社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

複合ウェーブレットトラジェクトリモデルに基づくテキスト音声合成

北条　伸克†　　亀岡　弘和††　　嵯峨山茂樹†

† 東京大学大学院情報理工学系研究科
〒 113–8656 東京都文京区本郷 7–3–1
†† NTT コミュニケーション科学基礎研究所
〒 243–0198 神奈川県厚木市森の里若宮 3–1
E-mail: †{hojo,kameoka,sagayama}@hil.t.u-tokyo.ac.jp

あらまし　本稿は，高品質なテキスト音声合成を目指し，複合ウェーブレットモデル（composite wavelet model; CMW）と隠れマルコフモデル（hidden Markov　model; HMM）の統合モデルに，フォルマント周波数軌跡のモデルを組み込む．ケプストラム特徴量による従来の HMM 音声合成方式では，モデル学習時におけるケプストラム特徴量の平均化がスペクトルの周波数方向の平滑化の原因となり，一般に buzzy な合成音声へ劣化する原因となった．これに対し，フォルマント周波数に相当する特徴量の平均化はスペクトルの平滑化をもたらさないと期待される．このような観点から，我々は，CWM と HMM の統合モデルによる音声合成方式を過去に提案した．一方で，従来の CWM と HMM の統合モデルは，スペクトル系列の区分定常な生成モデルであり，ダイナミクスのモデル化に課題があった．本稿は，CWM がフォルマント周波数に相当するパラメータを持つことに着目し，フォルマント周波数軌跡の生成モデルを CWM と HMM の統合モデルに組み込み，スペクトル系列のダイナミクスをモデル化する．本稿では，実験を通して提案モデルがフォルマント周波数軌跡を十分に推定することを定性的に確認した．
キーワード　隠れマルコフモデル，複合ウェーブレットモデル，フォルマント周波数軌跡，補助関数法

Text-to-speech synthesis based on composite wavelet trajectory model

Nobukatsu HOJO†, Hirokazu KAMEOKA††, and Shigeki SAGAYAMA†

† Graduate School of Information Science and Technology, The University of Tokyo,
Hongo 7–3–1, Bunkyo, Tokyo, 113-8656, Japan
†† NTT Communication Science Laboratories
Morinosato 3–1, Wakamiya, Atsugi-shi, Kanagawa, 243-0198, Japan
E-mail: †{hojo,kameoka,sagayama}@hil.t.u-tokyo.ac.jp

Abstract　This paper integrates the formant frequency trajectory model into the combination of composite wavelet model (CWM) and the hidden Markov model (HMM) for text-to-speech synthesis (TTS) systems. We proposed TTS system based on CWM features to avoid the over-smoothing of spectra which is often caused by the conventional TTS system using cepstral features. Since CWM features represent formant frequencies, we model spectral dynamics by the formant frequency trajectory model and incorporate it into the TTS system which we proposed. We qualitatively confirmed through experiments that the proposed model estimated the formant frequency trajectories adequately.
Key words　hidden Markov model, composite wavelet model, formant frequency trajectory, auxiliary function

1. はじめに

本稿では，統計モデルに基づくテキスト音声合成の高品質化を目指し，フォルマント周波数軌跡モデルに基づく，スペクトルダイナミクスの生成モデル化を行う．統計モデルに基づくテキスト音声合成では，音声における様々な性質や挙動をいかに適切に生成モデルの形で記述できるかが合成音声の品質を左右する．特に，音声のダイナミクスを適切にモデル化することは，高品質なテキスト音声合成のために重要である．例えば，スペクトル包絡特徴量の動的特徴には話者の個人性が現れることが

知られている [1].

トラジェクトリ HMM [2] では，音素やコンテキストラベルなどの離散的なシンボルと連続的に時間変化する音声スペクトルとの関係を，メルケプストラムの静的，動的特徴を出力する隠れマルコフモデル (Hidden Markov Model; HMM) を用いてモデル化することにより，高品質な音声合成が可能となった．我々は，スペクトル包絡系列を生成モデル化する際，その時間変化を，トラジェクトリ HMM と同様にスペクトル特徴量のダイナミクスのモデルにより表現することを目指す一方で，スペクトル包絡系列の情報において特に重要であるフォルマント周波数軌跡に着目し，その生成過程を直接モデル化する手法を検討している．例えば，我々は，複合ウェーブレットモデル (Composite Wavelet Model; CWM) [3]〜[5] のフォルマント周波数に相当するパラメータの時間軌跡をモデル化することによりスペクトル包絡系列の生成過程をモデル化した [6].

フォルマント周波数軌跡のモデルを考える際，無声音のスペクトル包絡をいかに表現するかが問題となる．声帯を振動源，声道をフィルタとして有声音の生成過程を捉えたとき，フォルマント周波数は声道における共振周波数として捉えられ，発話する音素系列に対応した声道の形状変化に基づき時間方向になめらかに変化すると考えられる．[6] の手法は，このようなフォルマントの性質に着目し，提案されたものであった．一方で，呼気の摩擦や破裂などにより生成される無声音のスペクトル包絡では，その生成過程が上記の有声音のものとは異なり，異なるダイナミクスが仮定されるべきであると考えられる．したがって，フォルマントによりスペクトル包絡系列全体を表現するような生成モデルにおいては，特に無声音区間において表現に無理があり，フォルマント周波数軌跡も不自然なものとなる可能性がある．

本稿では，フォルマントに相当するスペクトル包絡ピークを持つ有声音スペクトルモデルと，平坦な形状を持つ無声音スペクトルモデルとの重畳によりモデル化されるスペクトル包絡の時系列としてスペクトル包絡系列全体が表現可能であると仮定し，有声音，無声音のスペクトルモデルを，それぞれ異なる生成過程を経る CWM パラメータセットにより表現するモデルを提案する．

2. 音声スペクトル生成過程の確率モデル化

2.1 複合ウェーブレットモデル

スペクトル包絡における各ピークを Gauss 分布関数で近似的に表現できるとすると，スペクトル包絡全体を Gauss 分布関数の重ね合わせ，すなわち混合 Gauss 分布関数モデル（Gaussian Mixture Model; GMM）で表現することができる．このようなスペクトル包絡の表現を，複合ウェーブレットモデル（Composite Wavelet Model; CWM）[3], [4] と呼ぶ．CWM におけるスペクトル包絡モデル $\phi_{\omega,l}$ は

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

で与えられる．ただし，ω, l は周波数と時刻のインデックス，

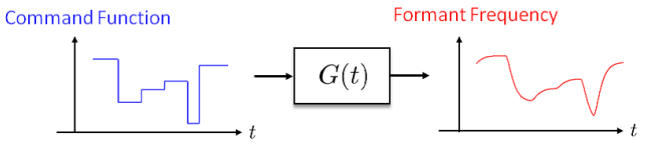


図 1　線形系によるフォルマント周波数軌跡の生成過程

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，$\mu_{k,l}, \sigma_{k,l}^2, w_{k,l}$ 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．

2.2 複合ウェーブレットトラジェクトリモデル

2.2.1 フォルマント軌跡の生成過程に関する仮説

声帯振動が共振することによって生じるフォルマントの周波数軌跡には声道の運動に伴う何らかの物理的な制約が付随すると考えられるが，フォルマント周波数軌跡が藤崎の F_0 パターン生成過程モデル [7] と同様のメカニズムによって生じると仮定する．具体的には，フォルマント周波数の対数の時間軌跡を，図 1 に示すように音素区間ごとに一定値をとる階段状の指令関数（以後，音素指令関数）にインパルス応答

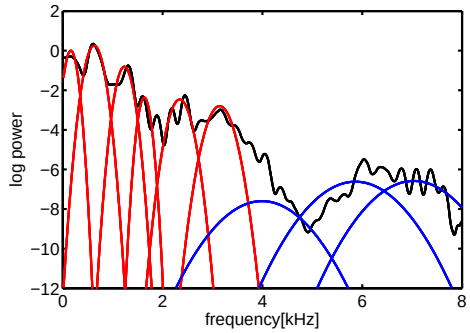
$$G(t) = \begin{cases} \alpha^2 t e^{-\alpha t} & (t \geqq 0) \\ 0 & (t < 0) \end{cases} \tag{2}$$

が畳み込まれ（α は固有角周波数），二次線形系の出力として生じたものと考える．なお，同様のフォルマント周波数軌跡のモデルが，規則音声合成 [8]，フォルマント周波数軌跡推定 [9] において過去に提案されている．また，二次線形系の仮定が置かれた他の音声生成過程モデルの例として，音素認識を目的とした調音運動の動的モデルが提案されている [10].

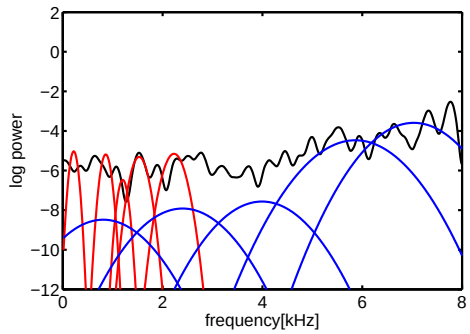
2.2.2 確率モデルの定式化

前述のフォルマント軌跡モデルに基づきスペクトル包絡系列をモデル化する手法を考える．ここで，1 章で述べたように，有声音のものとは性質の異なる無声音のスペクトル包絡をもフォルマントの存在を前提としてモデル化することは不自然である．したがって，有声音，無声音の双方が現れるスペクトル包絡系列をモデル化するためには，例えば，無声音のスペクトル包絡を別途モデル化する方法が考えられる．また，有声子音では，呼気の摩擦や破裂等と声帯の振動が同時に生じるため，そのスペクトル包絡は有声音，無声音の双方の特徴を併せ持つ可能性がある．

以上より，スペクトル包絡系列の生成過程は，有声音，無声音の特徴をそれぞれ持つスペクトルや，その双方の特徴を併せ持つスペクトルが各時刻で生成される過程として見なすことができる．このようなスペクトル包絡系列の生成過程を自然にモデル化するためには，例えば，有声音，無声音のスペクトルを別途モデル化し，それらが重畳されたものをスペクトル包絡系列のモデルとする方法が可能であると考えられる．複合ウェーブレットトラジェクトリモデルでは，有声音，無声音のスペクトルを，それぞれ異なる生成過程を経る CWM パラメータのセットによりモデル化し，それらを重畳することによ



(a) 音素/a/



(b) 音素/s/

図 2　有声音スペクトルモデル (赤) と無声音スペクトルモデル (青) の CWM パラメータの例

りスペクトル包絡系列のモデルとする．便宜的に，以後 GMM の分散パラメータ σ^2 の逆数 ρ をパラメータとみなし，有声音，無声音のスペクトルモデルの CWM パラメータセットをそれぞれ，$\{\boldsymbol{\rho}^{(v)}, \boldsymbol{w}^{(v)}, \boldsymbol{\mu}^{(v)}\}$，$\{\boldsymbol{\rho}^{(u)}, \boldsymbol{w}^{(u)}, \boldsymbol{\mu}^{(u)}\}$ とする．ここで $m = 1, \ldots, M$, M は有声音スペクトルモデルの GMM 混合数，$n = 1, \ldots, N$, N は無声音スペクトルモデルの GMM 混合数であり，$\boldsymbol{\rho}^{(v)} = \{\rho^{(v)}_{m,l}\}_{m,l}$，などとした．特に $\boldsymbol{\mu}^{(v)}$ はフォルマント周波数軌跡に対応すると見なすことができる．

提案モデルのパラメータ例を図 2 に示した．有声音/a/のスペクトルは赤で示された有声音スペクトルモデルのパラメータにより，無声音/s/のスペクトルは青で示された無声音スペクトルモデルのパラメータにより主にモデル化される．

有声音スペクトルの生成過程は，2.2.1 節のフォルマント軌跡の生成過程に関する仮説に基づき定式化される．音素指令関数の生成過程は，音素に対応する HMM 状態の出力分布として定式化することが可能である．次に，2.2.1 節で議論した様に，生成された指令関数 $\{u_{m,l}\}_{m,l}$ に二次線形系のインパルス応答が畳み込まれて Gauss 分布関数の平均値の軌跡 $\{\mu^{(v)}_{m,l}\}_{m,l}$ が生じるとする．提案モデルでは，それぞれの生成過程を，

$$P(u_{m,l}|s_l) = \mathcal{N}(u_{m,l}; m_{m,s_l}, \eta^2_{m,s_l}) \tag{3}$$

$$P(\mu^{(v)}_{m,l}|\{u_{m,l}\}_l) = \mathcal{LN}(G_{m,l} \overset{l}{*} u_{m,l}, \nu^2_m) \tag{4}$$

とする．ここで $\mathcal{LN}(x; \mu, \sigma^2)$ は対数正規分布であり，$\log x$ が正規分布 $\mathcal{N}(x; \mu, \sigma^2)$ に従うことと等価である．また，各フォルマントの帯域幅やパワーに対応するパラメータの生成過程を，

$$P(w^{(v)}_{m,l}|s_l) = \mathrm{Gamma}(w^{(v)}_{m,l}; a^{(v)}_{m,s_l}, b^{(v)}_{m,s_l}) \tag{5}$$

$$P(\rho^{(v)}_{m,l}|s_l) = \mathrm{Gamma}(\rho^{(v)}_{m,l}; c^{(v)}_{m,s_l}, d^{(v)}_{m,s_l}) \tag{6}$$

の様にモデル化する．ただし，$\mathrm{Gamma}(x; a, b)$ はガンマ分布

$$\mathrm{Gamma}(x; a, b) = x^{a-1} \frac{\exp(-x/b)}{\Gamma(a)\ b^a} \tag{7}$$

である．

以上のモデル化により，各フォルマント周波数，帯域幅及びパワーの先験的情報を，それぞれ (3)，(5) 及び (6) のパラメータの設定により，モデルに組み込むことが可能である．

続いて，無声音スペクトルの生成過程をモデル化する．提案モデルでは，

$$P(w^{(u)}_{m,l}|s_l) = \mathrm{Gamma}(w^{(u)}_{m,l}; a^{(u)}_{m,s_l}, b^{(u)}_{m,s_l}) \tag{8}$$

$$P(\rho^{(u)}_{m,l}|s_l) = \mathrm{Gamma}(\rho^{(u)}_{m,l}; c^{(u)}_{m,s_l}, d^{(u)}_{m,s_l}) \tag{9}$$

とモデル化し，各 Gauss 関数の平均パラメータは時間変化しないと仮定し，以後 $\mu^{(u)}_n$ と表記する．提案モデルでは特に，無声音スペクトル包絡の平坦な形状を表現するため，$\rho^{(u)}_{n,l}$ は，$\rho^{(u)}_{n,l}$ に比べより帯域幅が広くなるように生成確率を与える．

最後に，各時刻の CWM パラメータが与えられたときに観測スペクトル包絡 $y_{\omega,l}$ が生じる確率を

$$\phi_{\omega,l} = \sum_{m=1}^{M} \psi^{(v)}_{m,\omega,l} + \sum_{n=1}^{N} \psi^{(u)}_{n,\omega,l} \tag{10}$$

$$\psi^{(v)}_{m,\omega,l} = w^{(v)}_{m,l} \sqrt{\frac{\rho^{(v)}_{m,l}}{2\pi}} \exp\left(-\frac{\rho^{(v)}_{m,l}(\omega - \mu^{(v)}_{m,l})^2}{2} \right) \tag{11}$$

$$\psi^{(u)}_{n,\omega,l} = w^{(u)}_{n,l} \sqrt{\frac{\rho^{(u)}_{n,l}}{2\pi}} \exp\left(-\frac{\rho^{(u)}_{n,l}(\omega - \mu^{(u)}_{n})^2}{2} \right) \tag{12}$$

$$P(y_{\omega,l}|\boldsymbol{\rho}^{(v)}, \boldsymbol{w}^{(v)}, \boldsymbol{\mu}^{(v)}, \boldsymbol{\rho}^{(u)}, \boldsymbol{w}^{(u)}) = \mathrm{Poisson}(y_{\omega,l}; \phi_{\omega,l}) \tag{13}$$

とする．ここで $\mathrm{Poisson}(x; \lambda)$ は Poisson 分布である．なお，この仮定の下での λ の最尤推定問題は，スペクトル間の近さを測る尺度の一つとして近年音響信号処理分野で多用される I ダイバージェンスと呼ぶ歪み尺度を規準とした x と λ の最適フィッティング問題と等価となることが知られている [11]。

以上の提案モデルの推定すべきパラメータをまとめて $\Theta = \{\boldsymbol{\rho}^{(v)}, \boldsymbol{w}^{(v)}, \boldsymbol{\mu}^{(v)}, \boldsymbol{u}, \boldsymbol{\rho}^{(u)}, \boldsymbol{w}^{(u)}, \boldsymbol{\theta}^{(v)}, \boldsymbol{\theta}^{(u)}, \boldsymbol{s}\}$ とする．ここで新たに $\boldsymbol{u} = \{u_{m,l}\}_{m,l}, \boldsymbol{s} = \{s_l\}_l, \boldsymbol{\theta}^{(v)} = \{m_{m,i}, \nu^2_{m,i}, a^{(v)}_{m,i}, b^{(v)}_{m,i}, c^{(v)}_{m,i}, d^{(v)}_{m,i}\}_{m,i}, \boldsymbol{\theta}^{(u)} = \{a^{(u)}_{n,i}, b^{(u)}_{n,i}, c^{(u)}_{n,i}, d^{(u)}_{n,i}\}_{m,i}$ とした．次章では，観測スペクトル包絡系列 $Y = \{y_{\omega,l}\}_{\omega,l}$ が与えられた下での事後確率 $P(\Theta|Y)$ を最大化するパラメータ推定アルゴリズムについて述べる．

3. パラメータ推定アルゴリズム

$P(\Theta|\boldsymbol{y})$ を最大化する Θ を解析的に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．この時 $\log P(\Theta|\boldsymbol{Y})$ は，

$$\log P(\Theta|\boldsymbol{Y}) \overset{c}{=} \log P(\boldsymbol{Y}|\Theta) + \log P(\Theta) \tag{14}$$

$$\log P(\Theta) \overset{c}{=} \log P(\boldsymbol{\rho}^{(v)}|\boldsymbol{\theta}^{(v)}, \boldsymbol{s}) + \log P(\boldsymbol{w}^{(v)}|\boldsymbol{\theta}^{(v)}, \boldsymbol{s})$$

$$+\log P(\boldsymbol{u}|\boldsymbol{\theta}^{(v)},\boldsymbol{s})+\quad\log P(\boldsymbol{\mu}^{(v)}|\boldsymbol{u})$$
$$+\log P(\boldsymbol{\rho}^{(u)}|\boldsymbol{\theta}^{(u)},\boldsymbol{s})+\log P(\boldsymbol{w}^{(u)}|\boldsymbol{\theta}^{(u)},\boldsymbol{s})$$
$$+\log P(\boldsymbol{s}) \tag{15}$$

と書ける．ここで，$\stackrel{c}{=}$ は定数部分を除いた場合の等号を意味する．先述のように，$-\log P(\boldsymbol{Y}|\Theta)$ は定数項を除けば観測スペクトル包絡 $y_{\omega,l}$ とスペクトル包絡モデル $\phi_{\omega,l}$ との間の I ダイバージェンスと等しく [11]，さらに，

$$\begin{aligned}\sum_{\omega} w\sqrt{\frac{\rho}{2\pi}}\exp\left(\frac{\rho(\omega-\mu)^2}{2}\right)&\\ \simeq\int_{-\infty}^{\infty}\sqrt{\frac{\rho}{2\pi}}w\exp\left(\frac{\rho(w-\mu)^2}{2}\right)\mathrm{d}\omega&\\ =w&\end{aligned} \tag{16}$$

となることを用いれば，

$$\begin{aligned}-\log P(\boldsymbol{Y}|\Theta)&\stackrel{c}{=}\sum_{\omega,l}\left(y_{\omega,l}\log\frac{y_{\omega,l}}{\phi_{\omega,l}}-y_{\omega,l}+\phi_{\omega,l}\right)\\ &\stackrel{c}{=}\sum_{\omega,l}\left(\phi_{\omega,l}-y_{\omega,l}\log\phi_{\omega,l}\right)\\ &\simeq\sum_{m,l}w_{m,l}^{(v)}+\sum_{n,l}w_{n,l}^{(u)}-\sum_{\omega,l}y_{\omega,l}\log\phi_{\omega,l}\end{aligned} \tag{17}$$

が言える．この式の $-y_{\omega,l}\log\phi_{\omega,l}$ の項に対し，負の対数関数の凸性を利用し、Jensen の不等式を用いることで

$$\begin{aligned}-y_{\omega,l}\log\phi_{\omega,l}\leqq&-y_{\omega,l}\sum_{m}\gamma_{m,\omega,l}^{(v)}\log\frac{\psi_{m,\omega,l}^{(v)}}{\gamma_{m,\omega,l}^{(v)}}\\ &-y_{\omega,l}\sum_{n}\gamma_{n,\omega,l}^{(u)}\log\frac{\psi_{n,\omega,l}^{(u)}}{\gamma_{n,\omega,l}^{(u)}}\end{aligned} \tag{18}$$

のように上界関数を設計することができる．また，$-\log P(\boldsymbol{\mu}^{(v)}|\boldsymbol{u})$ の $\left(\sum_{\tau}G_{m,l-\tau}u_{m,\tau}\right)^2$ の項に対し，二次関数の凸性を利用し，同様に Jensen の不等式を用いることで

$$\left(\sum_{\tau}G_{m,l-\tau}u_{m,\tau}\right)^2\leqq\sum_{\tau}\frac{\left(G_{m,l-\tau}u_{m,\tau}\right)^2}{\lambda_{\tau,m,l}} \tag{19}$$

のように上界関数を設計することができる [12]．さらに $-\log P(\boldsymbol{\mu}^{(v)}|\boldsymbol{u})$ の中の $(\log\mu_{m,l}^{(v)})^2$ の項に関しては，

$$\begin{aligned}(\log\mu_{m,l}^{(v)})^2\leqq&\frac{1}{\mu_{m,l}^{(v)}}+\left(\frac{2\log\xi_{m,l}}{\xi_{m,l}}+\frac{1}{\xi_{m,l}^2}\right)\mu_{m,l}^{(v)}\\ &+|\log\xi_{m,l}|^2-2\log\xi_{m,l}-\frac{2}{\xi_{m,l}}\end{aligned} \tag{20}$$

のように上界関数を設計することができる [13]．ここで $\gamma_{m,\omega,l}^{(v)},\gamma_{n,\omega,l}^{(u)},\lambda_{\tau,m,l},\xi_{m,l}$ は補助変数である．式 (18)〜(20) より，$-\log p(\Theta|Y)$ の上界関数を設計することができ，これを補助関数として補助関数法を適用することができる．まず，補助変数の更新式は，上述の不等式の等号成立条件，すなわち，

$$\gamma_{m,\omega,l}^{(v)}=\frac{\psi_{m,\omega,l}^{(v)}}{\phi_{\omega,l}} \tag{21}$$

$$\gamma_{n,\omega,l}^{(u)}=\frac{\psi_{n,\omega,l}^{(u)}}{\phi_{\omega,l}} \tag{22}$$

$$\lambda_{\tau,m,l}=\frac{G_{m,l-\tau}u_{m,\tau}}{\sum_{\tau'}G_{m,l-\tau'}u_{m,\tau'}} \tag{23}$$

$$\xi_{m,l}=\mu_{m,l}^{(v)} \tag{24}$$

で与えられる．

各時刻の CWM パラメータについて，$\boldsymbol{\mu}^{(v)}$ 以外の更新式は，

$$u_{m,l}=\frac{\frac{m_{m,s_l}}{\eta_{m,s_l}^2}+\sum_{\tau\geqq l}\frac{G_{k,\tau-l}\log\mu_{m,\tau}}{\nu_m^2}}{\frac{1}{\eta_{m,s_l}^2}+\sum_{\tau\geqq l}\frac{G_{m,\tau-l}^2}{\nu_m^2\lambda_{l,m,\tau}}} \tag{25}$$

$$\rho_{m,l}^{(v)}=\frac{2(c_{m,s_l}^{(v)}-1)+\sum_{\omega}y_{\omega,l}\gamma_{m,\omega,l}^{(v)}}{\frac{2}{d_{m,s_l}^{(v)}}+\sum_{\omega}y_{\omega,l}\gamma_{m,\omega,l}^{(v)}(\omega-\mu_{m,l}^{(v)})^2} \tag{26}$$

$$\rho_{n,l}^{(u)}=\frac{2(c_{n,s_l}^{(u)}-1)+\sum_{\omega}y_{\omega,l}\gamma_{n,\omega,l}^{(u)}}{\frac{2}{d_{n,s_l}^{(u)}}+\sum_{\omega}y_{\omega,l}\gamma_{n,\omega,l}^{(u)}(\omega-\mu_{n}^{(u)})^2} \tag{27}$$

$$w_{m,l}^{(v)}=\frac{a_{m,s_l}^{(v)}-1+\sum_{\omega}y_{\omega,l}\gamma_{m,\omega,l}^{(v)}}{\frac{1}{b_{m,s_l}^{(v)}}+1} \tag{28}$$

$$w_{n,l}^{(u)}=\frac{a_{n,s_l}^{(u)}-1+\sum_{\omega}y_{\omega,l}\gamma_{n,\omega,l}^{(u)}}{\frac{1}{b_{n,s_l}^{(u)}}+1} \tag{29}$$

となる．$\boldsymbol{\mu}^{(v)}$ については，

$$p_3\mu_{m,l}^{(v)3}+p_2\mu_{m,l}^{(v)2}+p_1\mu_{m,l}^{(v)}+p_0=0 \tag{30}$$

$$p_3=\sum_{\omega}y_{\omega,l}\gamma_{m,\omega,l}^{(v)}\rho_{m,l}^{(v)} \tag{31}$$

$$p_2=\frac{2\xi_{m,l}\log\xi_{m,l}+1}{2\nu_m^2\xi_{m,l}^2}-\sum_{\omega}y_{\omega,l}\gamma_{m,\omega,l}^{(v)}\rho_{m,l}^{(v)}\omega \tag{32}$$

$$p_1=1-\frac{1}{\nu_m^2}\sum_{\tau\leqq l}G_{m,l-\tau}u_{m,\tau} \tag{33}$$

$$p_0=-\frac{1}{2\nu_m^2} \tag{34}$$

とおいて，式 (30) の正の解のうち $-\log P(\Theta|\boldsymbol{Y})$ を最も小さくする $\mu_{k,l}^{(v)}$ を選べばよい．

各状態出力分布のパラメータは，まず，$\{m_{m,i},\eta_{m,i}^2\}_{m,i}$ について，

$$m_{m,i}=\frac{1}{N_i}\sum_{l\in\mathcal{T}_i}u_{m,l} \tag{35}$$

$$\eta_{m,i}^2=\frac{1}{N_i}\sum_{l\in\mathcal{T}_i}u_{m,l}^2-(\frac{1}{N_i}\sum_{l\in\mathcal{T}_i}u_{m,l})^2 \tag{36}$$

となる．ただし，$\mathcal{T}_i=\{l|s_l=i\}$，$N_i$ は状態 $s_l=i$ となる時刻 l の数である．$\{a_{m,i}^{(v)},b_{m,i}^{(v)},c_{m,i}^{(v)},d_{m,i}^{(v)}\}_{m,l}$, $\{a_{n,i}^{(u)},b_{n,i}^{(u)},c_{n,i}^{(u)},d_{n,i}^{(u)}\}_{n,l}$ については，更新式は閉形式としては求まらないが，

$$\log a_{m,i}^{(v)}-\psi(a_{m,i}^{(v)})=\log(\frac{1}{N_i}\sum_{l\in\mathcal{T}_i}\rho_{m,l}^{(v)})-\frac{1}{N_i}\sum_{l\in\mathcal{T}_i}\rho_{m,l}^{(v)} \tag{37}$$

$$\log c_{m,i}^{(v)}-\psi(c_{m,i}^{(v)})=\log(\frac{1}{N_i}\sum_{l\in\mathcal{T}_i}w_{m,l}^{(v)})-\frac{1}{N_i}\sum_{l\in\mathcal{T}_i}w_{m,l}^{(v)} \tag{38}$$

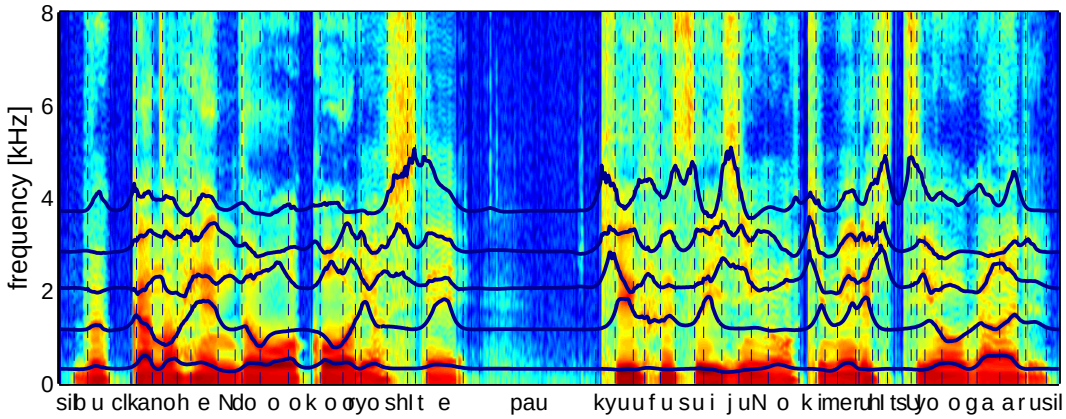


図 3　音声「物価の変動を考慮して，給付水準を決める必要がある」のスペクトル包絡系列 (カラーマップ表示) と提案法による推定フォルマント周波数軌跡 (実線) (点線は音素境界)

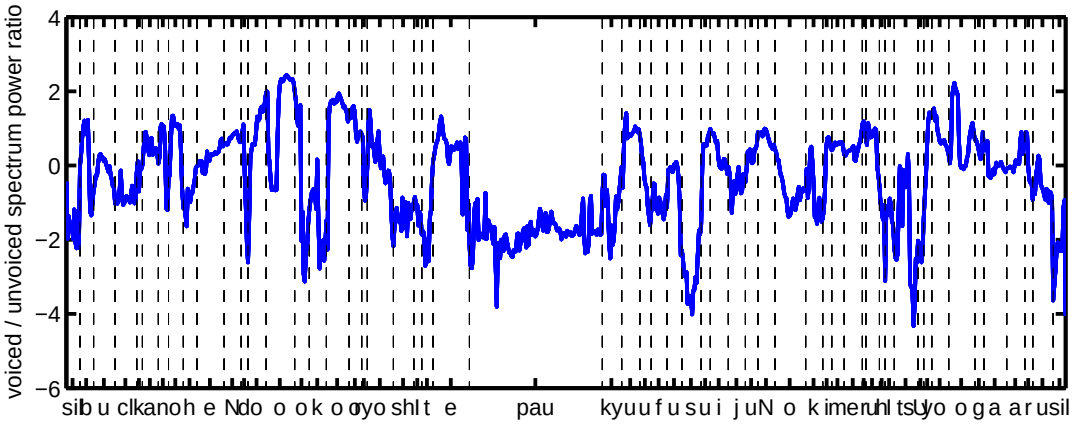


図 4　音声「物価の変動を考慮して，給付水準を決める必要がある」の有声音スペクトルモデルと無声音スペクトルモデルのパワー比の対数（実線）(点線は音素境界)

$$b_{m,i}^{(v)} = \frac{\frac{1}{N_i}\sum_{l\in\mathcal{T}_i}\rho_{m,l}^{(v)}}{a_{m,i}^{(v)}} \tag{39}$$

$$d_{m,i}^{(v)} = \frac{\frac{1}{N_i}\sum_{l\in\mathcal{T}_i} w_{m,l}^{(v)}}{c_{m,i}^{(v)}} \tag{40}$$

の根を求めることにより $\{a_{m,i}^{(v)}, b_{m,i}^{(v)}, c_{m,i}^{(v)}, d_{m,i}^{(v)}\}_{m,l}$ を更新することができる．また，$\{a_{n,i}^{(u)}, b_{n,i}^{(u)}, c_{n,i}^{(u)}, d_{n,i}^{(u)}\}_{n,l}$ についても同様である．ここで，$\psi(a)$ は digamma 関数

$$\psi(a) = \frac{\partial\Gamma(a)}{\partial a}/\Gamma(a) \tag{41}$$

である．

また，$P(\boldsymbol{s})$ については，通常の Viterbi Alignment による更新が可能である．

以上の更新則を十分な回数反復することで，$P(\Theta|\boldsymbol{Y})$ を局所最大化するパラメータ Θ を推定することができる．

4. 実　　験

提案モデルによるテキスト音声合成が可能であることを確認するための予備実験として，提案モデルのパラメータ推定を行い，動作を確認した．ただし，音響モデルは 1 状態の left to right HMM を仮定し，簡単のため，状態系列 $\boldsymbol{s}$ は音素ラベルにより与え固定した．また，従来の HMM 音声合成 [14] と同様に，提案モデルにおいても，フルコンテキストラベルに基づく各状態の状態出力分布に対する決定木によるクラスタリングを行うことが考えられるが，本実験では，簡単のため，音素を HMM の各状態とした．フォルマント CWM パラメータセットの混合数は M=5，CWM パラメータセットの混合数は N=7, $\alpha_m = 30$ $(m = 1\cdots M)$ とした．スペクトル包絡の解析は，STRAIGHT [15] により行った．

HTS2.1 のデモスクリプト [16] に同梱された男性話者の音声 (サンプリング周波数 16kHz· サンプルサイズ 16bit) のうち，450 文を学習データとして用いた．A04 文「物価の変動を考慮して，給付水準を決める必要がある」に対するパラメータ推定結果を図 3 および図 4 に示した．図 3 は，学習データから得られたスペクトル包絡系列に対し，推定されたパラメータ $\{\mu_{k,l}^{(v)}\}_{k,l}$ の時間軌跡を重ねて描いた図である．各時刻のスペクトル包絡ピークとパラメータ軌跡が重なることから，フォルマント周波数に相当するパラメータが正しく推定されたと考えられる．また，図 4 には，有声音スペクトルモデルと無声音スペクトルモデルのパワーの比の対数

$$\log\left(\frac{\sum_{m=1}^{M} w_{m,l}^{(v)}}{\sum_{n=1}^{N} w_{n,l}^{(u)}}\right) \tag{42}$$

を時刻に対しプロットした図を示した．有声音の音素区間では有声音スペクトルモデルのパワー比が大きく，無声音の音素区間に対しては，その逆となる傾向が現れた．この傾向から，有

声音の観測スペクトルは主に有声音スペクトルモデルにより，無声音の観測スペクトルは主に無声音スペクトルによりモデル化されていると考えられる．有声音，無声音の異なる生成過程を，二つのスペクトル生成モデルで表現した提案モデルにより，それぞれの生成過程のダイナミクスを表現するようなパラメータ系列が得られていることが期待される．

5. おわりに

本稿では，フォルマント周波数軌跡モデルに基づくスペクトル包絡系列の生成モデルとして，有声音，無声音のスペクトルをそれぞれ異なる生成過程を経た CWM パラメータセットによりモデル化し，それらの重畳によりスペクトル包絡系列をモデル化する，複合ウェーブレットトラジェクトリモデルを提案した．実験を通して，提案モデルが十分に観測スペクトルを表現し，フォルマント周波数軌跡が推定可能であることを確認した．今後は，テキスト音声合成実験を行い，品質を評価する予定である．

文　献

[1] 嵯峨山茂樹，板倉文忠，“音声の動的尺度に含まれる個人性情報，”日本音響学会春季研究発表会講演論文集，no.3–2–7，pp.589–590，1979.

[2] K. Tokuda, H. Zen, and T. Kitamura, “Trajectory modeling based on hmms with the explicit relationship between static and dynamic features,” Proc. of Eurospeech, pp.837–840, 2004.

[3] 槻武也，松本恭輔，酒向慎司，嵯峨山茂樹，“複合ウェーブレットモデルに基づく音声の分析合成，”電子情報通信学会技術研究報告，vol.105，no.372，pp.1–6，2005.

[4] P. Zolfaghari and T. Robinson, “Formant analysis using mixtures of gaussians,” in Proc. ICSLP ’96, vol.2, pp.1229–1232, 1996.

[5] N. Hojo, H. Kameoka, and S. Sagayama, “Text-to-speech synthesizer based on combination of composite wavelet and hidden markov models,” eigth ISCA Workshop on Speech Synthesis, pp.●●–●●, 2013.

[6] 吉里幸太，北条伸克，亀岡弘和，齋藤大輔，嵯峨山茂樹，“フォルマント周波数軌跡を潜在パラメータとした音声スペクトル生成過程の確率モデル，”日本音響学会春季研究発表会講演論文集，no.1–7–7，pp.277–280，2013.

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

[8] L.R. Rabiner, “Speech synthesis by rule: an acoustic domain approach,” PhD thesis, M. I. T, 1967.

[9] 板橋秀一，横山晶一，“線形 2 次系モデルによるホルマント軌跡の記述とセグメンテ-ション，”電子技術総合研究所彙報，vol.40，no.6，pp.p530–541，1976.

[10] 誉田雅彰，“調音モデルにもとづく音声の特徴抽出に関する研究，”PhD thesis，早稲田大学，1977.

[11] H. Kameoka, “Statistical approach to multipitch analysis,” PhD thesis, The University of Tokyo, 2007.

[12] 亀岡弘和，中谷智広，吉岡拓也，“音声のスパース性と非負制約つき畳み込みモデルに基づくパワースペクトル領域残響除去，”日本音響学会秋季研究発表会講演論文集，no.3-8-10，pp.705–708，2008.

[13] H. Kameoka, M. Nakano, K. Ochiai, Y. Imoto, K. Kashino, and S. Sagayama, “Constrained and regularized variants of non-negative matrix factorization incorporating music-specific constraints,” Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference onIEEE, pp.5365–5368 2012.

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

[15] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, “Restructuring speech representations using a pitch-adaptive time-freqency smoothing and an instantaneous-frequency-based f0 extraction,” Speech Communication, vol.27, no.3–4, pp.187–207, 1999.

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