統計的パラメトリック音声合成における変調スペクトルを考慮したパラメータ生成法*

○ 高道 慎之介（奈良先端大/CMU）, 戸田 智基（奈良先端大）,
Alan W. Black（CMU）, 中村 哲 (奈良先端大)

1 はじめに

統計的パラメトリック音声合成における音質劣化の一因は，生成パラメータ系列の過剰な平滑化である．我々はこれまでに，過剰な平滑化を定量化する特徴量として，系列内変動（GV： Global Variance）[1, 2] の拡張である変調スペクトル（MS: Modulation Spectrum）に着目し，ポストフィルタ処理においてその有効性を示している [3]．本稿では，HMM 音声合成 [4] と GMM 声質変換 [2] において，変調スペクトルを考慮したパラメータ生成法を提案する．実験的評価により，提案法の音質改善効果を示す．

2 従来のパラメータ生成法

HMM 音声合成では入力テキストのコンテキストと出力音声のパラメータ系列から，GMM 声質変換では入出力音声のパラメータから，各統計モデル $\boldsymbol{\lambda}$ を学習する．生成時には，入力コンテキスト又は入力音声パラメータ系列 $\boldsymbol{X}$ から次式の出力確率密度関数 $L(\boldsymbol{y})$ を計算し，静的・動的特徴量間の制約条件下の尤度最大化により，出力パラメータ系列 $\boldsymbol{y}$ を生成する．

$$L(\boldsymbol{y}) = P(\boldsymbol{W}\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{q}, \boldsymbol{\lambda}) = \mathcal{N}\left(\boldsymbol{W}\boldsymbol{y}; \boldsymbol{\mu}_{\boldsymbol{q}}, \boldsymbol{\Sigma}_{\boldsymbol{q}}\right) \quad (1)$$

ただし，$\boldsymbol{y} = \left[\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_T^\top\right]^\top$ は T フレームの音声パラメータ系列，$\boldsymbol{y}_t = [y_t(1), \cdots, y_t(D)]^\top$ は時刻 t における D 次元の音声パラメータ，$\boldsymbol{W}$ 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 [5] である．出力確率密度関数は，準最適な HMM の状態系列又は GMM の分布系列 $\boldsymbol{q}$ による近似を行うことで，平均ベクトル $\boldsymbol{\mu}_{\boldsymbol{q}}$ と共分散行列 $\boldsymbol{\Sigma}_{\boldsymbol{q}}$ から成る正規分布 $\mathcal{N}\left(\cdot; \boldsymbol{\mu}_{\boldsymbol{q}}, \boldsymbol{\Sigma}_{\boldsymbol{q}}\right)$ で与えられる [5].

パラメータ系列の GV $\boldsymbol{v}(\boldsymbol{y}) = [v(1), \cdots, v(D)]^\top$ の d 番目の要素 $v(d)$ は次式で定義される．

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

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アルゴリズム [1, 2] では，式 (1) 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．

$$L_{\mathrm{v}}(\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{W}\boldsymbol{y}; \boldsymbol{\mu}_{\boldsymbol{q}}, \boldsymbol{\Sigma}_{\boldsymbol{q}}\right)\mathcal{N}\left(\boldsymbol{v}(\boldsymbol{y}); \boldsymbol{\mu}_{\mathrm{v}}, \boldsymbol{\Sigma}_{\mathrm{v}}\right)^{w_{\mathrm{v}}} \quad (3)$$

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 尤度の重みを表す．

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 [3]．パラメータ系列 $\boldsymbol{y}$ の MS $\boldsymbol{s}(\boldsymbol{y}) = [\boldsymbol{s}(1), \cdots, \boldsymbol{s}(D)]^\top$ を次式で定義する．

$$\boldsymbol{s}(d) = \left[s_d(0), \cdots, s_d(D_{\mathrm{s}} - 1)\right]^\top \quad (4)$$

$$s_d(f) = R_{d,f}^2 + I_{d,f}^2 \quad (5)$$

$$= \left(\sum_{t=1}^{T} y_t(d)\cos kt\right)^2 + \left(\sum_{t=1}^{T} y_t(d)\sin kt\right)^2 \quad (6)$$

ただし，$2D_{\mathrm{s}}$ は離散フーリエ変換（DFT）のタップ長，$k = -\pi f/D_{\mathrm{s}}$ 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，$\boldsymbol{\Sigma}_{\mathrm{s}}^{-1} = \left[\boldsymbol{p}_{\mathrm{s}}^{(1)}, \cdots, \boldsymbol{p}_{\mathrm{s}}^{(D)}\right]$ とする．$\boldsymbol{p}_{\mathrm{s}}^{(d)}$ は，$\boldsymbol{s}(d)$ の精度に対応する行列である．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．

$$L_{\mathrm{s}}(\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{W}\boldsymbol{y}; \boldsymbol{\mu}_{\boldsymbol{q}}, \boldsymbol{\Sigma}_{\boldsymbol{q}}\right)\mathcal{N}\left(\boldsymbol{s}(\boldsymbol{y}); \boldsymbol{\mu}_{\mathrm{s}}, \boldsymbol{\Sigma}_{\mathrm{s}}\right)^{\omega_{\mathrm{s}}} \quad (7)$$

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 $\hat{\boldsymbol{y}}^{(i+1)} = \hat{\boldsymbol{y}}^{(i)} + \alpha\left(\partial \log L_{\mathrm{s}}/\partial \boldsymbol{y}\right)|_{\boldsymbol{y}=\hat{\boldsymbol{y}}^{(i)}}$ 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．

$$\frac{\partial \log L_{\mathrm{s}}}{\partial \boldsymbol{y}} = \omega_{\mathrm{s}}\left(-\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{W}\boldsymbol{y} + \boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{\mu}_{\boldsymbol{q}}\right) + \left[\boldsymbol{s}_1'^\top, \cdots, \boldsymbol{s}_t'^\top, \cdots, \boldsymbol{s}_T'^\top\right]^\top \quad (8)$$

$$\boldsymbol{s}_t' = [s_t(1), \cdots, s_t(D)]^\top \quad (9)$$

$$s_t(d) = \left(\boldsymbol{s}(\boldsymbol{y}) - \boldsymbol{\mu}_{\mathrm{s}}\right)^\top \boldsymbol{p}_{\mathrm{s}}^{(d)}\boldsymbol{f}_t(d) \quad (10)$$

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

$$f_{t,d}(f) = -2\left(R_{d,f}\cos kt + I_{d,f}\sin kt\right) \quad (12)$$

初期化時には，式 (1) の最大化で生成したパラメータ系列の MS $s_d(f)$ を，次式のように変形する．

$$s_d(f)' = \frac{\sigma_{d,f}}{\sigma'_{d,f}}\left(s_d(f) - \mu'_{d,f}\right) + \mu_{d,f} \quad (13)$$

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．提案法による過強調を緩和するため，反復後の生成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．

HMM 音声合成においては，MSD-HMM [6] 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．ただし [3] と同様に，式 (6) 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

*Parameter Generation Algorithm Considering Modulation Spectrum in Statistical Parametric Speech Synthesis, by TAKAMICHI, Shinnosuke (NAIST/CMU), TODA, Tomoki (NAIST), Black Alan (CMU), NAKAMURA, Satoshi (NAIST)

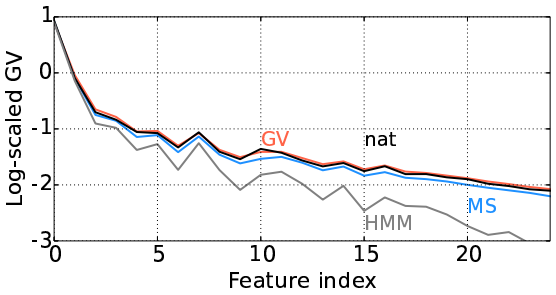

Fig. 1　メルケプストラム系列の GV

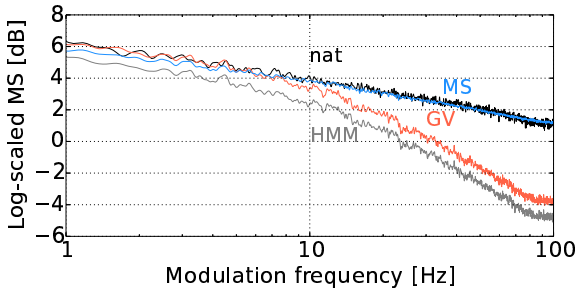

Fig. 2　9 次のメルケプストラム系列の MS

／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 (1)，式 (3)，式 (7) 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．

4　実験的評価

4.1　実験条件

CMU ARCTIC 音声データベース [7] 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 [8] 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 [9] を使用する．MS を計算する DFT のタップ数は 8192 点とする．詳細な実験条件は，表 1 参照とする．

GV を考慮したパラメータ生成法（“GV”）と提案法（“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，非周期成分の生成には “GV” を使用する．

4.2　実験結果

各主観評価結果を Fig. 3 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 [12] 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．

Table 1　実験条件

	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	男性及び女性
音響モデル	5 状態 HSMM	64 混合 GMM
提案法	スペクトル，F_0	スペクトル

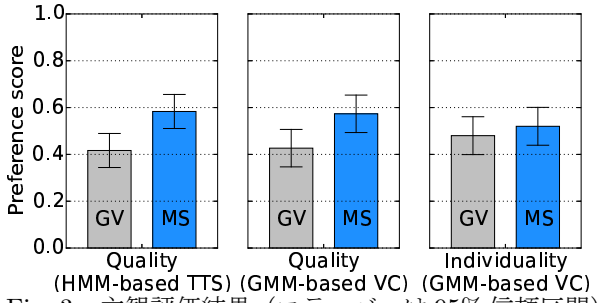

Fig. 3　主観評価結果（エラーバーは 95% 信頼区間）

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 [10]，連続 F_0 モデル [11] への導入，及びポストフィルタ処理 [3, 12, 13] 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] T. Toda *et al.*, *IEICE on Trans. Inf. and Syst.*, Vol. E90-D, No. 5, pp. 816–824, 2007.
[2] T. Toda *et al.*, *IEEE. Trans. on Audio, Speech and Language Processing* Vol. 15, No. 8, pp. 2222–2235, 2007.
[3] S. Takamichi *et al.*, *Proc. ICASSP*, pp. 290–294, 2014.
[4] K. Tokuda *et al.*, *Proc. IEEE*, Vol. 111, No. 5, pp. 1234–1252, 2013.
[5] K. Tokuda *et al.*, *Proc. ICASSP*, pp. 1315–1318, 2000.
[6] K. Tokuda *et al.*, *IEICE Trans. on Inf. and Syst.*, Vol. E85-D, No. 3, pp. 455–464, 2002.
[7] J. Kominek *et al.*, *Tech. Rep.* CMU-LTI-03-177, 2003.
[8] H. Kawahara *et al.*, *Speech Comm.*, Vol. 27, No. 3–4, pp. 187–207, 1999.
[9] 大谷 他，信学論，Vol. J91-D，No. 4，pp. 1082–1091，2008.
[10] H. Zen *et al.*, *Computer Speech and Language*, Vol. 21, No. 1, pp. 153–173, 2007.
[11] K. Yu *et al.*, *IEEE Trans. on Audio, Speech and Language*, Vol. 19, No. 5, pp. 1071–1079, 2011.
[12] S. Takamichi *et al.*, *Proc. APSIPA ASC*, 2014.
[13] S. Takamichi *et al.*, *Proc. GlobalSIP*, pp. 710–714, 2014.