複合ウェーブレットモデル分析合成系に基づく HMM 音声合成*

☆北条伸克 [1]，△南賢太郎 [2]， 齋藤大輔 [1]， 亀岡弘和 [1,3]， 嵯峨山茂樹 [1]
([1] 東大 情報理工, [2] 東大 工, [3]NTT CS 研)

1 はじめに

本稿では，HMM 音声合成において，その音声分析合成系として複合ウェーブレットモデル (Composite Wavelet Model; CWM) [1] を用いる手法を提案する．CWM では，混合ガウス関数モデル (Gaussian Mixture Model; GMM) によりスペクトル包絡を近似し，音声の特徴を GMM のパラメータで表現する (以下 CWM パラメータ)．合成時には，CWM パラメータに基づいて Gabor wavelet を生成し，ピッチ周期で重ね合わせることで音声波形を合成する．本稿では，CWM パラメータを特徴量とし，それに基づく波形生成を行う HMM 音声合成について議論する．

対話調の音声や感情音声といった多様な音声の合成を実現する上では，加工性に優れた合成手法が望まれる．音声の特徴がパラメトリックなモデルで表現された音声合成法は，パラメータの操作によって合成音の柔軟な操作が可能である．例えば LPC，PARCOR，LSP [2]，メルケプストラム合成 [3] などのパラメトリックな音声合成手法は，音声符号化の他，HMM 音声合成における分析合成系として利用される．

上述のパラメトリックな分析で得られる特徴量はスペクトル包絡の表現に違いがある．LSP は，スペクトルの周波数方向の関係に着眼した特徴表現であり，フォルマント周波数と凡そ対応している．このためパラメータの補間特性に優れており，また統計的学習に伴うスペクトル平滑化の影響が少ない．しかし，統計的学習後に得られたパラメータが安定性条件を満たす保証がなく，構成されたフィルタが不安定になる問題が内在する．一方，メルケプストラムに代表されるケプストラム型の特徴量は，スペクトル包絡の振幅に着眼した特徴といえる．特徴量の次元間の独立性が高く，ガウス分布を出力分布とするような統計的学習に適している反面，学習にともなうスペクトルの平滑化が顕著な問題の一つである．

LSP やメルケプストラムを用いた HMM 音声合成は，近年では波形接続型音声合成と同等以上の了解度の音声を合成可能である [4]．しかし合成音声の自然性については，改善の余地がある．これらのパラメトリックな音声合成の多くは巡回型フィルタを用いた手法であり，これらの手法ではフィルタの時間特性に関する問題が指摘されている [1]．を本研究では，LSP，メルケプストラムの双方の特徴量の利点を統合可能であり，かつ巡回型フィルタの問題点を解決しうる手法として，CWM 分析によって得られる特徴量を用いた HMM 音声合成に関して，その検討を行った．

2 CWM 分析合成系

2.1 複合ウェーブレットモデル (CWM)

巡回型フィルタでは，共振特性の Q 値が高いフィルタを倍音成分が共振周波数と一致する音源によって駆動した場合，時間制御特性の悪い音声が生成されることで，後続音に悪影響を与え，自然性劣化の一因となる [1]．CWM に基づく分析合成は，この問題を解決しうる FIR フィルタによる合成手法である [1]．CWM では，時間領域における Gabor wavelet と周波数領域におけるガウス関数の関係に着眼し，スペクトル包絡を GMM によって表現した上で，それに対応する重畳された Wavelet 関数を基本波形として，ピッチ周期毎に重ね合わせることで合成を実現する．また CWM ではスペクトル包絡が GMM で近似されるため，LSP が表現する周波数方向の関係性とケプストラムが表現する振幅方向の関係性を同時にモデル化していると解釈でき，双方の特徴量の利点を統合しうる．

2.2 CWM による音声分析

本章では，CWM を用いた音声分析及び分析によって得られたパラメータからの波形生成について述べる [1]．CWM では，音声スペクトル包絡 GMM で近似し，このパラメータがスペクトル包絡の特徴量表現とみなされる．この際，補助関数法によってスペクトル包絡と GMM によるモデルスペクトルとの間の距離尺度を逐次的に最小化することで，スペクトルパラメータが抽出される．本研究では距離尺度として，I-divergence

$$I(Y|F) = \sum_{\omega,t}[Y_{\omega,t}\log\frac{Y_{\omega,t}}{F_{\omega,t}} - Y_{\omega,t} + F_{\omega,t}] \qquad (1)$$

を用いた．ただし $Y_{\omega,t}$, $F_{\omega,t}$ は, それぞれ時刻 t における観測およびモデルスペクトル包絡を表す．$F_{\omega,t}$ は以下のような GMM で表現される．

$$F_{\omega,t} = \sum_{k=1}^{K}\frac{w_k}{\sqrt{2\pi\sigma_k^2}}\exp\left[-\frac{(\omega-\mu_k)^2}{2\sigma_k^2}\right] \qquad (2)$$

K は GMM の混合数を表し，本稿では以下これを CWM の混合数と呼ぶ．最終的に得られた CWM パラメータ $\mu_k, \sigma_k, w_k (k = 1, \ldots, K)$ を連結したものがスペクトル特徴量となる．なお，本研究では観測スペクトルを STRAIGHT 分析により得られたスペクトルとし，GMM 推定の際の μ_k の初期値として，$2K$ 次の LSP 解析によって得られたスペクトル対の平均を用いた．

*HMM-based speech synthesis using speech analysis based on Composite Wave Model. by HOJO Nobukatsu, MINAMI Kentaro, SAITO Daisuke, KAMEOKA Hirokazu, and SAGAYAMA Shigeki (The University of Tokyo)

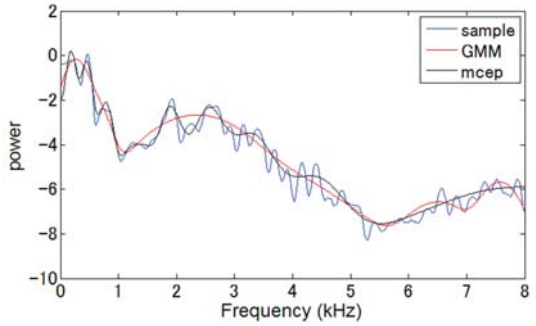


Fig. 1 音素/e/のスペクトル包絡近似; CWM (K=10), 24 次メルケプストラム

Fig. 1 に男声話者が発声した「あらゆる現実を...」の音素/e/の中心部分のスペクトル包絡，$K = 10$ の CWM を用いたモデルスペクトル，および 24 次のメルケプストラムによるスペクトル包絡近似を示す．Fig. 1 からは，メルケプストラムと比較すると低周波領域の近似精度は劣るものの，CWM はスペクトル包絡全体を緩やかに近似していることが確認される．

2.3 CWM 音声波形合成

CWM 特徴量と基本周波数情報を用いて，音声波形を合成する手法について述べる．2.1 節で挙げた巡回型フィルタの問題点を解決する方法として，GMM 包絡近似の逆フーリエ変換から得られる FIR 型フィルタを用いる．

有声音部分では，式 (3) で示されるようにガウス関数の逆フーリエ変換がガボール関数，つまりガウス関数と三角関数の積であることを考慮すると，GMM の逆フーリエ変換はガボール関数の重ね合わせで表現される [1].

$$\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(\omega-\mu)^2}{2\sigma^2}\right] \Leftrightarrow \frac{1}{2\pi}\exp\left[-\frac{\sigma^2 t^2}{2}+j\mu t\right] \quad (3)$$

時間領域において Gabor wavelet を基本周波数に対応する時間間隔で並べることにより有声音部分の波形が合成される．これは FIR フィルタを基本周波数に対応したデルタ列で駆動することに相当する．無声音部分では，CWM パラメータから得られたスペクトル包絡の各周波数 bin に乱数を掛け合わせ，零位相の逆フーリエ変換を行い，得られた波形に対しさらに時間領域で標準偏差が τ であるガウス窓を掛け合わせることで wavelet が得られる．乱数の掛け合わせはフィルタを雑音源で駆動することと等価であり，無声音部分では波形の非周期性が高くなることを考慮した合成方法である．また，ガウス窓の掛け合わせは，周波数特性を保ちつつ，合成した無声部の wavelet が 1 フレーム長 (64 ms) に広がることにより歯切れの悪い音声になることを防ぐ効果がある．この wavelet を等間隔 Δt で配置することにより無声音部分の波形が合成される．本研究では，$\tau = 7.5\,\mathrm{ms}$，$\Delta t = 5.5\,\mathrm{ms}$ とした．

3 HMM 学習のための CWM 特徴量抽出

3.1 フレーム単位特徴量抽出の問題点

本章では，CWM 分析合成系に基づく HMM 音声合成と，そのための特徴量抽出の改良について述べる．前章で述べた CWM 特徴量の抽出は，フレーム単位の処理であった．一方，一般に GMM のパラメータ推定は要素分布の順序の任意性が存在する．すなわちフレーム単位で独立にパラメータ推定を行った場合，異なるフレームにおいて，同一のインデックス k と要素分布の一貫性が保証されない．HMM 音声合成では複数のフレームから合成のための音響モデルを構築するため，この不整合によって品質が劣化する可能性がある．

3.2 CWM 特徴量の時間遷移確率モデル

3.1 節で述べた問題を解決するために，μ_k の時間変動について新たに確率モデルを導入する．μ_k の時間変動が平均 0 の正規分布に従うと仮定する．つまり，ある平均値特徴量が時刻 t にて $\mu_k^{(t)}$ であるとき，時刻 $t+1$ にて $\mu_k^{(t+1)}$ へ変化する遷移確率を，

$$P(\mu_k^{(t+1)}|\mu_k^{(t)}) = \mathcal{N}(\mu_k^{(t+1)}; \mu_k^{(t)}, \nu_k^2) \quad (4)$$

と定める．ここで，各インデックス k について，正規分布の分散 ν_k^2 は許容される μ_k の時間変動の程度を表す．この新たな確率モデルの導入は，抽出される平均値パラメータを時間方向に平滑化し，結果として順序が交代する確率を低減する効果を持つ一方で，遷移確率モデルの導入により，近似 GMM と，観測スペクトログラムとの誤差が増大する．

3.3 CWM 特徴量抽出アルゴリズムの改良

前述の時間遷移確率モデルを導入した場合，CWM 特徴量の抽出は観測スペクトログラムとモデルスペクトログラムとの違いを表す以下の目的関数の最小化によって実現される．

$$J(Y|F) = I(Y|F) + \sum_k^K \frac{1}{2\nu_k^2}\sum_{t=1}^{T-1}(\mu_k^{(t+1)} - \mu_k^{(t)})^2 \quad (5)$$

ただし T は音声から得られたスペクトルの全フレーム数である．[6] と同様に，Jensen の不等式を用いて補助関数を導入する．このとき，μ_k を除くパラメータについては，2 章と同じ更新式である．一方 μ_k については，時間遷移確率モデルを考慮した以下の式により更新される．時系列ベクトルを $\boldsymbol{\mu_k} = (\mu_k^{(1)}, \mu_k^{(2)}, \cdots, \mu_k^{(T)})^{\mathrm{T}}$ とすると，$\{\mu_k^{(t)}\}$ の更新式は，

$$\boldsymbol{\mu_k^*} = \frac{1}{2}(D_k + E_k)^{-1}F_k \quad (6)$$

と書ける．ただし，

$$D_k = \frac{1}{2\nu_k^2}\left\{D^{(i,j)}\right\}_{i,j} \quad (i,j \in [1,T]) \quad (7)$$

$$D^{(i,j)} = \begin{cases} 1 & (i=j=\{1,T\}) \\ 2 & (i=j\in[2,T-1]) \\ -1 & (|i-j|=1) \\ 0 & (\text{other}) \end{cases} \quad (8)$$

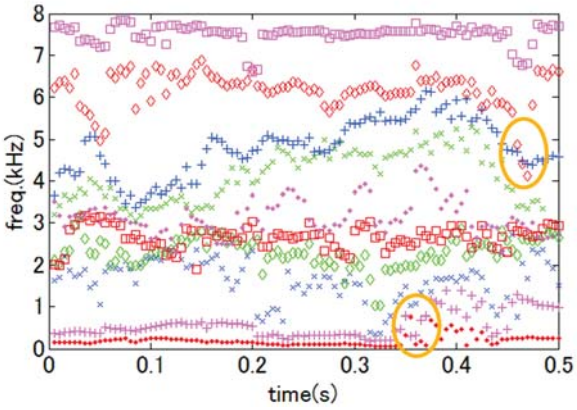


Fig. 2　Chain を導入しない場合の $\{\mu_k\}$ 抽出結果（混合数:10）　同一色かつ同一シンボルの列が 1 つの平均値特徴量 μ_k に対応する．楕円により強調した箇所に順序の入れ替えが確認される．

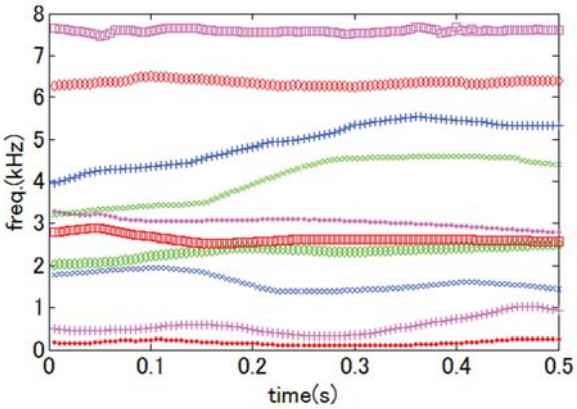


Fig. 3　Chain を導入した場合の $\{\mu_k\}$ 抽出結果 (混合数:10) 同一色かつ同一シンボルの列が 1 つの平均値特徴量 μ_k に対応する．

$$E_k = \frac{1}{2\sigma_k^2}\left\{E_k^{(i,j)}\right\}_{i,j} \quad (i,j \in [1,T]) \tag{9}$$

$$E_k^{(i,j)} = \begin{cases} \sum_\omega Y(\omega,i)\lambda_k(\omega,i) & (i=j) \\ 0 & (i \neq j) \end{cases} \tag{10}$$

$$F_k = \frac{1}{\sigma_k^2}(F_k^{(i)})_i \quad (i \in [1,T]) \tag{11}$$

$$F_k^{(i)} = \sum_\omega \omega Y(\omega,i)\lambda_k(\omega,i) \tag{12}$$

である．ただし，λ_k は補助関数法における補助変数である．

時間遷移確率モデルを導入しない場合の特徴量抽出の結果を Fig. 2 に，時間遷移確率モデルを導入した場合の結果を Fig. 3 に示した．用いた音声，GMM の混合数，初期値の設定は同一である．ν_k の値については，μ_k の時間変化の標準偏差がメル周波数軸上でほぼ一定となるよう定めた．μ_k の時間変化が平滑化され，順序の交代が減少しているとが確認できる．

4　CWM 特徴量を用いた HMM 音声合成

4.1　合成手順

CWM を分析合成系とする HMM 音声合成の手順を以下に示す．

1. 学習に用いる音声について，STRAIGHT 分析によって，基本周波数とスペクトル包絡を抽出した．3 章で述べたアルゴリズムによりスペクトル包絡から CWM 特徴量抽出を行った．
2. 各フレーム毎の GMM のパラメータ (μ_k, σ_k, w_k) と，さらにその 1 階，2 階の時間微分量 Δ, Δ^2, を併せてスペクトルパラメータとして CWM 特徴量を得る．さらに前のステップで得た基本周波数について，対数基本周波数，その 2 階までの時間微分を特徴量として連結した．
3. 前述の特徴量を用いて，HMM 音声合成のためのモデル学習を行う．学習されたモデルを用いて，評価用の文章に対して対応する CWM 特徴量と基本周波数の系列を生成する．
4. 2 章の CWM を用いた波形合成アルゴリズムにより，実際に音声を合成する．

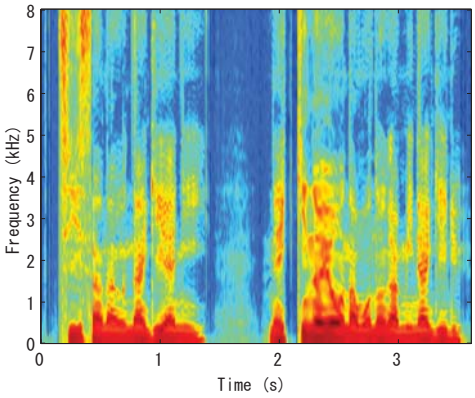
4.2　音声合成実験

CWM による特徴量表現を用いた統計的音声合成の音声再現性を確認するため，前述の合成手順に従い，実際に HMM 音声合成を行った．学習には，HTS 2.1 のデモスクリプト [7] に同梱された男性話者 1 名の音声（サンプリング周波数 16 kHz ・サンプルサイズ 16 bit ）のうち 450 文を用い，評価文章として 53 文を用いた[1]．フレームシフト 5ms で STRAIGHT 分析を行い，スペクトル包絡および基本周波数を抽出した後 K=10 として，CWM 特徴量を抽出した．音響モデルは 5 状態 left-to-right の隠れセミマルコフモデルであり，対数基本周波数については MSD-HMM とした．HMM の学習には HTS 2.1 を用いた [8]．CWM に基づく波形合成ではフレーム長 64 ms，フレームシフト 5 ms として合成を行った．なお特徴量系列生成の際，系列内変動（Global Variance; GV）を考慮しないパラメータ生成を行った．

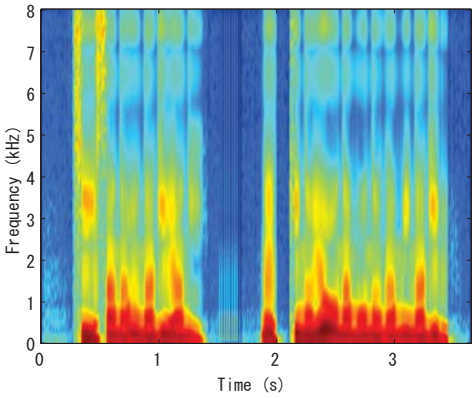
4.3　実験結果と考察

実験結果の例として，ATR の J02 文「泥棒でも入ったかと，一瞬僕は思った」について，サンプル音声と合成音声のスペクトログラムを，それぞれ Fig. 4(a), Fig. 4(b) にそれぞれ示す．また，先頭の音素/o/の中央付近のフレームについて，(1) サンプル音声のスペクトル包絡 (青) (2) CWM 特徴量により合成した音声のスペクトル包絡 (赤) (3) 従来法で合成した音声のスペクトル包絡 (黒) を Fig. 5 に示す．ただし，従来法とは，24 次のメルケプストラムによる手法である [3]．提案手法による合成音声は，スペクトル包絡・スペクトログラムがおおむね再現されていることが確認された．聴感上，特に有声音・無声音が混ざった音声について，メルケプストラムによる合成音声では，無声部のガウス雑音駆動の音声が後の有声部の

[1]発話内容は ATR 音素バランス分であり，学習データがサブセット A-I，評価文章がサブセット J に対応している．



(a) サンプル音声



(b) 合成音声

Fig. 4　スペクトログラム (「小さな鰻屋に熱気の様なものがみなぎる」)

音声に影響し，雑音が混ざったような音になる傾向があったが，提案手法ではこのような傾向が改善された．これは，FIR 型フィルタを用いた効果であると考えられる．一方で，提案法により合成した音声は，サンプル音声やメルケプストラムによる合成音声に比べ，こもったような印象を与える音声になった．主な原因は，Fig. 5 に確認されるような，フォルマントの平滑化であると考えられる．フォルマントが平滑化する傾向は，他の母音についても同様に確認された．この原因としては，

- 要素分布の対応を一貫させるために，特徴量の時間遷移モデルのみでは不十分であり，特徴量が平均化した
- CWM の混合数がスペクトル包絡を表現するためには不十分であった
- 特徴量の時間遷移モデルの導入が悪影響を及ぼした

などが考えられ，品質向上のためには，w_k に対する G V の導入や，特徴量抽出手法について改善の余地がある．

5　おわりに

本研究では，巡回型フィルタの時間制御特性の問題を解決し，LSP，ケプストラムの双方の特徴量の利点

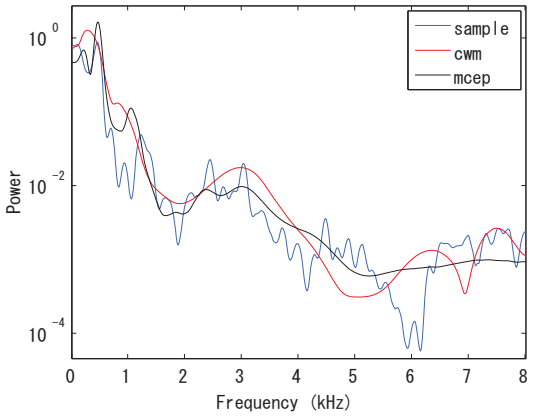


Fig. 5　音素/o/のスペクトル包絡再現結果

を統合する可能なモデルである CWM 分析合成系を利用し，これに基づく特徴量を用いた HMM 音声合成を検討した．提案法により合成された音声と，実際に読み上げられた音声のスペクトログラムは類似しており，HMM 音声合成が可能であることが確認された．特に有声音・無声音が混ざった音声について，駆動雑音の影響が抑えられ，FIR フィルタを用いる効果が現れたと考えられる．一方，フォルマントが平滑化し，音声がこもる傾向があり，本手法の合成音声の品質劣化の原因となっていると考えられる．特徴量抽出手法の更なる改良や，w_k に対する GV の導入などの工夫を行うことで，高品質な音声合成が可能になると考えられる．音声の品質向上に取り組むとともに，合成音声の主観評価実験を行い従来法と比較するなど，品質に対する定量評価を行う予定である．

参考文献

[1] 槻武也他，“複合ウェーブレットモデルによる音声合成の検討，”日本音響学会 2006 年春季研究発表会講演論文集，2-11-7，2006.
[2] 管村昇他，“線スペクトル対（LSP）音声分析合成方式による音声情報圧縮，”電子通信学会論文誌，J64-A，pp. 599-606，1981.
[3] 徳田恵一他，“動的特徴量を用いた HMM からの音声パラメータ生成アルゴリズム，”日本音響学会誌，vol.53，no.3，pp.192-200，1997.
[4] K. Hashimoto *et al.*, Proc. Blizzard Challenge Workshop 2011.
[5] Hideki Kawahara *et al.*, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0extraction, ”Speech Communication, Vol. 27, No. 3-4, pp. 187-207, 1999.
[6] 亀岡弘和他，“スペクトル包絡と調波構造の合成関数モデルによる音声分析，”日本音響学会 2005 年秋季研究発表会講演論文集，2-6-4，2005.
[7] http://hts.sp.nitech.ac.jp/
[8] 全 炳河他，“HMM 音声合成システム (HTS) の開発，”情報処理学会研究報告，SLP 音声言語情報処理 2007 (129)，301-306，2007-12-20，2007.