Vol.2015-SLP-105 No.1
2015/2/27

統計的パラメトリック音声合成のための変調スペクトルを考慮した音声パラメータ生成アルゴリズム

高道 慎之介[1,2,a)]　戸田 智基[1]　ブラック アラン[2]　中村 哲[1]

概要：統計的パラメトリック音声合成において生成パラメータの過剰な平滑化は，合成音声の音質を著しく劣化させる．これに対して，系列内変動を考慮したパラメータ生成法の有効性が知られている．近年，我々は，系列内変動の拡張である変調スペクトルが，過剰な平滑化をより高精度に捉える特徴量であることを明らかにしている．本稿では，過剰な平滑化をより一層緩和する手法として，変調スペクトルを考慮した音声パラメータ生成法を提案する．実験的評価により，系列内変動を考慮したパラメータ生成法と比較して，提案法により合成音声の音質が向上する事を示す．

キーワード：統計的パラメトリック音声合成，HMM 音声合成，GMM 声質変換，過剰な平滑化，系列内変動，変調スペクトル，パラメータ生成

Speech Parameter Generation Algorithm Considering Modulation Spectrum for Statistical Parametric Speech Synthesis

Shinnosuke Takamichi[1,2,a)]　Tomoki Toda[1]　Alan W. Black[2]　Satoshi Nakamura[1]

Abstract: This paper proposes a novel speech parameter generation algorithm considering modulation spectrum for statistical parametric speech synthesis. The over-smoothing effect that is observed in generated speech parameter trajectories deteriorates the synthetic speech quality. A parameter generation algorithm considering Global Variance (GV) is known as an efficient approach to alleviating the over-smoothing effect but this effect still remains to be addressed. Recently, we have found the Modulation Spectrum (MS) which is regarded as an extension of the GV is capable of more sensitively detecting the over-smoothing effect than the GV. To further alleviate the over-smoothing effect, the proposed algorithm integrates the MS into the parameter generation. The experimental results demonstrate that the proposed parameter generation algorithm considering the MS yields significant improvements in synthetic speech quality compared to the conventional parameter generation algorithm considering the GV.

Keywords: statistical parametric speech synthesis, HMM-based text-to-speech synthesis, GMM-based voice conversion, over-smoothing, global variance, modulation spectrum, parameter generation

1. はじめに

入力情報から音声を生成する音声合成技術は，我々のコミュニケーション能力や身体機能を拡張する可能性を秘めている．本稿で取り扱う，テキストから音声を合成するテキスト音声合成（TTS: Text-To-Speech） [1] と言語情報を保持しつつ声質を変換する声質変換（VC: Voice Conversion） [2] は，音声合成技術の代表例である．1990 年代に提案され 2000 年代に急速に普及した統計的パラメトリック音声合成方式 [3], [4] は，合成器構築の容易さ及びその汎用性の高さ [5], [6] から，身障補助 [7], [8]，言語教育支援 [9], [10]，アミューズメント [11], [12] 等に向けて広く研究されている．

[1] 奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan.
[2] カーネギーメロン大学 言語技術研究所
Language Technologies Institute, Carnegie Mellon University (CMU), United States.
[a)] shinnosuke-t@is.naist.jp

一方で，統計的パラメトリック音声合成方式における合成音声の音質は，自然音声と比較して著しく劣化する傾向にある [13], [14]．その要因は，分析部・学習部・生成部の各々に存在する [15] が，特に生成部では，統計処理による音声パラメータ系列の過剰な平滑化が大きな要因である．パラメータ系列の系列内変動（GV: Global Variance）[4], [16] 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り，現在も広く応用されている [17], [18]．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する．

近年，我々は，GV の拡張形である変調スペクトル（MS: Modulation Spectrum）が，過剰な平滑化の定量化に一層効果的であること，また，生成パラメータ系列の MS を補償するポストフィルタが，音質改善に有効であることを明らかにしている [19], [20], [21]．しかしながらこの処理は，従来のパラメータ生成処理から独立したものであり，生成基準を無視してパラメータ系列を過剰に変形する．また，モデル適応 [6] や同時最適化学習 [17], [22] 等に代表される統計的パラメトリック音声合成方式の恩恵を受けられない．

本稿では，HMM（Hidden Markov Model）音声合成 [3] と GMM（Gaussian Mixture Model）声質変換 [4] において，MS を考慮した音声パラメータ生成法を提案する．提案法は，HMM 尤度又は GMM 尤度と，MS 尤度から成る目的関数を最大化するように，音声パラメータ系列を生成する．MS 尤度の使用により，生成パラメータの MS は自然音声パラメータの MS に近づくように補償される．実験的評価により，従来の GV を考慮したパラメータ生成法を超える音質改善効果が提案法により得られることを示す．

2. 統計的パラメトリック音声合成

HMM 音声合成と GMM 声質変換の学習部及び生成部について論述する．ただし，生成部以降の記述を簡単化するため，各合成方式間で対応するもの（例えば，HMM 音声合成における入力テキストのコンテキストと，GMM 声質変換における入力音声の特徴量）は変数名を共有する．

2.1 学習部

2.1.1 HMM 音声合成

HMM 音声合成では入力テキストのコンテキスト系列 $\boldsymbol{X}$ と出力音声の特徴量系列 $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_t^\top, \cdots, \boldsymbol{Y}_T^\top\right]^\top$ から，HMM のパラメータセット $\boldsymbol{\lambda}$ を学習する．出力音声の特徴量は，インデックス q の HMM 状態において，次式の出力確率密度関数でモデル化される．

$$P\left(\boldsymbol{Y}_t \middle| \boldsymbol{X}, q, \boldsymbol{\lambda}\right) = \mathcal{N}\left(\boldsymbol{Y}_t; \boldsymbol{\mu}_q^{(Y|X)}, \boldsymbol{\Sigma}_q^{(Y|X)}\right) \quad (1)$$

ただし，$\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top, \Delta\Delta\boldsymbol{y}_t^\top\right]^\top$ は，時刻 t における出力音声の静的特徴量 $\boldsymbol{y}_t = \left[y_t\left(1\right), \cdots, y_t\left(d\right), \cdots, y_t\left(D\right)\right]$ とその一次と二次の動的特徴量 $\Delta\boldsymbol{y}_t$，$\Delta\Delta\boldsymbol{y}_t$ の結合ベクトルを表す．T は総フレーム数，d は次元のインデックス，D は $\boldsymbol{y}_t$ の次元数を表す．HMM パラメータセットは Q 個の HMM 状態から成り，状態インデックス q の HMM 状態は，平均ベクトル $\boldsymbol{\mu}_q^{(Y|X)}$ と共分散行列 $\boldsymbol{\Sigma}_q^{(Y|X)}$ の正規分布 $\mathcal{N}\left(\cdot; \boldsymbol{\mu}_q^{(Y|X)}, \boldsymbol{\mu}_q^{(Y|X)}\right)$ を有する．F_0 パターンと状態継続長は，それぞれ多空間確率分布 HMM（Multi-Space probability Distribution HMM: MSD-HMM）[23] と状態継続長モデル [24] でモデル化される．

2.1.2 GMM 声質変換

GMM 声質変換では入力音声の特徴量系列 $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_t^\top, \cdots, \boldsymbol{X}_T^\top\right]^\top$ と出力音声の特徴量系列 $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_t^\top, \cdots, \boldsymbol{Y}_T^\top\right]^\top$ から，GMM のパラメータセット $\boldsymbol{\lambda}$ を学習する．入出力音声の特徴量は，次式の結合確率密度関数でモデル化される．

$$P\left(\left[\begin{array}{c}\boldsymbol{X}_t\\ \boldsymbol{Y}_t\end{array}\right]\middle|\boldsymbol{\lambda}\right) = \sum_{q=1}^{Q}\alpha_q\mathcal{N}\left(\left[\begin{array}{c}\boldsymbol{X}_t\\ \boldsymbol{Y}_t\end{array}\right]; \boldsymbol{\mu}_q^{(X,Y)}, \boldsymbol{\Sigma}_q^{(X,Y)}\right) \quad (2)$$

$$\boldsymbol{\mu}_q^{(X,Y)} = \left[\begin{array}{c}\boldsymbol{\mu}_q^{(X)}\\ \boldsymbol{\mu}_q^{(Y)}\end{array}\right], \boldsymbol{\Sigma}_q^{(X,Y)} = \left[\begin{array}{cc}\boldsymbol{\Sigma}_q^{(XX)} & \boldsymbol{\Sigma}_q^{(XY)}\\ \boldsymbol{\Sigma}_q^{(YX)} & \boldsymbol{\Sigma}_q^{(YY)}\end{array}\right] \quad (3)$$

ただし，$\boldsymbol{X}_t$ と $\boldsymbol{Y}_t$ はそれぞれ，時刻 t における入出力音声の特徴量であり，$\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$ と表される．GMM パラメータセットは Q 個の混合要素を有し，インデックス q の混合要素は，混合重み α_q，平均ベクトル $\boldsymbol{\mu}_q^{(X,Y)}$ と共分散行列 $\boldsymbol{\Sigma}_q^{(X,Y)}$ から成る正規分布で表される．$\boldsymbol{\mu}_q^{(X,Y)}$ は，入出力音声の特徴量の平均ベクトルである $\boldsymbol{\mu}_q^{(X)}$ と $\boldsymbol{\mu}_q^{(Y)}$ から成る．また，$\boldsymbol{\Sigma}_q^{(X,Y)}$ は，入出力音声の特徴量の共分散行列である $\boldsymbol{\Sigma}_q^{(XX)}$ 及び $\boldsymbol{\Sigma}_q^{(YY)}$ と，相互共分散行列である $\boldsymbol{\Sigma}_q^{(YX)}$ 及び $\boldsymbol{\Sigma}_q^{(XY)}$ から成る．

2.2 生成部

生成時にはまず，入力コンテキスト系列又は入力音声パラメータ系列 $\boldsymbol{X}$ から，対応する HMM 又は GMM 系列を構築する．次に，次式に示すように，HMM 音声合成における状態継続長モデルの尤度最大化 [25]，又は，GMM 声質変換における周辺化 GMM の事後確率最大化 [4] により，準最適な HMM の状態系列又は GMM の分布系列 $\hat{\boldsymbol{q}} = [\hat{q}_1, \cdots, \hat{q}_t, \cdots, \hat{q}_T]$ を決定する．

$$\hat{\boldsymbol{q}} = \underset{\boldsymbol{q}}{\operatorname{argmax}}\, P\left(\boldsymbol{q}\middle|\boldsymbol{X}, \boldsymbol{\lambda}\right) \quad (4)$$

ただし，$\hat{q}_t$ は時刻 t における準最適な状態又は分布 $\hat{q}$ である．合成音声のパラメータ系列は，静的・動的特徴量間の制約条件下で，次式の出力確率密度関数 $L\left(\boldsymbol{y}\right)$ を最大化することで得られる．

$$L\left(\boldsymbol{y}\right) = P\left(\boldsymbol{W}\boldsymbol{y}\middle|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}\right) = \mathcal{N}\left(\boldsymbol{W}\boldsymbol{y}; \boldsymbol{\mu}_{\hat{\boldsymbol{q}}}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{q}}}\right) \quad (5)$$

ただし，$\boldsymbol{y} = \left[\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top\right]^\top$ は音声パラメータ系列，$\boldsymbol{W}$ 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 [25] である．出力確率密度関数は正規分布として表現され，その平均ベクトル $\boldsymbol{\mu}_{\hat{\boldsymbol{q}}} = \left[\boldsymbol{\mu}_{\hat{q}_1,1}^\top, \cdots, \boldsymbol{\mu}_{\hat{q}_t,t}^\top, \cdots, \boldsymbol{\mu}_{\hat{q}_T,T}^\top\right]^\top$ と共分散行列 $\boldsymbol{\Sigma}_{\hat{\boldsymbol{q}}} = \mathrm{diag}\left[\boldsymbol{\Sigma}_{\hat{q}_1}, \cdots, \boldsymbol{\Sigma}_{\hat{q}_t}, \cdots, \boldsymbol{\Sigma}_{\hat{q}_T}\right]$ の各要素はそれぞれ，次式で与えられる．

$$\boldsymbol{\mu}_{\hat{q},t} = \begin{cases} \boldsymbol{\mu}_{\hat{q}}^{(Y|X)} & \text{(HMM)} \\ \boldsymbol{A}_{\hat{q}}\boldsymbol{X}_t + \boldsymbol{b}_{\hat{q}} & \text{(GMM)} \end{cases} \tag{6}$$

$$\boldsymbol{\Sigma}_{\hat{q},t} = \begin{cases} \boldsymbol{\Sigma}_{\hat{q}}^{(Y|X)} & \text{(HMM)} \\ \boldsymbol{\Sigma}_{\hat{q}}^{(YY)} - \boldsymbol{A}_{\hat{q}}\boldsymbol{\Sigma}_{\hat{q}}^{(XX)}\boldsymbol{A}_{\hat{q}}^\top & \text{(GMM)} \end{cases} \tag{7}$$

$$\boldsymbol{A}_{\hat{q}} = \boldsymbol{\Sigma}_{\hat{q}}^{(YX)}\boldsymbol{\Sigma}_{\hat{q}}^{(XX)^{-1}},\ \boldsymbol{b}_{\hat{q}} = \boldsymbol{\mu}_{\hat{q}}^{(Y)} - \boldsymbol{A}_{\hat{q}}\boldsymbol{\mu}_{\hat{q}}^{(X)} \tag{8}$$

式 (5) で生成されるパラメータ系列は過剰な平滑化の影響を強く受けるが，GV を考慮することでその影響を緩和できる．パラメータ系列の GV $\boldsymbol{v}(\boldsymbol{y}) = \left[v(1), \cdots, v(d), \cdots, v(D)\right]^\top$ の d 番目の要素 $v(d)$ は次式で定義される．

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t(d) - \bar{y}(d)\right)^2 \tag{9}$$

ここで，$\bar{y}(d) = (1/T)\sum_{t=1}^{T} y_t(d)$ である．GV を考慮したパラメータ生成法 [4], [16] では，式 (5) に GV 尤度を組み込んだ次式の目的関数 $L_{\mathrm{v}}(\boldsymbol{y})$ の最大化により，パラメータ系列を生成する．

$$L_{\mathrm{v}}(\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{W}\boldsymbol{y}; \boldsymbol{\mu}_{\hat{\boldsymbol{q}}}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{q}}}\right)\mathcal{N}\left(\boldsymbol{v}(\boldsymbol{y}); \boldsymbol{\mu}_{\mathrm{v}}, \boldsymbol{\Sigma}_{\mathrm{v}}\right)^{w_{\mathrm{v}}} \tag{10}$$

ただし，$\boldsymbol{\mu}_{\mathrm{v}}$ と $\boldsymbol{\Sigma}_{\mathrm{v}}$ はそれぞれ，GV の平均ベクトルと共分散行列であり，学習データから推定される．w_{v} は GV 尤度の重みを表し，本稿では，$\boldsymbol{W}\boldsymbol{y}$ の次元数を $\boldsymbol{v}(\boldsymbol{y})$ の次元数で割った値に設定する．

3. 提案するパラメータ生成法

3.1 変調スペクトル（MS）

MS は，GV の拡張であり，パラメータ系列のパワースペクトルとして定義される [19]．パラメータ系列 $\boldsymbol{y}$ の MS $\boldsymbol{s}(\boldsymbol{y}) = \left[\boldsymbol{s}(1), \cdots, \boldsymbol{s}(d), \cdots, \boldsymbol{s}(D)\right]^\top$ は次式で表される．

$$\boldsymbol{s}(d) = \left[s_d(0), \cdots, s_d(f), \cdots, s_d(D_{\mathrm{s}} - 1)\right]^\top \tag{11}$$

$$s_d(f) = R_{d,f}^2 + I_{d,f}^2 \tag{12}$$

$$= \left(\sum_{t=1}^{T} y_t(d)\cos kt\right)^2 + \left(\sum_{t=1}^{T} y_t(d)\sin kt\right)^2 \tag{13}$$

ただし，$2D_{\mathrm{s}}$ は離散フーリエ変換（DFT: Discrete Fourier Transform）のタップ長，f は変調周波数のインデックス，$k = -\pi f/D_{\mathrm{s}}$ は変調周波数を表す．また，MS の確率密度関数を，平均ベクトル $\boldsymbol{\mu}_{\mathrm{s}}$ と共分散行列 $\boldsymbol{\Sigma}_{\mathrm{s}}$ から成る正規分布 $\mathcal{N}\left(\boldsymbol{s}(\boldsymbol{y}); \boldsymbol{\mu}_{\mathrm{s}}, \boldsymbol{\Sigma}_{\mathrm{s}}\right)$ で表し，更に，$\boldsymbol{\Sigma}_{\mathrm{s}}^{-1} = \left[\boldsymbol{p}_{\mathrm{s}}^{(1)}, \cdots, \boldsymbol{p}_{\mathrm{s}}^{(d)}, \cdots, \boldsymbol{p}_{\mathrm{s}}^{(D)}\right]$ とする．$\boldsymbol{p}_{\mathrm{s}}^{(d)}$ は，DD_{s}-by-D_{s} の行列である．本稿では発話毎に MS を計算し，平均ベクトルと共分散行列を推定する．

3.2 変調スペクトルを考慮したパラメータ生成法

提案するパラメータ生成法では，次式の目的関数 $L_{\mathrm{s}}(\boldsymbol{y})$ を最大化する．

$$L_{\mathrm{s}}(\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{W}\boldsymbol{y}; \boldsymbol{\mu}_{\hat{\boldsymbol{q}}}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{q}}}\right)\mathcal{N}\left(\boldsymbol{s}(\boldsymbol{y}); \boldsymbol{\mu}_{\mathrm{s}}, \boldsymbol{\Sigma}_{\mathrm{s}}\right)^{\omega_{\mathrm{s}}} \tag{14}$$

ここで ω_{s} は MS 尤度の重みであり，本稿では，$\boldsymbol{W}\boldsymbol{y}$ の次元数を $\boldsymbol{s}(\boldsymbol{y})$ の次元数で割った値に設定する．$L_{\mathrm{s}}(\boldsymbol{y})$ の最大化問題を解析的に解くのは困難であるため，最急降下法 $\hat{\boldsymbol{y}}^{(i+1)} = \hat{\boldsymbol{y}}^{(i)} + \alpha\left(\partial \log L_{\mathrm{s}}/\partial \boldsymbol{y}\right)\big|_{\boldsymbol{y}=\hat{\boldsymbol{y}}^{(i)}}$ により，反復的にパラメータ系列を生成する．i は反復インデックス，α は学習係数である．ここで，一次微分は次式で与えられる．

$$\frac{\partial \log L_{\mathrm{s}}}{\partial \boldsymbol{y}} = -\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{W}\boldsymbol{y} + \boldsymbol{W}^\top\boldsymbol{\Sigma}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{\mu}_{\hat{\boldsymbol{q}}} + \omega_{\mathrm{s}}\left[\boldsymbol{s}_1'^\top, \cdots, \boldsymbol{s}_t'^\top, \cdots, \boldsymbol{s}_T'^\top\right]^\top \tag{15}$$

$$\boldsymbol{s}_t' = \left[s_t(1), \cdots, s_t(D)\right]^\top \tag{16}$$

$$s_t(d) = \left(\boldsymbol{s}(\boldsymbol{y}) - \boldsymbol{\mu}_{\mathrm{s}}\right)^\top \boldsymbol{p}_{\mathrm{s}}^{(d)}\boldsymbol{f}_t(d) \tag{17}$$

$$\boldsymbol{f}_t(d) = \left[f_{t,d}(0), \cdots, f_{t,d}(D_{\mathrm{s}} - 1)\right]^\top \tag{18}$$

$$f_{t,d}(f) = -2\left(R_{d,f}\cos kt + I_{d,f}\sin kt\right) \tag{19}$$

3.2.1 初期化法

初期化時には [4], [16] と同様に，HMM 尤度又は GMM 尤度を最大化してパラメータ系列を生成し，他の尤度を最大化するようにそのパラメータ系列を変形する．本稿では，MS 尤度を上昇させるように変形するため，MS を補償するポストフィルタ [19] を採用し，次式のように MS $s_d(f)$ を変形する．

$$s_d(f)' = \frac{\sigma_{d,f}}{\sigma_{d,f}'}\left(s_d(f) - \mu_{d,f}'\right) + \mu_{d,f} \tag{20}$$

ただし，$\mu_{d,f}$ と $\sigma_{d,f}$ はそれぞれ，$s_d(f)$ の平均と標準偏差である．$\mu_{d,f}'$ と $\sigma_{d,f}'$ は，生成パラメータ系列の MS の平均及び標準偏差であり，学習データに対する生成パラメータから事前推定される．初期パラメータ系列 $\hat{\boldsymbol{y}}^{(0)}$ は，変形された MS と変形前のパラメータ系列の位相から計算される．過強調を緩和するため，反復後のパラメータ系列に対して 50Hz カットオフのローパスフィルタを施す [*1]．

3.2.2 F_0 パターン生成への適用

HMM 音声合成においては，MSD-HMM を用いた F_0 パターン生成に対しても，提案法を適用する．有声／無声境界における不連続な遷移を認めるという条件の下，連結された有声フレームにおける F_0 を生成する．本稿では，連結された有声フレームの F_0 系列から MS を計算する．た

[*1] 予備主観評価の結果，50Hz 以上の MS を除去した音声と分析再合成音の間に音質の差はないことが明らかになっている．

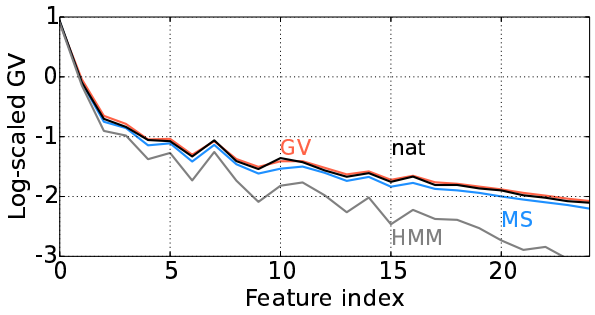

図 1 メルケプストラム系列の GV の例

Fig. 1 An example of the GV of the mel-cepstral coefficient sequences.

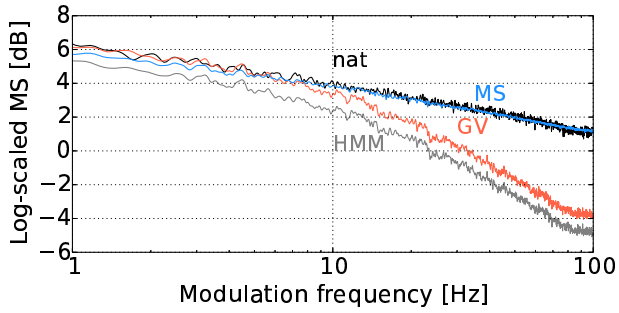

図 2 9 次のメルケプストラム系列の MS の例

Fig. 2 An example of the MS of the 9-th mel-cepstral coefficient sequences.

だし [19] と同様に，式 (13) の $y_t(d)$ を $y_t(d) - \bar{y}(d)$ に変形し，系列平均を 0 とした MS を計算する．なお，上述の初期化法は，有声／無声境界において F_0 パターンを過剰に変形するため，初期 F_0 パターンの生成には，従来の GV に基づく初期化法を用いる．

3.3 考察

パラメータ生成基準に MS を組み込んだ提案法により，ポストフィルタ処理とは異なり，HMM 尤度および GMM 尤度を考慮しながら生成パラメータ系列の MS を補償することが可能となる．また，MS は GV を内包するため，提案法は暗黙的に GV も補償する．Fig. 1 と Fig. 2 にそれぞれ，自然音声パラメータ（“nat”）と生成パラメータの GV と MS の例を示す．“HMM”，“GV”，“MS” はそれぞれ，式 (5)，式 (10)，式 (14) で生成したパラメータ系列の GV 及び MS である．提案法により，MS のみならず GV も補償されていることが確認できる．一方で “GV” は，明らかに “nat” と異なる MS を有する．これは，GV が，MS を変調周波数で平均化した値のみを表現しているためである．

MS を補償するポストフィルタは，従来の生成基準を無視した変形を行うため，過剰に強調された音声を生成する．一方で提案法は，従来の生成基準と MS 尤度を同時に最大化してパラメータ系列を生成する．また，提案法は数学的に定式化されているため，コンテキスト依存モデリング，適応，同時最適化による学習等が容易である．

4. 実験的評価

4.1 実験条件

CMU ARCTIC 音声データベース [26] から学習データと評価データを選択する．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 [27] 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 [28] を使用する．MS を計算する DFT のタップ数は 8192 点とする．GMM 声質変換では対数 F_0 を線形変換する．詳細な実験条件は表 1 に示す．

以下の音声パラメータ系列を比較する．

HMM/GMM： 式 (5) で生成したパラメータ系列
GV： 式 (10) で生成したパラメータ系列
MS： 式 (14) で生成したパラメータ系列 **（提案法）**
nat： 自然音声パラメータ系列

提案法の影響を客観的に評価するため，パラメータ系列に対する HMM/GMM 尤度，GV 尤度，MS 尤度を計算する．また，考察のために対数 MS $\log s_d(f)$ の尤度も計算する．HMM/GMM 尤度では T で正規化した値を，MS 尤度と対数 MS の尤度では D_s で正規化した値を示す．次に，提案法による音質改善効果を確認するため，“GV” と “MS” を比較する．評価として，HMM 音声合成と GMM 声質変換における音質に関するプリファレンス AB テスト，及び，GMM 声質変換における話者性に関する XAB テストを実施する．XAB テストのリファンレス音声は，ターゲット話者の分析再合成音である．被験者数は，HMM 音声合成の評価では 8 人，GMM 声質変換の評価では 6 人である．非周期成分における提案法の知覚的影響は小さいため，非周期成分には “GV” を使用する．

4.2 客観評価結果

Fig. 3 から Fig. 5 にそれぞれ，HMM 音声合成におけるスペクトルパラメータ系列と F_0 系列，また，GMM 声質変換におけるスペクトルパラメータ系列に対する，HMM/GMM 尤度，GV 尤度，MS 尤度を示す．

HMM/GMM 尤度と GV 尤度： Fig. 3 から，提案法で生成した各パラメータ系列に対する HMM/GMM 尤度（“MS”）は，従来の “HMM/GMM” と “GV” よりも低く

表 1 実験条件
Table 1 Experimental conditions.

	HMM 音声合成	GMM 声質変換
学習データ	593 文	50 文
評価データ	100 文	100 文
話者	男性及び女性	男性及び女性
音響モデル	5 状態 HSMM	64 混合 GMM
提案法の適用	スペクトル，F_0	スペクトル

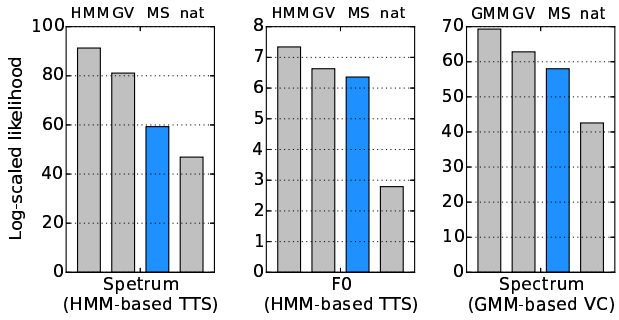

図 3 各パラメータ系列の HMM/GMM 尤度

Fig. 3 HMM/GMM likelihoods for parameter sequences.

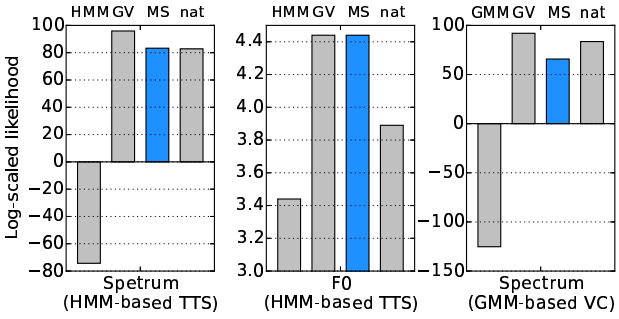

図 4 各パラメータ系列の GV 尤度

Fig. 4 GV likelihoods for parameter sequences.

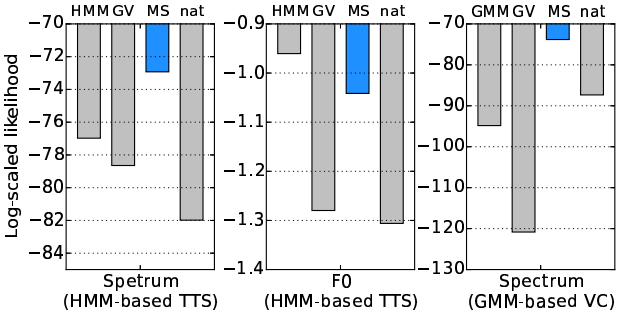

図 5 各パラメータ系列の MS 尤度

Fig. 5 MS likelihoods for parameter sequences.

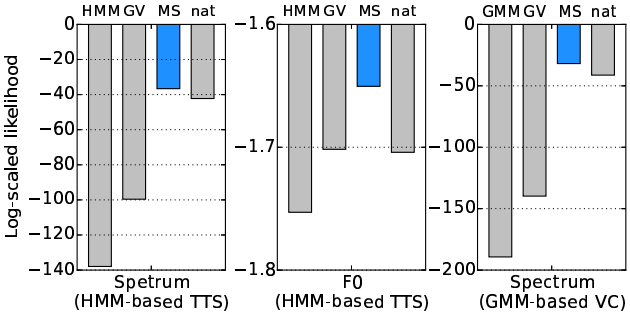

図 6 各パラメータ系列の対数 MS の尤度

Fig. 6 Log-MS likelihoods for parameter sequences.

なっているが，依然として “nat” よりも高い値を保っている．また Fig. 4 から，“MS” は，“GV” と同様に GV 尤度

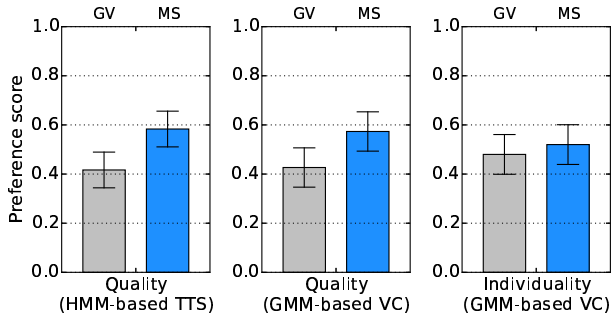

図 7 音質に関する主観評価結果（エラーバーは 95% 信頼区間）

Fig. 7 Results of the subjective evaluations with 95% confidence interval on speech quality.

を大幅に改善していることが分かる．以上の結果から，提案法は従来基準による効果を保持しているといえる．

MS 尤度と対数 MS の尤度： Fig. 5 から，スペクトルパラメータにおける “MS” は “HMM” と “GV” よりも高い尤度となっており，提案法の効果が確認できる．一方で F_0 の場合，同じ初期化法だが異なる目的関数を持つ “GV” と “MS” で，“MS” の方が高い尤度となっているため，提案法の反復的最大化により MS 尤度が改善されることが分かる．しかしながら，スペクトルと F_0 の両方において，“HMM” の MS 尤度は “nat” の尤度よりも高い．この現象を考察するために，Fig. 6 に対数 MS の尤度を示す．こちらの結果では “nat” の尤度が高いため，妥当な結果であるといえる．これらの結果から，MS の確率密度関数は，対数 MS 領域における正規分布でより良くモデル化されることが分かる．ただし，提案法によるパラメータ生成処理における音質改善効果という面では，MS のモデル化と対数 MS のモデル化の間に大きな差はないことを，予備実験により確認している．

4.3 主観評価結果

各主観評価結果を Fig. 7 に示す．HMM 音声合成と GMM 声質変換の音質評価において提案法のスコアが上昇しているため，提案法の音質改善効果が確認できる．一方で，GMM 声質変換の話者性のスコアには有意な差が見られない．同様の傾向が [20] で得られている事から，変調周波数成分において，個人性知覚に関わる音響的手掛かりの内，HMM や GMM，GV では表現しきれないものは小さいと考えられる．

5. まとめ

本稿では，統計的パラメトリック音声合成の音質改善を目的として，変調スペクトルを考慮したパラメータ生成法を提案し，実験的評価により提案法の音質改善効果を確認した．今後は，トラジェクトリ学習 [29]，連続 F_0 モデル [30] への導入，及びポストフィルタ [19] との比較を行う．

謝辞 本研究の一部は，JSPS 特別研究員奨励費 26 · 10354，JSPS 科研費 26280060，及び，頭脳循環を加速する若手研究者戦略的海外派遣プログラムの助成を受け実施した．

参考文献

[1] Y. Sagisaka. Speech synthesis by rule using an optimal selection of non-uniform synthesis units. In *Proc. ICASSP*, pp. 679–682, New York, U.S.A, Apr. 1988.

[2] Y. Stylianou, O. Cappe, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 2, pp. 131–142, Mar. 1988.

[3] K. Tokuda, Y Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura. Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, Vol. 101, No. 5, pp. 1234–1252, 2013.

[4] T. Toda, A. W. Black, and K. Tokuda. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, No. 8, pp. 2222–2235, 2007.

[5] K. Oura, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda. Tying covariance matrices to reduce the footprint of HMM-based speech synthesis systems. In *Proc. INTERSPEECH*, pp. 1759–1762, Brighton, U. K., 2009.

[6] 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.

[7] J. Yamagishi, C. Veaux, S. King, and S. Renals. Speech synthesis technologies for individuals with vocal diabilities: Voice banking and reconstruction. *Acoust. Sci. technol.*, Vol. 33, pp. 1–5, 2012.

[8] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. A hybrid approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation. *IEICE Trans. on Inf. and Syst.*, Vol. E97-D, No. 6, pp. 1429–1437, Jun. 2014.

[9] S. Aryal and R. G.-Osuna. Can voice conversion be used to reduce non-native accents? In *Proc. ICASSP*, pp. 7929–7933, Florence, Italy, May 2014.

[10] 高道慎之介, 大島悠司, 戸田智基, Neubig Graham, Sakti Sakriani, 中村哲. 日本人英語のための音声合成技術を用いた英語学習支援の検討. 教育システム情報学会研究報告, Vol. 29, No. 5, pp. 111–116, Jan. 2015.

[11] K. Shirota, K. Nakamura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda. Integration of speaker and pitch adaptive training for HMM-based singing voice synthesis. In *Proc. ICASSP*, pp. 2578–2582, Florence, Italy, May 2014.

[12] K. Kobayashi, T. Toda, H. Doi, T. Nakano, M. Goto, G. Neubig, S. Sakti, and S. Nakamura. Voice timbre control based on perceived age in singing voice conversion. *IEICE Trans. on Inf. and Syst.*, Vol. E97-D, No. 6, pp. 1419–1428, Jun. 2014.

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

[14] Y. Stylianou. Voice transformation: A survey. In *Proc. ICASSP*, pp. 3585–3588, Taipei, Taiwan, Apr. 2009.

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

[16] 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.

[17] H. Hwang, Y. Tsao, H. Wang, Y. Wang, and S. Chen. Incorporating global variance in the training phase of GMM-based voice conversion. In *Proc. APSIPA*, pp. 1–6, Kaohsiung, Taiwan, Oct. 2013.

[18] S. Takamichi, T. Toda, Y. Shiga, S. Sakti, G. Neubig, and S. Nakamura. Parameter generation methods with rich context models for high-quality and flexible text-to-speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, Vol. 8, No. 2, pp. 239–250, May 2014.

[19] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. A postfilter to modify the modulation spectrum in HMM-based speech synthesis. In *Proc. ICASSP*, pp. 290–294, Florence, Italy, May 2014.

[20] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura. Modulation spectrum-based post-filter for gmm-based voice conversion. In *Proc. APSIPA ASC*, Siem Reap, Cambodia, Dec. 2014.

[21] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura. Modified modulation spectrum-based post-filter for HMM-based speech synthesis. In *Proc. GlobalSIP*, pp. 710–714, Atlanta, United States, Decc. 2014.

[22] T. Toda and S. Young. Trajectory training considering global variance for HMM-based speech synthesis. In *Proc. ICASSP*, pp. 4025–4028, Taipei, Taiwan, Aug. 2009.

[23] K. Tokuda, T. Masuko, B. Miyazaki, and T. Kobayashi. Multi-space probability distribution HMM. *IEICE Trans., Inf. and Syst.*, Vol. E85-D, No. 3, pp. 455–464, 2002.

[24] 吉村貴克, 徳田恵一, 益子貴史, 小林隆夫, 北村正. Hmm に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化. 電子情報通信学会論文誌, Vol. J83-D-II, No. 5, pp. 2099–2107, Nov. 2000.

[25] 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.

[26] J. Kominek and A. W Black. The CMU ARCTIC speech databases for speech synthesis research. In *Tech. Rep. CMU-LTI-03-177*, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, U.S.A, 2003.

[27] 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.

[28] 大谷大和, 戸田智基, 猿渡洋, 鹿野清宏. STRAIGHT 混合励振源を用いた混合正規分布モデルに基づく最尤声質変換法. 電子情報通信学会論文誌, Vol. J91-D, No. 4, pp. 1082–1091, Apr. 2008.

[29] H. Zen, Y. Nankaku, and K. Tokuda. Continuous stochastic feature mapping based on trajectory HMMs. *IEEE Trans.*, Vol. 19, pp. 417–430, Jan. 2011.

[30] 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.