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制約付きトラジェクトリ学習*

○ 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の音質改善には，音声パラメータ生成部における変調スペクトル（MS: Modulation Spectrum）補償が有効である [1, 2]．しかしながら，生成部における補償処理は，短遅延合成処理を必要とするシステムに不向きである．そこで本稿では，HMM 音声合成 [3] と GMM 声質変換 [4] において，静的・動的特徴量の制約下で系列をモデル化するトラジェクトリモデル [5] を，変調スペクトル制約下で学習する手法を提案する．実験的評価により，従来の短遅延合成処理能力を維持しつつ，音質を改善できることを示す．

2 従来の枠組み

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 声質変換では入力・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特徴量系列 $\boldsymbol{X}$ から決定される次式の目的関数 $L(\boldsymbol{y})$ を，静的・動的特徴量間の制約条件下で最大化することで，パラメータ系列 $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}$ を生成する．

$$L(\boldsymbol{y}) = P(\boldsymbol{W}\boldsymbol{y}|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{W}\boldsymbol{y}; \boldsymbol{E}_{\hat{\boldsymbol{q}}}, \boldsymbol{D}_{\hat{\boldsymbol{q}}}) \quad (1)$$

$$\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} = \boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{r}_{\hat{\boldsymbol{q}}} = (\boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{W})^{-1}\boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{E}_{\hat{\boldsymbol{q}}} \quad (2)$$

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 [6] である．準最適な HMM の状態系列又は GMM の分布系列 $\hat{\boldsymbol{q}} = [\hat{q}_1, \cdots, \hat{q}_T]$ を用いた近似により，目的関数は，平均ベクトル $\boldsymbol{E}_{\hat{\boldsymbol{q}}} = [\boldsymbol{E}_{\hat{q}_1,1}, \cdots, \boldsymbol{E}_{\hat{q}_T,T}]$ と精度行列 $\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1} = \mathrm{diag}\left[\boldsymbol{\Sigma}_{\hat{q}_1}^{-1}, \cdots, \boldsymbol{\Sigma}_{\hat{q}_T}^{-1}\right]$ から成る正規分布 $\mathcal{N}(\cdot; \boldsymbol{E}_{\hat{\boldsymbol{q}}}, \boldsymbol{D}_{\hat{\boldsymbol{q}}})$ で与えられる．q_t は時刻 t における HMM 状態もしくは GMM 分布インデックス q であり，$1 \leq q \leq Q$ とする．Q は HMM 状態もしくは GMM 分布の総数である．$\boldsymbol{\mu}_{q,t} = \left\{\boldsymbol{\mu}_q\ (\text{HMM}),\ \boldsymbol{A}_q\boldsymbol{X}_t + \boldsymbol{b}_q\ (\text{GMM})\right\}$ と $\boldsymbol{\Sigma}_q^{-1}$ はそれぞれ，時刻 t における q に対応する平均ベクトル及び精度行列である．$\boldsymbol{\mu}_q$，$\boldsymbol{A}_q$，$\boldsymbol{b}_q$ はそれぞれ，q に対応する平均ベクトル，変換行列，及び，バイアスベクトルであり，$\boldsymbol{X}_t$ は，$\boldsymbol{X}$ における時刻 t のベクトルである．コレスキー分解 [6] や短遅延生成アルゴリズム [7] により，$\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}$ は解析的に生成される．

3 提案する学習法

3.1 トラジェクトリモデル

トラジェクトリ HMM の学習部 [5] では，次式の目的関数 L_{trj} を最大化するように $\boldsymbol{\lambda}$ を推定する．

$$L_{\mathrm{trj}} = P(\boldsymbol{y}|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}) = \mathcal{N}\left(\boldsymbol{y}; \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}, \boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}\right) \quad (3)$$

一方で GMM 声質変換において，結合確率密度関数のトラジェクトリ GMM が提案されている [8] が，本稿では式 (3) と同様に，条件付き確率密度関数としてトラジェクトリ GMM を定式化する．

3.2 変調スペクトル（MS）

MS は，系列内変動（GV: Global Variance）[9, 4] 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 [1]．パラメータ系列 $\boldsymbol{y}$ の MS $\boldsymbol{s}(\boldsymbol{y}) = [\boldsymbol{s}(1), \cdots, \boldsymbol{s}(D)]^\top$ は次式で表される．

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

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

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

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．D_{s}' は MS の次元数を表し，$D_{\mathrm{s}}' \leq D_{\mathrm{s}}$ である．本稿では，音声知覚に強く影響する低変調周波数帯域の MS のみを取り扱う．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 $\mathcal{N}(\boldsymbol{s}(\boldsymbol{y}); \boldsymbol{\mu}_{\mathrm{s}}, \boldsymbol{\Sigma}_{\mathrm{s}})$ で表し，更に，$\boldsymbol{\Sigma}_{\mathrm{s}}^{-1} = [\boldsymbol{p}_{\mathrm{s}}^{(1)}, \cdots, \boldsymbol{p}_{\mathrm{s}}^{(D)}]$ 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．

3.3 変調スペクトル制約付きトラジェクトリ学習

提案法では，次式の目的関数 L_{ms} を最大化するように $\boldsymbol{\lambda}$ を推定する．

$$L_{\mathrm{ms}} = \mathcal{N}\left(\boldsymbol{y}; \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}, \boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}\right) \mathcal{N}\left(\boldsymbol{s}(\boldsymbol{y}); \boldsymbol{s}(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}), \boldsymbol{\Sigma}_{\mathrm{s}}\right)^{\frac{\omega_{\mathrm{s}} T}{D_{\mathrm{s}}'}} \quad (7)$$

ω_{s} は MS 尤度の重みであり，本稿では 1.0 に設定する．ここで，$\boldsymbol{\xi}_{\mathrm{A}} = [\boldsymbol{A}_1^\top, \cdots, \boldsymbol{A}_Q^\top]^\top$, $\boldsymbol{\xi}_{\mathrm{b}} = [\boldsymbol{b}_1^\top, \cdots, \boldsymbol{b}_Q^\top]^\top$, $\boldsymbol{\mu} = [\boldsymbol{\mu}_1^\top, \cdots, \boldsymbol{\mu}_Q^\top]^\top$, $\boldsymbol{\Sigma}^{-1} = [\boldsymbol{\Sigma}_1^{-1}, \cdots, \boldsymbol{\Sigma}_Q^{-1}]^\top$ を，各々全ての $\boldsymbol{A}_q$, $\boldsymbol{b}_q$, $\boldsymbol{\mu}_q$, $\boldsymbol{\Sigma}_q^{-1}$ を結合したパラメータとする．$\boldsymbol{\lambda}$ は，最急降下法 $\boldsymbol{\xi}_{\mathrm{A}}^{(i+1)} = \boldsymbol{\xi}_{\mathrm{A}}^{(i)} + \alpha\left(\partial \log L_{\mathrm{ms}} / \partial \boldsymbol{\xi}_{\mathrm{A}}\right)|_{\boldsymbol{\xi}_{\mathrm{A}} = \boldsymbol{\xi}_{\mathrm{A}}^{(i)}}$ にて反復的に推定される．i は反復インデックス，α は学習係数である．各結合パラメータに関する一次微分は次式で与えられる．

$$\frac{\partial \log L_{\mathrm{ms}}}{\partial \boldsymbol{\xi}_{\mathrm{A}}} = \boldsymbol{S}_{\hat{\boldsymbol{q}}}^\top \mathrm{diag}^{-1}[\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{W}\left(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right)\boldsymbol{X}^\top + \frac{\omega_{\mathrm{s}} T}{D_{\mathrm{s}}'}\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{W}\boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{s}_{\hat{\boldsymbol{q}}}\boldsymbol{X}^\top] \quad (8)$$

$$\frac{\partial \log L_{\mathrm{ms}}}{\partial \boldsymbol{\xi}_{\mathrm{b}}} = \frac{\partial \log L_{\mathrm{ms}}}{\partial \boldsymbol{\mu}} = \boldsymbol{S}_{\hat{\boldsymbol{q}}}^\top (\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{W}\left(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right) + \frac{\omega_{\mathrm{s}} T}{D_{\mathrm{s}}'}\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{W}\boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{s}_{\hat{\boldsymbol{q}}}) \quad (9)$$

$$\frac{\partial \log L_{\mathrm{ms}}}{\partial \boldsymbol{\Sigma}^{-1}} = \frac{1}{2}\boldsymbol{S}_{\hat{\boldsymbol{q}}}^\top \mathrm{diag}^{-1}[\boldsymbol{W}(\boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1} + \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}^\top - \boldsymbol{y}\boldsymbol{y}^\top) - \boldsymbol{E}_{\hat{\boldsymbol{q}}}\left(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} - \boldsymbol{y}\right)^\top \boldsymbol{W}^\top - \boldsymbol{W}\left(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} - \boldsymbol{y}\right)\boldsymbol{E}_{\hat{\boldsymbol{q}}}^\top + \frac{2\omega_{\mathrm{s}} T}{D_{\mathrm{s}}'}\boldsymbol{W}\boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{s}_{\hat{\boldsymbol{q}}}\left(\boldsymbol{E}_{\hat{\boldsymbol{q}}} - \boldsymbol{W}\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right)] \quad (10)$$

---

*Modulation Spectrum-Constrained Trajectory Training Algorithm in Statistical Parametric Speech Synthesis, by TAKAMICHI, Shinnosuke (NAIST/CMU), TODA, Tomoki (NAIST), Black Alan (CMU), NAKAMURA, Satoshi (NAIST)

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

ただし，$\boldsymbol{S}_{\hat{\boldsymbol{q}}}$ は $\hat{\boldsymbol{q}}$ を表現する行列である [5]．$\boldsymbol{s}_{\hat{\boldsymbol{q}}}$ は次式で与えられる．

$$\boldsymbol{s}_{\hat{\boldsymbol{q}}} = [\boldsymbol{s}_1'^{\top}, \cdots, \boldsymbol{s}_T'^{\top}]^{\top}, \quad (11)$$

$$\boldsymbol{s}_t' = [s_t(1), \cdots, s_t(D)]^{\top}, \quad (12)$$

$$s_t(d) = 2\boldsymbol{f}_t(d)\,\boldsymbol{p}_{\mathrm{s}}^{(d)\top}(\boldsymbol{s}(\boldsymbol{y}) - \boldsymbol{s}(\hat{\boldsymbol{y}})), \quad (13)$$

$$\boldsymbol{f}_t(d) = [f_{t,d}(0), \cdots, f_{t,d}(D_{\mathrm{s}}' - 1)]^{\top}, \quad (14)$$

$$f_{t,d}(f) = \hat{R}_{d,f}\cos kt + \hat{I}_{d,f}\sin kt, \quad (15)$$

ただし，$\hat{R}_{d,f}$ と $\hat{I}_{d,f}$ は $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}$ から計算される．本稿では，通常のトラジェクトリ学習による結果を初期値として，最急降下法による反復更新を行う．HMM 音声合成における F_0 モデリングには連続 F_0 モデル [10] を使用し，[1] と同様に系列平均を 0 とした MS を計算する．GMM 声質変換においては，結合 GMM パラメータの内，分布重みと，入力特徴量の平均ベクトル及び共分散行列については，更新しない．

3.4　考察

学習部における MS 制約により生成パラメータ $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}$ の変調スペクトルは補償されるため，生成部における補償は不要である．故に，式 (2) の生成処理を使用しつつ，音質改善が可能である．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．

4　実験的評価

4.1　実験条件

CMU ARCTIC 音声データベース [11] 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 [12] 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 [13] を使用する．ω_{s} は 1.0 に設定する．D_{s} は 2048 とする．D_{s}' は，スペクトルにおいて $D_{\mathrm{s}}/2$ (= 50Hz) [2]，F_0 において $D_{\mathrm{s}}/10$ (= 10Hz) [1] に設定する．詳細な実験条件は，表 1 参照とする．

以下の学習法を比較する．パラメータ生成法には式 (2) を使用する．

BASIC：通常の学習 [3, 4]
TRJ：トラジェクトリ学習 [5][1]
GV：GV 制約付き（トラジェクトリ）学習 [14, 15][2]
MS：MS 制約付きトラジェクトリ学習　（提案法）

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．被験者は，各評価に対して 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，非周期成分の学習には “BASIC” を使用する．

[1] トラジェクトリ HMM は [5] による定式化，トラジェクトリ GMM は本稿による定式化である．
[2] [15] はトラジェクトリ GMM を使用していない．

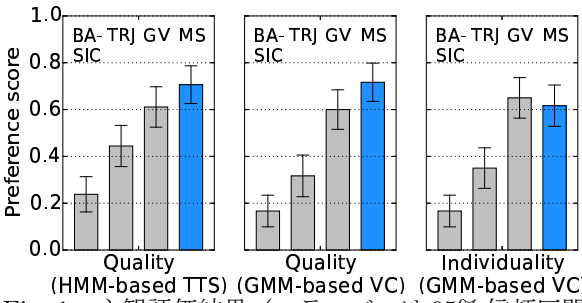

Fig. 1　主観評価結果（エラーバーは 95% 信頼区間）

4.2　実験結果

各主観評価結果を Fig. 1 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（“MS”）のスコアが最も高い．一方で，“MS” の話者性のスコアは，“TRJ” より高いものの “GV” と同程度のスコアとなっている．これらの結果より，提案する変調スペクトル制約付き学習により，解析的なパラメータ生成処理を保持しつつ，変調スペクトル補償による音質改善効果が得られることが明らかになった．

5　まとめ

本稿では，統計的パラメトリック音声合成において，単遅延合成処理能力を維持しつつ音質を改善させる手法として，変調スペクトル制約付きトラジェクトリ学習法を提案し，実験的評価により有効性を確認した．今後は，任意話者音声合成への応用を検討する．

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] S. Takamichi *et al.*, *Proc. ICASSP*, pp. 290–294, 2014.
[2] 高道 他, 音講論（春）, 2-2-2, 2015.
[3] K. Tokuda *et al.*, *Proc. IEEE*, Vol. 111, No. 5, pp. 1234–1252, 2013.
[4] T. Toda *et al.*, *IEEE. Trans. on Audio, Speech and Language Processing*, Vol. 15, No. 8, pp. 2222–2235, 2007.
[5] H. Zen *et al.*, *Computer Speech and Language*, Vol. 21, No. 1, pp. 153–173, 2007.
[6] K. Tokuda *et al.*, *Proc. ICASSP*, pp. 1315–1318, 2000.
[7] T. Muramatsu *et al.*, *Proc. INTERSPEECH*, pp. 1076–1079, 2008.
[8] H. Zen *et al.*, *IEEE Trans.*, Vol. 19, pp. 417–430, 2011.
[9] T. Toda *et al.*, *IEICE on Trans. Inf. and Syst.*, Vol. E90-D, No. 5, pp. 816–824, 2007.
[10] K. Yu *et al.*, *IEEE Trans. on Audio, Speech and Language*, Vol. 19, No. 5, pp. 1071–1079, 2011.
[11] J. Kominek *et al.*, *Tech. Rep.* CMU-LTI-03-177, 2003.
[12] H. Kawahara *et al.*, *Speech Comm.*, Vol. 27, No. 3–4, pp. 187–207, 1999.
[13] 大谷 他，信学論，Vol. J91-D，No. 4，pp. 1082–1091，2008.
[14] T. Toda *et al.*, *Proc. ICASSP*, pp. 4025–4028, 2009.
[15] H. Hwang *et al.*, *Proc. APSIPA ASC*, pp.1–6, 2014.