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.

HMM 音声合成における変調スペクトルに基づくポストフィルタ

高道慎之介† 戸田 智基† Graham Neubig† Sakriani Sakti† 中村 哲†

† 奈良先端科学技術大学院大学 〒 630–0192 奈良県生駒市高山町 8916–5
E-mail: †{shinnosuke-t,tomoki}@is.naist.jp

あらまし 隠れマルコフモデル (Hidden Markov Model : HMM) に基づく音声合成において，生成される音声パラメータ系列は過剰に平滑化される傾向にあり，合成音声の肉声感は劣化する．系列内変動 （Global Variance: GV）は，過剰な平滑化を定量的に表現する特徴量として広く使用されるが，GV を考慮して生成されたパラメータ系列は，未だに平滑化の影響を強く受ける．これに対し本稿では，過剰な平滑化の影響を表現する新たな特徴量としてパラメータ系列の変調スペクトル （Modulation Spectrum: MS）に着目し，生成パラメータ系列の MS を補償するポストフィルタを提案する．スペクトル及び F_0 に対して提案法を適用し，その有効性を実験的に評価する．評価結果から，提案法により音質改善効果が得られることを示す.

キーワード HMM 音声合成，過剰な平滑化，系列内変動，変調スペクトル，ポストフィルタ

Postfilter Based on Modulation Spectrum in HMM-Based Speech Synthesis

Shinnosuke TAKAMICHI†, Tomoki TODA†, Graham NEUBIG†, Sakriani SAKTI†, and Satoshi NAKAMURA†

† Nara Institute of Science and Technology, Tatayama-cho 8916–5, Ikoma, Nara, 630–0192 Japan
E-mail: †{shinnosuke-t,tomoki}@is.naist.jp

Abstract In this paper, we propose a postfilter based on a modulation spectrum for high-quality HMM-based speech synthesis. An over-smoothing effect that is observed in the generated speech parameter sequence is a main cause of quality degradation in HMM-based speech synthesis. A Global Variance (GV) is well-known as a better feature to capture the over-smoothing effect, and an effectiveness of the parameter generation algorithm considering the GV have been confirmed. However, the quality gap between natural speech and synthetic speech is still large. In this paper, we introduce a Modulation Spectrum (MS) of speech parameter trajectory as a new feature to effectively capture the over-smoothing effect. The generated speech parameter sequence is filtered to compensate its MS. Experimental results show that quality improvements by the proposed methods applied to spectral and F_0 components is yielded compared with conventional HMM-based speech synthesis.

Key words HMM-based speech synthesis, postfilter, modulation spectrum, over-smoothing, global variance

1. はじめに

テキスト音声合成（Text-To-Speech : TTS）は，任意のテキストから音声を合成する技術であり，音声をインターフェースとしたコミュニケーションシステムにおいて大きな役割を担っている．隠れマルコフモデル（Hidden Markov Model : HMM）に基づく音声合成技術 [1] は，統計的手法に基づく TTS の一つであり，声質制御の柔軟性 [2]〜[4] などの利点から，広く利用されている．HMM に基づく音声認識の分野で蓄積された手法を活用可能な点も，HMM 音声合成が利用される理由の一つである．一方で，統計処理による過剰な平滑化処理により，生成される音声パラメータ系列の詳細な特徴は失われ，合成音声の音質は，自然音声と比較して著しく劣化する [5] .

パラメータの系列内変動（Global Variance : GV）[6] は，過剰な平滑化を定量的に説明する特徴量として広く知られている．特徴量自体は，パラメータ系列の 2 次モーメントというシンプルな形式で表現されるものの，GV を考慮したパラメータ生成法は，平滑化の影響を比較的抑えたパラメータ系列を生成

可能である．しかしながら，その合成音声の音質は，自然音声の音質と比較すると未だに大きく劣化している．

本稿では，合成音声の音質改善を目的として，パラメータ系列の変調スペクトル （Modulation Spectrum: MS） に基づくポストフィルタを提案する．MS は，パラメータ系列のパワースペクトルとして定義され，音声知覚に関するスペクトルキュー [7] や音声認識におけるスペクトルパラメータ [8] として使用される特徴量である．HMM 音声合成において生成されるパラメータ系列の MS は，GV を考慮した場合においても，自然音声の MS と比較して大きく減衰する傾向にある．そこで，生成されたパラメータ系列の MS を補償するポストフィルタを適用する．ポストフィルタは，学習データ内の自然音声および合成音声のパラメータ系列から事前に学習される．また，スペクトルのみでなく，F_0 に対しても適用可能である．提案法により，GV を考慮したパラメータ生成法と比較し，合成音声の音質を改善できることを示す．

2. HMM 音声合成のパラメータ生成法

2.1 HMM 尤度最大化基準 [9]

HMM 音声合成では，自然音声のパラメータ系列からコンテキスト依存 HMM を学習する．生成時には，合成対象のテキストに対応する文 HMM を形成し，静的・動的特徴量間の明示的な制約条件の下で HMM 尤度を最大化することで，パラメータ系列を生成する．HMM 尤度のみを考慮したパラメータ生成は次式で示される．

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmax}}\, P\left(\boldsymbol{W}\boldsymbol{c}|\boldsymbol{\lambda}\right) \tag{1}$$

ただし，$\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \cdots, \boldsymbol{c}_t^\top, \cdots, \boldsymbol{c}_T^\top\right]^\top$ は T フレームの音声パラメータ系列，$\boldsymbol{c}_t = \left[c_t\left(1\right), \cdots, c_t\left(d\right), \cdots, c_t\left(D\right)\right]^\top$ は時刻 t における D 次元の音声パラメータ，d は次元のインデックス，$\boldsymbol{W}$ は動的特徴量の計算に用いる重み係数によって決定される行列 [9]，$\boldsymbol{\lambda}$ は HMM のパラメータセットを表す．

式 (1) により生成されるパラメータ系列は，自然音声パラメータ系列と比較して，過剰に平滑化される傾向にあり，こもった音質の合成音声を生み出す要因となる．

2.2 HMM 尤度・GV 尤度最大化基準 [6]

GV は，パラメータ系列の 2 次モーメントとして定義される．次式に示す通り，各次元のパラメータ系列の変動はスカラで表現される．

$$\boldsymbol{v}\left(\boldsymbol{c}\right) = \left[v\left(1\right), \cdots, v\left(d\right), \cdots, v\left(D\right)\right]^\top \tag{2}$$

$$v\left(d\right) = \frac{1}{T}\sum_{t=1}^{T}\left(c_t\left(d\right) - \frac{1}{T}\sum_{\tau=1}^{T} c_\tau\left(d\right)\right)^2 \tag{3}$$

生成時には，次式に示すように，HMM 尤度及び GV 尤度を最大化してパラメータ系列を生成する．

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmax}}\, P\left(\boldsymbol{W}\boldsymbol{c}|\boldsymbol{\lambda}\right) P\left(\boldsymbol{v}\left(\boldsymbol{c}\right)|\boldsymbol{\lambda}_v\right)^w \tag{4}$$

ただし，$\boldsymbol{\lambda}_v$ は GV の確率密度関数のパラメータセット，w は GV 尤度の重みを表す．GV の確率密度関数は，自然音声のパラメータ系列から学習される．

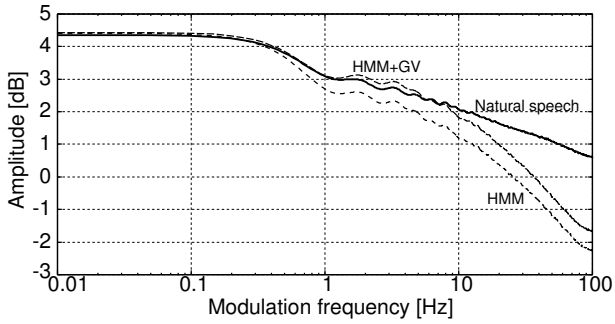


図 1　横軸を対数変調周波数とした第 10 次メルケプストラム系列の変調スペクトルの例

Fig. 1　An example of natural/generated 10th mel-cepstral coefficient sequences of log-scaled modulation frequency.

式 (1) による生成パラメータの GV は，通常，自然音声のパラメータ系列の GV より小さくなる傾向にある．一方で，式 (4) による生成パラメータ系列の GV は，GV 尤度を考慮することで補償され，合成音声の音質は大きく改善する．しかしながら，その音質は，依然として自然音声のものには及ばない．

3. パラメータ系列の変調スペクトル

MS は，本来，パラメータ系列をフーリエ変換した値を表す [10] が，本稿では，その対数パワースペクトルを MS と呼ぶ．パラメータ系列 $\boldsymbol{c}$ に対する変調スペクトル $\boldsymbol{s}\left(\boldsymbol{c}\right)$ を次式で定義する．

$$\boldsymbol{s}\left(\boldsymbol{c}\right) = \left[\boldsymbol{s}\left(1\right)^\top, \cdots, \boldsymbol{s}\left(d\right)^\top, \cdots, \boldsymbol{s}\left(D\right)^\top\right]^\top \tag{5}$$

$$\boldsymbol{s}\left(d\right) = \left[s_d\left(0\right), \cdots, s_d\left(m\right), \cdots, s_d\left(M\right)\right]^\top \tag{6}$$

ただし，$s_d\left(m\right)$ は，d 次元目のパラメータ系列 $\left[c_1\left(d\right), \cdots, c_t\left(d\right), \cdots, c_T\left(d\right)\right]^\top$ に対する，周波数インデックス m の MS，M は離散フーリエ変換（Discrete Fourier Transform : DFT）のサンプル数の半分を表す．本稿では，系列長が $2M$ になるように零詰めをしたパラメータ系列の MS を計算する．

図 1 に，式 (1)（“HMM”）と式 (4)（“HMM+GV”）で生成された第 10 次メルケプストラム系列の MS の平均を示す．比較のため，自然音声（“Natural speech”）の同系列の MS の平均も示す．いずれの MS も，低変調周波数にパワーが集中していることが分かる．“HMM” の MS は，自然音声のパラメータ系列の MS と比較して，大きく減衰していることが確認できる．これは，隠れマルコフモデルによる時間方向の平滑化によって生じるものであると予想される．“HMM+GV” の MS は，GV の導入により比較的補償され，低変調周波数の MS は “Natural speech” の MS に接近するが，それ以外の周波数においては，未だに大きく減衰していることが確認できる．

以上の結果より，MS が音質に寄与することが予想され，また，MS の直接的な補償により，合成音声の音質改善がもたらされると期待される．

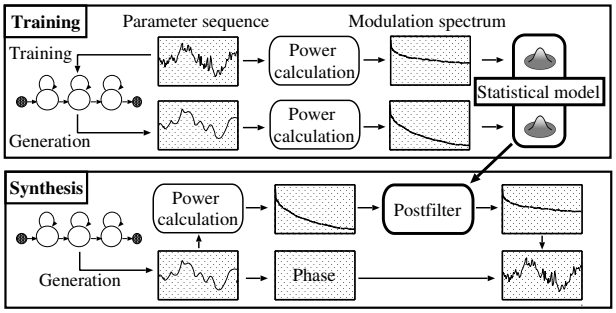


図 2　提案するポストフィルタの学習・生成手順

Fig. 2　A schematic diagram of proposed training/synthesis processes.

4.　変調スペクトルに基づくポストフィルタ

生成パラメータ系列の MS を補償するポストフィルタを提案する．図 2 に示す手順の通り，ポストフィルタは学習データを用いて事前学習する．

4.1　学　習　部

自然音声のパラメータ系列から，次式に示す確率密度関数を学習する．

$$P\left(\boldsymbol{s}\left(\boldsymbol{c}\right)|\boldsymbol{\lambda}_s\right)=\mathcal{N}\left(\boldsymbol{s}\left(\boldsymbol{c}\right);\boldsymbol{\mu}^{(\mathrm{N})},\boldsymbol{\Sigma}^{(\mathrm{N})}\right) \tag{7}$$

ただし，$\mathcal{N}\left(\cdot;\boldsymbol{\mu}^{(\mathrm{N})},\boldsymbol{\Sigma}^{(\mathrm{N})}\right)$ は平均 $\boldsymbol{\mu}^{(\mathrm{N})}=\left[\mu_{1,0}^{(\mathrm{N})},\cdots,\mu_{D,M}^{(\mathrm{N})}\right]^{\top}$ と対角共分散行列 $\boldsymbol{\Sigma}^{(\mathrm{N})}=\mathrm{diag}\left[\left(\sigma_{1,0}^{(\mathrm{N})}\right)^2,\cdots,\left(\sigma_{D,M}^{(\mathrm{N})}\right)^2\right]$ の正規分布，$\mu_{d,m}^{(\mathrm{N})}$ と $\left(\sigma_{d,m}^{(\mathrm{N})}\right)^2$ は $s_d\left(m\right)$ の平均と分散，$\boldsymbol{\lambda}_s$ は MS の確率密度関数のパラメータセットを表す．同様に，HMM 音声合成で生成されたパラメータ系列から正規分布 $\mathcal{N}\left(\cdot;\boldsymbol{\mu}^{(\mathrm{G})},\boldsymbol{\Sigma}^{(\mathrm{G})}\right)$ を学習する．なお，自然音声のパラメータ系列と生成パラメータ系列間の継続長の違いが MS に影響することを避けるために，正規分布の学習に用いる生成パラメータ系列は，自然音声の継続長において生成する．

4.2　生　成　部

生成されたパラメータ系列 $\boldsymbol{c}$ の MS に対して次式のポストフィルタを適用する．

$$\begin{aligned}s'_d\left(m\right)\;=\;&\left(1-k\right)s_d\left(m\right)\\&+\;k\left[\frac{\sigma_{d,m}^{(\mathrm{N})}}{\sigma_{d,m}^{(\mathrm{G})}}\left(s_d\left(m\right)-\mu_{d,m}^{(\mathrm{G})}\right)+\mu_{d,m}^{(\mathrm{N})}\right]\end{aligned} \tag{8}$$

ただし，k はポストフィルタ強度係数 $(0\leqq k\leqq 1)$ を表す．フィルタ後の MS は，$k=1$ の際には自然音声のパラメータ系列の MS に近い値となり，$k=0$ の際にはポストフィルタ処理前と等価となる．ポストフィルタ後のパラメータ系列は，式 (8) で計算された MS と，フィルタ処理前のパラメータ系列の周波数位相特性から計算する．

4.3　F0 パラメータへの適用

提案するポストフィルタは連続遷移する時系列に対するフーリエ変換を用いるため，F_0 系列に対する直接的な適用は不可能である．これに対して本稿では，無声区間においても連続的な F_0 パターンが観測されるものとしてモデル化を行う手法である，連続 F_0 モデル [11] を導入する．[12] と同様に，無声区間に対してスプライン法に基づく補間処理を行うことで，連続的な F_0 パターンを生成する．ただし，無音区間の F_0 推定は補外処理となるため，スプライン法による推定精度の劣化が予想される．故に，有音区間のみの連続 F_0 パターンを用いて MS を計算する (注1)．推定された連続 F_0 モデルに対しては，知覚的な影響の小さいマイクロプロソディ [13] による MS の変動を避けるために，低域通過フィルタ（Low Pass Filter : LPF）を適用する．また．連続 F_0 パターンを推定する前に，零詰め処理による不連続な遷移を避けるために，系列平均を 0 にするように F_0 パターンにバイアスをかける．

合成時には，生成 F_0 パターンの連続化を行う前に，系列の平均 F_0 と有声/無声区間を抽出する．ポストフィルタ後の連続 F_0 パターンは，スペクトルパラメータの場合と同様に計算され，最終的な F_0 パターンは，この連続 F_0 パターンに対して，平均 F_0 と有声/無声区間を復元することで生成する．マイクロプロソディが除去されているため，フィルタ強度係数 k を 0 に設定した場合でも，フィルタ処理前後の F_0 パターンは一致しないことに注意する．

4.4　GV ポストフィルタとの関連性

パラメータ系列の GV を補償するポストフィルタ [14] との関連性について議論する．GV ポストフィルタでは，生成パラメータに対して次式の線形変換を行う．

$$\hat{c}_t\left(d\right)=\sqrt{\frac{\mu_d^{(\mathrm{GV,N})}}{\mu_d^{(\mathrm{GV,G})}}}\left\{c_t\left(d\right)-\left\langle c_t\left(d\right)\right\rangle\right\}+\left\langle c_t\left(d\right)\right\rangle \tag{9}$$

ただし，$\mu_d^{(\mathrm{GV,N})}$，$\mu_d^{(\mathrm{GV,G})}$ はそれぞれ，学習データにおける d 次元目の自然音声パラメータ及び合成音声パラメータの GV の平均，$\left\langle c_t\left(d\right)\right\rangle$ は，d 次元目の合成音声パラメータの平均を表す．このポストフィルタでは時系列の分散値のみを補償するため，図 1 に示したように依然として MS には大きな違いが生じており，自然音声のパラメータ系列に含まれる時間的変動の再現は困難である．一方で，提案法では全変調周波数の MS を直接補償するため，自然音声のパラメータ系列のような時間的変動を含んだパラメータ系列を生成することができる．そのため，提案法による音質改善が予想される．

パーセバルの定理より，フーリエ変換前後のパワーは一致する．式 (3) で定義された GV は，直流成分を除いた系列のパワーに相当する．一方で，MS は系列のパワースペクトルであるため，直流成分を除いた全変調周波数の MS の和は GV と一致する (注2)．つまり MS とは，直流成分を除いた要素の和が GV に一致するような特徴量であると解釈出来る．また，GV ポストフィルタ処理は，音声パラメータの各次元において，直流成分を除いた全変調周波数の MS を一律に定数倍することに相当し，ポストフィルタ係数 1.0 における提案法の変換処理に，

(注1)：無音区間の F_0 推定法には，系列の平均 F_0 による置換，または，最近傍の有声フレームの F_0 による置換等が考えられるが，テストデータに対する MS 尤度を計算した結果，尤度が最も高くなった上記の手法を用いる．

(注2)：正確には，線形パワーの和が GV と一致する．

次式の制約を付与することに等しい (注3).

$$\mu_{d,m}^{(\cdot)} = 0, \quad \sigma_{d,m}^{(\cdot)} = \begin{cases} 1 & m = 0 \\ \mu_d^{(\mathrm{GV},\cdot)} & \text{otherwise} \end{cases} \tag{10}$$

一方で，提案法では，各変調周波数において異なる倍率を用いて MS を補償することが可能である．

5. 実験的評価

5.1 実験条件

学習データは女性話者による ATR 音素バランス文 [15] A-I セット 450 文とする．評価データは同 J セット 53 文を使用する．学習データのサンプリング周波数は 16 kHz，フレームシフトは 5 ms とする．スペクトル特徴量は，STRAIGHT 分析 [16] による 0 次から 24 次のメルケプストラム係数，音源特徴量は，対数 F_0，5 周波数帯域における平均非周期成分 [17], [18] を使用する．5 状態 left-to-right 型の隠れセミマルコフモデル（Hidden Semi-Markov Model : HSMM） [19] の学習を行う．変調スペクトルにおける DFT のサンプル数は 4096 点とする．これは，学習・評価データのパラメータ系列のフレーム数を十分に超える値である．マイクロプロソディを除去するための LPF のカットオフ周波数は，10Hz とする (注4).

以下に示す手法を用いて評価を行う．“HMM+GV+MS” に用いる MS の確率密度関数の学習には，GV を考慮して生成されたパラメータを使用することに注意する．

“HMM”： 式 (1) で生成

“HMM+MS”： 式 (1) で生成したパラメータ系列に対して提案法を適用

“HMM+GV”： 式 (4) で生成

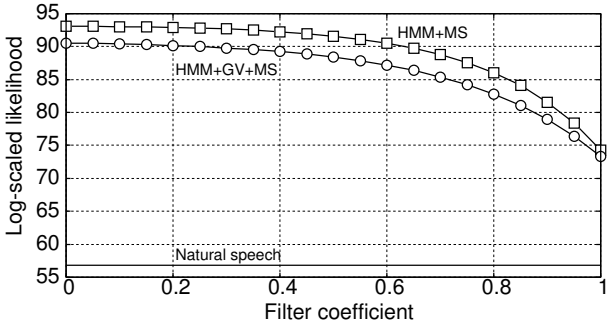
“HMM+GV+MS”： 式 (4) で生成したパラメータ系列に対して提案法を適用

まず，ポストフィルタ強度係数を決定するための評価を行う．ポストフィルタ強度係数を 0 から 1 まで 0.05 刻みで変化させ，ポストフィルタ処理後のパラメータ系列に対する HMM 尤度，GV 尤度及び MS 尤度を計算する．同時に，自然音声（Natural speech）のパラメータ系列に対する尤度も計算する．次に，提案法による音質改善効果を対比較実験により評価する (注5)．評価者には，ランダムに再生された音声から音質の高い方を強制選択させた．評価は 8 人の受聴者に対するプリファレンススコアとする．客観・主観評価はスペクトル・F_0 毎に行い，提案法を適用しない音声パラメータは，“HMM” を使用する．

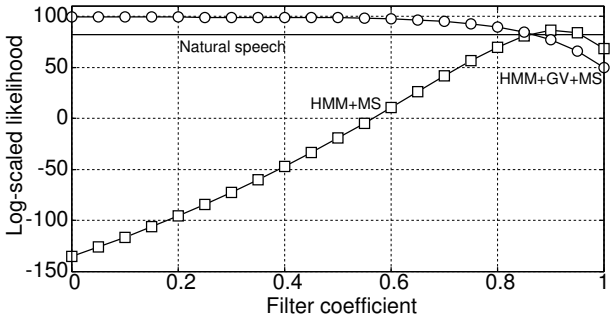
5.2 HMM・GV・MS 尤度を用いた客観評価結果

ポストフィルタ強度係数を変化させた時の，ポストフィルタ後のスペクトルパラメータ系列に対する HMM 対数尤度を図 3(a) に，GV 対数尤度を図 3(b) に，MS 対数尤度を図 3(c) に示す．図 3(a) から，ポストフィルタ強度係数を大きくするに従い，生成パラメータ系列に対する HMM 尤度は大きく減少することがわかる．しかしながら，その尤度は自然音声のパラメータ系列に対する HMM 尤度よりも依然として大きい．一方，図 3(b) から，ポストフィルタ強度係数を大きくするに従い GV 尤度は変化し，ポストフィルタ強度係数を 0.85 に設定した場合に，“HMM+MS” と “HMM+GV+MS” の両方の尤度が自然音声に接近していることがわかる．対して図 3(c) から，生成パラメータ系列に対する MS 尤度は，自然音声のパラメータ系列に対する尤度よりも常に小さいことがわかる．以上の結果から，提案法の全尤度が自然音声の尤度よりも大きい係数は存在しないため，提案法と自然音声の GV 尤度が一致する係数 0.85 を，スペクトルパラメータのフィルタ強度係数として設定する．

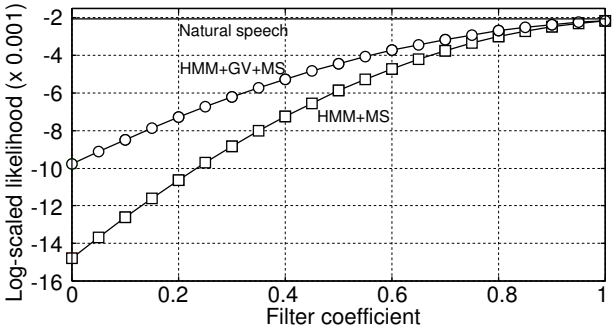
同様に，ポストフィルタ後の F_0 パラメータ系列に対する HMM・GV・MS 尤度をそれぞれ図 4(a)，図 4(b)，図 4(c) に示す．ポストフィルタ強度係数を大きくした時の各尤度の変化



(a) HMM 尤度



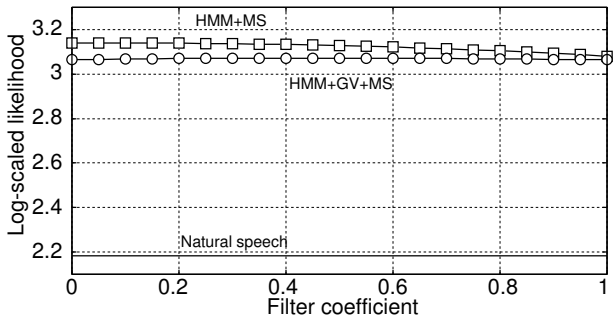
(b) GV 尤度



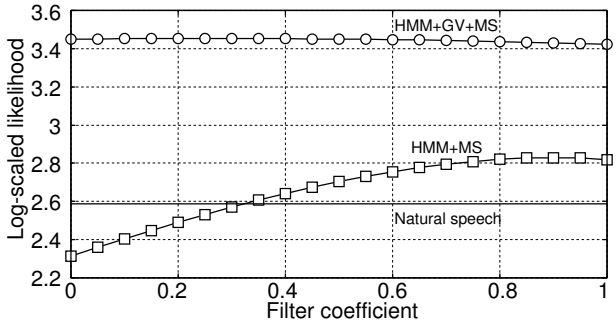
(c) MS 尤度

図 3 フィルタ後のスペクトルパラメータに対する尤度
Fig. 3 Likelihoods for filtered spectral parameter sequences.

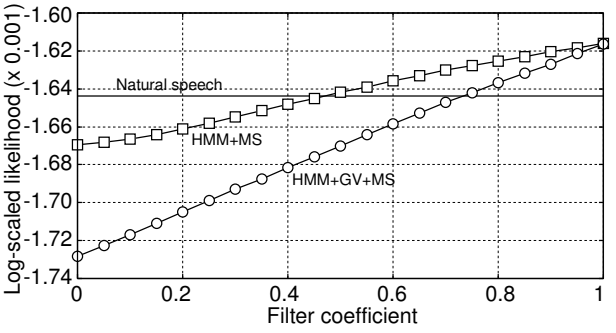
(注3)：MS を線形のパワースペクトルして扱った場合．
(注4)：いくつかのカットオフ周波数において MS の確率密度関数の学習精度を評価した結果，10Hz のカットオフ周波数が比較的良い性能となった．
(注5)：音声サンプルは，`http://isw3.naist.jp/~shinnosuke-t/sample_mspf.html` で受聴可能である．



(a) HMM 尤度



(b) GV 尤度



(c) MS 尤度

図 4　フィルタ後の F_0 系列に対する尤度

Fig. 4　Likelihoods for filtered F_0 contours.

の傾向は，スペクトルパラメータの場合と同じであることが確認出来る．しかしながら，フィルタ強度係数 0.75 を超えると，“HMM+MS” 及び “HMM+GV+MS” の全尤度が．“natural speech” の尤度を超えている事がわかる．また，係数を 1.0 に設定すると，MS 尤度が最大となることがわかる．以上の結果から，F_0 におけるポストフィルタ強度係数を 1.0 に設定する．

5.3　音質に関する主観評価結果

スペクトルパラメータに提案法を適用した時の，音質の主観評価結果を図 5 に示す．また，パラメータ系列及びスペクトログラムの例をそれぞれ，図 6 と図 8 に示す．“HMM” による生成パラメータ系列に対して提案法を適用することで，スコアが著しく上昇し，“HMM+GV” と同等の音質が得られることが分かる．また，“HMM+GV” におけるパラメータ系列に対する提案法の適用により，スコアは更に上昇することがわかる．以上の結果から，スペクトルパラメータに対する提案法の有効性が示された．

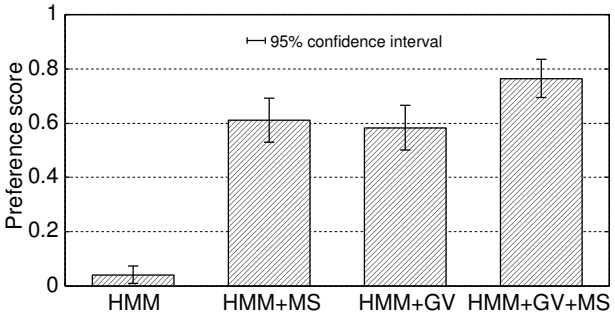


図 5　スペクトルパラメータにおける主観評価結果

Fig. 5　Preference score for spectral component.

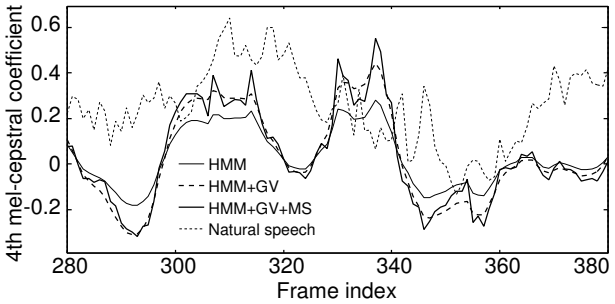


図 6　生成された 4 次のメルケプストラム系列

Fig. 6　Examples of natural and generated 4th mel-cepstral coefficient sequence.

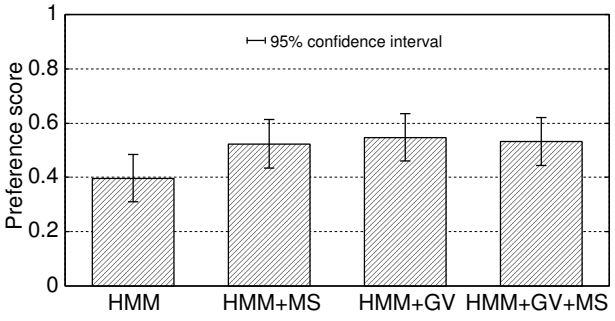


図 7　F_0 パラメータにおける主観評価結果

Fig. 7　Preference score for F_0 component.

同様に，F_0 パラメータにおける主観評価結果を図 7 に示す．“HMM+MS” 及び “HMM+GV” のスコアが “HMM” のスコアよりも高いことから，F_0 に対しても提案法の有効性が示された．自然音声及び合成音声の F_0 パターンは共に緩やかに遷移するため，F_0 に対する提案法による音質改善効果が，スペクトルパラメータの場合よりも小さくなったと思われる．

6. ま と め

本稿では，HMM 音声合成の音質改善を目的として，生成パラメータ系列の変調スペクトルを補償するポストフィルタを提案し，スペクトル及び F_0 における音質改善効果を実験的評価により示した．今後は，変調スペクトルを考慮したパラメータ生成法の検討を行う．

謝辞　本研究の一部は，JSPS 科研費 22680016 の助成を受け実施したものである．

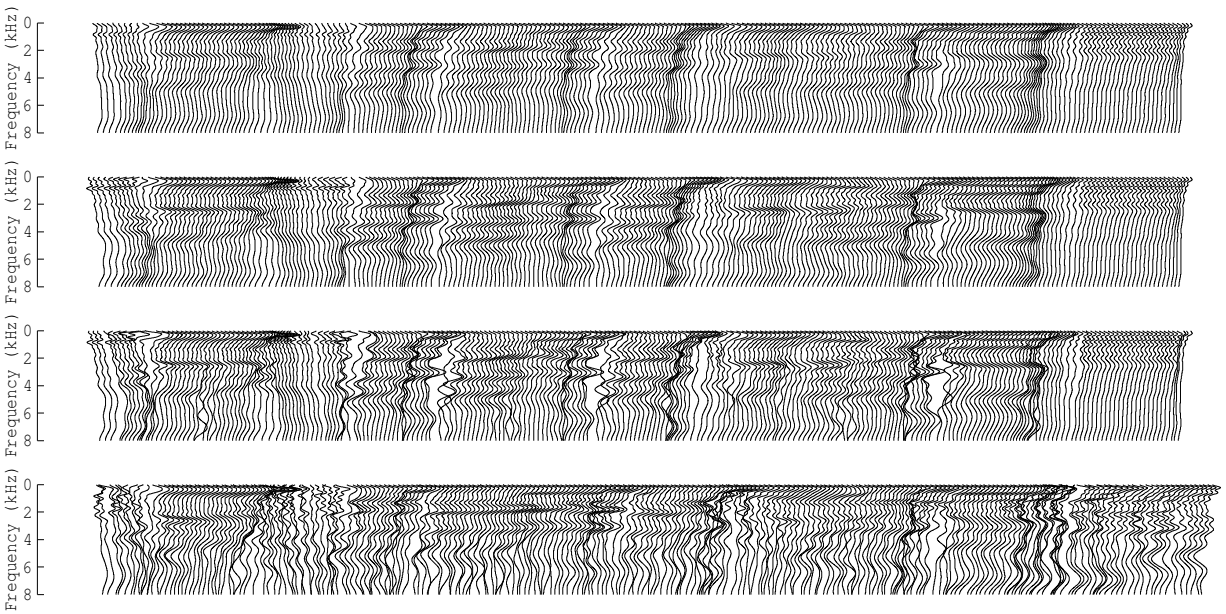


図 8　スペクトログラム（上から，“HMM”，“HMM+GV”，“HMM+GV+MS”，自然音声を表す）

Fig. 8　Spectrogram (representing “HMM”, “HMM+GV”, “HMM+GV+MS”, and natural speech from top down.)

文　　献

[1] H. Zen, K. Tokuda, and A. Black. Statistical parametric speech synthesis. *Speech Commun.*, Vol. 51, No. 11, pp. 1039–1064, 2009.

[2] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura. Speaker interpolation for HMM-based speech synthesis system. *J. Acoust. Soc. Jpn. (E)*, Vol. 21, No. 4, pp. 199–206, 2000.

[3] J. Yamagishi and T. Kobayashi. Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. *IEICE Trans., Inf. and Syst.*, Vol. E90-D, No. 2, pp. 533–543, 2007.

[4] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi. A style control technique for HMM-based expressive speech synthesis. *IEICE Trans., Inf. and Syst.*, Vol. E90-D, No. 9, pp. 1406–1413, 2007.

[5] S. King and V. Karaiskos. The blizzard challenge 2011. In *Proc. Blizzard Challenge workshop*, Turin, Italy, Sept. 2011.

[6] T. Toda and K. Tokuda. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Trans.*, Vol. E90-D, No. 5, pp. 816–824, 2007.

[7] R. Drullman, J .M. Festen, and R. Plomp. Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. of America*, Vol. 95, pp. 2670–2680, 1994.

[8] S. Thomas, S. Ganapathy, and H. Hermansky. Phoneme recgnition usng spectral envelop and modulation frequency features. In *Proc. ICASSP*, pp. 4453–4456, Taipei, Taiwan, April 2009.

[9] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. ICASSP*, pp. 1315–1318, Istanbul, Turkey, June 2000.

[10] L. Atlas and S. A.Shamma. Joint acoustic and modulation frequency. *EURASIP Journal on Applied Signal Processing*, Vol. 7, pp. 668–675, 2003.

[11] K. Yu and S. Young. Continuous F0 modeling for HMM based statistical parametric speech synthesis. *IEEE Trans. Audio, Speech and Language*, Vol. 19, No. 5, pp. 1071–1079, 2011.

[12] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. A hybrid approach to electrolaryngeal speech enhansement based on spectral subtraction and statistical voice conversion. In *Proc. INTERSPEECH*, pp. 3067–3071, Lyon, France, Sep. 2013.

[13] P. Taylor. *Text-To-Speech synthesis*. Cambridge Univ. Press, 2009.

[14] T. Toda, T. Muramatsu, and H. Banno. Implementation of conputationally efficient real-time voice conversion. In *Proc. INTERSPEECH*, Portland, Oregon, U.S., Sept. 2012.

[15] Y. Sagisaka, K. Takeda, M. Abe, S. Katagiri, T. Umeda, and H. Kuawhara. A large-scale Japanese speech database. In *ICSLP90*, pp. 1089–1092, Kobe, Japan, Nov. 1990.

[16] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Commun.*, Vol. 27, No. 3–4, pp. 187–207, 1999.

[17] H. Kawahara, Jo Estill, and O. Fujimura. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT” . In *MAVEBA 2001*, pp. 1–6, Firentze, Italy, Sept. 2001.

[18] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano. Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation. In *Proc. INTERSPEECH*, pp. 2266–2269, Pittsburgh, U.S.A., Sep. 2006.

[19] H. Zen, K. Tokuda, T. Kobayashi T. Masuko, and T. Kitamura. Hidden semi-markov model based speech synthesis system. *IEICE Trans., Inf. and Syst., E90-D*, No. 5, pp. 825–834, 2007.