mathrm{a}[k]$ をそれぞれフレーズ成分 $y_\mathrm{p}(t)$, フレーズ指令 $u_\mathrm{p}(t)$, アクセント成分 $y_\mathrm{a}(t)$, アクセント指令 $u_\mathrm{a}(t)$ 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．

$$y[k] \mid u_\mathrm{p}[k], u_\mathrm{a}[k] \sim \mathcal{N}(x[k], \upsilon^2_\mathrm{n}[k]), \tag{10}$$

$$x[k] = G_\mathrm{p}[k] * u_\mathrm{p}[k] + G_\mathrm{a}[k] * u_\mathrm{a}[k] + u_\mathrm{b}. \tag{11}$$

ここで $\upsilon^2_\mathrm{n}[k]$ 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．[8, 9] 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．

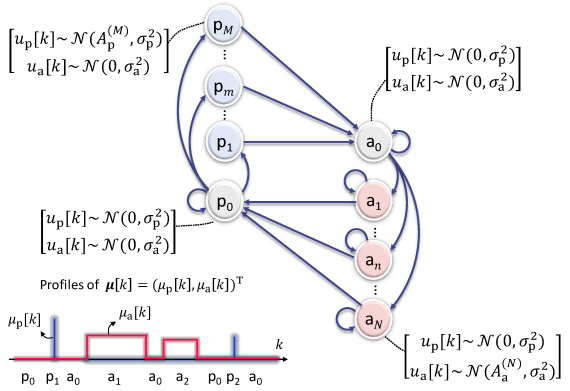

Fig. 3　指令列生成 HMM の状態遷移モデル [3].

[3] でも述べたように，藤崎モデルの指令列は言語情報との関連が深いので，フレーズ指令と呼気段落，アクセント指令とアクセント句が 1 対 1 に対応した状態へと拡張することによって，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た．

上述の確率モデルを，コンテキスト依存型のモデルとするため Fig. 3 に示すような HMM の状態遷移を考える．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: $P(\boldsymbol{o}[k]|s_k = i) = \mathcal{N}(\boldsymbol{c}_i[k], \boldsymbol{\Upsilon})$

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

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るために，Left-to-Right 型の HMM をコンテキストラベルをもとに設計することによって，制約を考慮しつつ，パラメータ学習において指令列の強度と状態系列を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 [15]．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，

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

と表記する．簡単のため $\phi_{i',i}$, μ_b, $\upsilon^2_{\mathrm{p},i}$, $\upsilon^2_{\mathrm{a},i}$, υ^2_b, $\upsilon^2_\mathrm{n}[k]$, 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，

$$P(\boldsymbol{y}|\theta,\boldsymbol{s}) = \frac{|\boldsymbol{\Sigma}^{-1}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\}, \tag{12}$$

$$\boldsymbol{\mu} = \boldsymbol{A}^{-1}\boldsymbol{\mu}_{\mathrm{p}} + \boldsymbol{B}^{-1}\boldsymbol{\mu}_{\mathrm{a}} + \mu_{\mathrm{b}}\boldsymbol{1},$$
$$\boldsymbol{\Sigma} = \boldsymbol{A}^{-1}\boldsymbol{\Sigma}_{\mathrm{p}}\left(\boldsymbol{A}^{\top}\right)^{-1} + \boldsymbol{B}^{-1}\boldsymbol{\Sigma}_{\mathrm{a}}\left(\boldsymbol{B}^{\top}\right)^{-1} + \boldsymbol{\Sigma}_{\mathrm{b}}.$$

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

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

であり，a_2, a_1, a_0 及び b_2, b_1, b_0 は

$$a_2 = (\psi-1)^2, a_1 = -2\psi(\psi-1), a_0 = \psi^2, \tag{13}$$

$$b_2 = (\varphi-1)^2, b_1 = -2\varphi(\varphi-1), b_0 = \varphi^2, \tag{14}$$

$$\psi = 1 + 1/\alpha t_0, \varphi = 1 + 1/\beta t_0,$$

である．詳しい導出は [8] を参照されたい．

続いて，指令列生成 HMM を用いて指令列パラメータを学習・生成する手法について述べる．[15] で報告したように，F_0 パターンを生成する場合にラベルからの指令列の時間ずれを考慮する必要があることが示唆されており，本節では指令列の強度およびラベルからの時間ずれをパラメータとして学習・生成する手法について検討する．

学習 F_0 パターンに対して前節で導出した Left-to-Right 型の指令列生成 HMM を用いて指令列の強度と時間ずれを推定し，推定された強度と時間ずれの両方に対してコンテキストクラスタリングを用いて決定木を学習するという多段的な手法を採用する．学習 F_0 パターンに対して指令列を推定するアルゴリズムに関しては，[9] と同様の方針で導出できるので，詳細は省略する．ここで，コンテキストクラスタリングにおけるノード分割基準を以下に示す．

$$\begin{aligned}\mathrm{MDL_p} = \frac{1}{2}\sum_{d=1}^{D_j}\Big\{ &\log(2\pi\sigma^2_{\mathrm{p},x,j}) + \frac{(x_{\mathrm{p},i}-\mu_{\mathrm{p},x,j})}{\sigma^2_{\mathrm{p},x,j}} \\ &+ \log(2\pi\sigma^2_{\mathrm{p},\tau,j}) + \frac{(\tau_{\mathrm{p},i}-\mu_{\mathrm{p},\tau,j})}{\sigma^2_{\mathrm{p},\tau,j}}\Big\} \\ &+ cJ\log W,\end{aligned} \tag{15}$$

$$\begin{aligned}\mathrm{MDL_a} = \frac{1}{2}\sum_{d=1}^{D_j}\Big\{ &\log(2\pi\sigma^2_{\mathrm{a},x,j}) + \frac{(x_{\mathrm{a},i}-\mu_{\mathrm{a},x,j})}{\sigma^2_{\mathrm{a},x,j}} \\ &+ \log(2\pi\sigma^2_{\mathrm{a},\tau,j}) + \frac{(\tau_{\mathrm{a},i}-\mu_{\mathrm{a},\tau,j})}{\sigma^2_{\mathrm{a},\tau,j}} \\ &+ \log(2\pi\sigma^2_{\mathrm{a},\Delta\tau,j}) + \frac{(\Delta\tau_{\mathrm{a},i}-\mu_{\mathrm{a},\Delta\tau,j})}{\sigma^2_{\mathrm{a},\Delta\tau,j}}\Big\} \\ &+ cJ\log W.\end{aligned} \tag{16}$$

ここで，$x_{\mathrm{p},i}$ は各葉ノードにおけるフレーズ指令の強度，$\tau_{\mathrm{p},i}$ は各葉ノードにおけるフレーズ指令生起時刻のラベルからの時間ずれ，$x_{\mathrm{a},i}$ は各葉ノードにおけるアクセント指令の強度，$\tau_{\mathrm{a},i}$ は各葉ノードにおけるアクセント指令開始時刻のラベルからの時間ずれ，$\Delta\tau_{\mathrm{a},i}$ は各葉ノードにおけるアクセント指令のデュレーションのラベルからの時間ずれ，J は葉ノード数，j は選択されたノードインデックス，D_j はノード j における占有状態数である．さらに，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，対応するコンテキストラベルと学習された決定木からパラメータ $\boldsymbol{\theta}, \boldsymbol{\tau}$ を得ることができる．ただし，$\boldsymbol{\tau}$ はフレーズなら生起時刻の時間ずれのみを表し，アクセントならばアクセント開始時刻とデュレーションの両方の時間ずれを表す．$\boldsymbol{\theta}$ は各指令に対応する強度パラメータである．以下の生成確率が最大となるように F_0 パターン $\boldsymbol{y}$ を生成すればよい．

$$\boldsymbol{y} = \underset{\boldsymbol{y}}{\mathrm{argmax}}\, P(\boldsymbol{y}|\boldsymbol{s},\boldsymbol{\theta}). \tag{17}$$

ここで，$\boldsymbol{s}$ は，コンテキストラベルと学習した時間ずれのパラメータ $\boldsymbol{\tau}$ を用いて決定できる．

4 テキストからの音声合成手法

入力テキストから得られる音素状態系列に対し，尤度最大化規範により CWM 特徴量系列を出力し，スペクトル包絡系列を得る．ただし,[1] と同様に音素状態を細かく分割し，さらに動的特徴量 (特徴量の 1 階,2 階の時間差分量) のモデルを用いた．同様に韻律情報に対しても，入力テキストから対応する指令列パラメータを生成することによって，合成 F_0 パターンを得ることができる．ただし，音素状態の duration に関するモデルが別途必要である．従って今回は duration に関して HTS2.1[11] の手法を採用した．

5 合成音声の評価実験

本節では，提案した F_0 パターン生成手法が妥当であることを示すための客観評価実験と提案した音声合成システムが自然な音声を実現できるかを示すために行った主観評価実験について述べる．

5.1 平均二乗誤差と相関係数による客観評価

合成 F_0 パターンの客観的な良さを測るために，実音声から観測された F_0 パターンと合成 F_0 パターンとの二乗平均平方根誤差 (RMSE) や相関係数 (Correlation) がよく使われており，以下のように表される．

$$\mathrm{RMSE} = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(y_i - x_i)^2}, \tag{18}$$

$$\mathrm{Corr} = \frac{T\sum_{i=1}^{T} y_i x_i - \sum_{i=1}^{T} y_i \sum_{i=1}^{T} x_i}{\sqrt{T\sum_{i=1}^{T}(y_i)^2 - y_i^2}\sqrt{T\sum_{i=1}^{T}(x_i)^2 - x_i^2}}. \tag{19}$$

ただし Corr は相関係数，y_i, x_i はそれぞれ時刻 t_i における観測 F_0[Hz]，生成 F_0[Hz] の対数値であり，T はフレーム数である．以上の基準を用いて提案法と従来法 [1] の客観評価を行った．両者の手法とも，HTS2.1[11] に同梱された男性話者の音声のうち，450 文を学習データとして用い，残りの 53 文を評価の為に生成した．提案法の初期推定のパラメータは，$t_0 = 5$ ms, $\alpha = 3.0$ rad/s, $\beta = 20.0$ rad/s, $\upsilon^2_{\mathrm{p}}[k] = 3^2$, $\upsilon^2_{\mathrm{a}}[k] = 0.3^2$, υ^2_{b} は有声区間の最小値とし，有声区間において $\upsilon^2_{\mathrm{n}}[k] = 10^2$，無声区間において $\upsilon^2_{\mathrm{n}}[k] = 0.1^2$，$\phi_{i',i}$ はラベルから決定される duration が平均値となるようなガウス分布から得られる値とし，コンテキストクラスタリングは全てのノードが分割されるまで行った．ただし，どちらの F_0 パターンも実音声と同じ duration のラベルを用いて生成した．従来手法による F_0 パターンは有声・無声区間の判定が実音声のものと異なる区間が存在するので，どちらかが無声となる区間では評価しなかった．以上の結果を下の表に示す．

Table.1 からもわかるとおり，提案法のほうがより望ましい結果を得られていると確認できた．

Table 1　従来法と提案法によって生成された F_0 パターンと観測 F_0 パターンとの RMSE と相関係数を比較した表．括弧内はその標準偏差を表している．RMSE の低い方，相関係数の高い方を太字で表示した．

手法	RMSE	Correlation
提案手法	**0.100(± 0.023)**	**0.720(± 0.458)**$\times 10^4$
従来手法 [1]	0.122(± 0.033)	0.288(± 0.188)$\times 10^4$

5.2　主観評価

本節では，提案した音声合成システムが高品質な合成音声を生成できるかどうかを確かめるために行った主観評価実験について述べる．より細かく性能を評価するため，以下の四手法を評価対象とした．

Table 2　主観評価の対象とする四手法．ただし，hts は従来手法 [1]，cwm は cwm-hmm 統合系モデルの略で 2 節で述べた手法，pfm は probabilistic fujisaki model の略で 3 節で述べた手法である．

手法名	スペクトル包絡系列	F_0 パターン
A	hts	hts
B	cwm	hts
C	hts	pfm
D	cwm	pfm

以上の四手法に対してプリファレンステスト (AB テスト) を実施して，合成音声の品質を評価した．ここで，品質を評価する項目に関しては，JEITA 規格 [14] における評価項目から選んだ四つの項目，具体的には (a) 音声の明瞭性：音質の良さ，(b) 音声の時間要因：日本語らしいリズム感，(c) 音声の抑揚：抑揚・アクセントの自然性，流暢さ，(d) 音声の総合的な良さ：全体的な良し悪し，を用いた．この四手法に関して，受聴者は 4 手法全ての組み合わせの音声を聴き，各項目について良いと思った方を選択させる．受聴者は男性 5 名とする．実験は音声は HTS2.1 のデモスクリプトの j セット中の 5 文に対して行った．各項目に関するプリファレンススコアを以下の表に示す．表の各行が他の各手法と比較した場合のプリファレンススコアを表している．また，二項検定により 5% 水準で統計的に有意にスコアが高いと評価されたスコアは太字で示してある．

Table.3 から分かるとおり，提案音声合成手法は項目 (a)〜(d) に関して，他の全ての手法と比較して有意に高いスコアを示した．また，手法 B や C は手法 A と比較して統計的に有意に高いとは言えなかったが，手法 D では他の全ての手法と比べて有意に高いスコアを示していることから，音韻情報と韻律情報の相乗的な効果により音声の自然性が飛躍的に良くなるという可能性が考えられる．

6　まとめ

本稿では，より高品質なテキスト音声合成システムを実現するため，CWM-HMM 統合モデルと F_0 パターン生成過程の確率モデルによる音声合成システムを提案し，その概要について述べた．性能評価においては，F_0 パターン生成手法について客観評価基準を用いて評価し，その妥当性を示した．さらに，主観評価実験によって，提案システムの効果を確認した．今後は，更に大規模な主観評価実験を行う予定である．

謝辞　本研究は JSPS 科研費 26280060，26730100 の助成を受けたものです．

Table 3　項目 (a)〜(d) に関する各手法におけるプリファレンススコア．太字で示されているスコアは二項検定により 5% 水準で統計的に有意に高いと評価されていることを表す．

(a) 音質の良さ

	A	B	C	D
A	-	0.500	0.400	0.033
B	0.500	-	0.400	0.167
C	0.600	0.600	-	0.200
D	**0.967**	**0.833**	**0.800**	-

(b) 日本語らしいリズム感

	A	B	C	D
A	-	0.467	0.433	0.300
B	0.533	-	0.400	0.233
C	0.567	0.600	-	0.300
D	**0.700**	**0.767**	**0.700**	-

(c) アクセントの自然性

	A	B	C	D
A	-	0.600	0.367	0.400
B	0.400	-	0.333	0.200
C	0.633	0.667	-	0.433
D	0.600	**0.800**	0.567	-

(d) 全体的な良し悪し

	A	B	C	D
A	-	0.500	0.400	0.033
B	0.500	-	0.400	0.167
C	0.600	0.600	-	0.200
D	**0.967**	**0.833**	**0.800**	-

参考文献

[1] K. Tokuda *et al.*,.in Proc. ICASSP, vol. 3, pp. 1315–1318, 2000.

[2] N. Hojo *et al.*, In *8th ISCA Workshop on Speech Synthesis*, pp.129–34,2013.

[3] K. Kadowaki *et al*, in Proc. *The 15th Annual Conference of the International Speech Communication Association (Interspeech 2014)*, pp. 2322-2326, Sep. 2014.

[4] H. Fujisaki, Raven Press, 1988.

[5] 門脇他, 日本音響学会春季研究発表会講演集, 3-6-17, pp. 361-364, Mar. 2014.

[6] 槻他, 信学技報, Vol. 105, No. 372, pp. 1–6, 2005.

[7] H. Kameoka *et al.*, IEEE Transactions on Audio, Speech and Language Processing, Vol. 18, No. 6, pp. 1507–1516, Aug. 2010.

[8] H. Kameoka *et al.*, in Proc. SAPA, pp. 43–48, 2010.

[9] K. Yoshizato*et al.*, in Proc. Interspeech, 2012.

[10] T. Mausko *et al.*, IEIC Technical Report, vol. 101, no. 323, pp. 41–42, 2001.

[11] `http://hts.sp.nitech.ac.jp/`

[12] H. Kawahara *et al.*, Speech Communication, vol. 27, no. 3, pp. 187–207, 1999.

[13] T. Yoshimura *et al.*, in Proc. of Eurospeech, pp. 2347–2350, 1999.

[14] `http://www.jeita.or.jp/`

[15] 門脇 健人，亀岡 弘和，日本音響学会 2014 年秋季研究発表会講演集，3-7-3，pp. 261-264，Sep. 2014.