HMM 音声合成におけるスペクトル・F0 の 分散共有フルコンテキストモデルによる音質改善*

☆ 高道慎之介, 戸田智基 (奈良先端大), 志賀芳則 (NICT),
Sakriani Sakti, Graham Neubig, 中村 哲 (奈良先端大)

1 はじめに

HMM 音声合成において，汎化による生成パラメータの過剰な平滑化は，音質劣化の一因となる．これに対し，我々は HMM 音声合成の利点を保持した素片選択型合成とのハイブリッド法として，分散共有フルコンテキストモデルを用いたパラメータ生成法を提案し，スペクトルパラメータにおいてその有効性を示している [1]．本稿では，より音質の高い合成音声を得るために，分散共有フルコンテキストモデルによる F_0 パターン生成法を提案する．実験的評価結果から，提案法により合成音声の音質が向上することを示す．

2 分散共有フルコンテキストモデルとパラメータ生成法

2.1 HMM 音声合成における状態共有モデル

HMM 音声合成において，考慮するコンテキスト情報 (フルコンテキスト) は膨大であり，学習データにおいて，各フルコンテキストはしばしば一つの音声素片のみに対応する．故に，各フルコンテキストに対するフルコンテキストモデルのスパース性は高く，未知音声に対する頑健性に乏しい．そこで，各フルコンテキスト要因に対する質問で構成される決定木により，音声パラメータ毎，HMM 状態毎にフルコンテキストモデルをクラスタリングして [2]，クラス c 毎に出力確率密度関数（状態共有モデル）b_c をモデル化する．

スペクトルパラメータ: スペクトルパラメータは，次式の出力確率密度関数を持つ連続 HMM でモデル化される．

$$b_c(\boldsymbol{o}_t) = \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \tag{1}$$

ただし，$\boldsymbol{o}_t = \left[\boldsymbol{c}_t^\top, \Delta\boldsymbol{c}_t^\top, \Delta\Delta\boldsymbol{c}_t^\top\right]^\top$ は，時刻 t における静的特徴量 $\boldsymbol{c}_t$ とその一次と二次の動的特徴量 $\Delta\boldsymbol{c}_t$, $\Delta\Delta\boldsymbol{c}_t$ の結合ベクトルを表し，$\mathcal{N}(\cdot; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ は，平均 $\boldsymbol{\mu}_c$，共分散行列 $\boldsymbol{\Sigma}_c$ を持った正規分布を表す．

F0 パラメータ: F_0 パラメータは，次式の出力確率密度関数を持つ多空間確率分布 HMM（MSD-HMM）[3] でモデル化される．

$$b_c(\boldsymbol{o}_t) = \begin{cases} w_c \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), & l_t = \mathrm{V} \\ 1 - w_c, & l_t = \mathrm{U} \end{cases} \tag{2}$$

ただし，l_t は，時刻 t における有声（V）/無声（U）ラベルを表し，w_c は有声空間の重みである．l_t は特徴量 $\boldsymbol{o}_t$ と同時に観測される．

状態共有モデルは，多数の素片を一つの分布でモデル化するため，未知音声に対する頑健性を持つ一方で，生成されるパラメータは過剰に平滑化される．

2.2 分散共有フルコンテキストモデル

未知音声に対する頑健性を保ちつつ，過剰な平滑化の影響を緩和する方法として，分散共有フルコンテキストモデルがある [4]．連続 HMM において，クラス c に属する要素番号 m の分散共有フルコンテキストモデルの出力確率密度関数 $b_{c,m}$ は，フルコンテキスト毎（概ね素片毎）の平均 $\boldsymbol{\mu}_{c,m}$ とクラスで共有する共分散行列 $\boldsymbol{\Sigma}_c$ を持つ正規分布 $\mathcal{N}(\cdot; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c)$ により，次式で示される．

$$b_{c,m}(\boldsymbol{o}_t) = \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c) \tag{3}$$

フルコンテキスト毎の $\boldsymbol{\mu}_{c,m}$ は，状態共有モデルを用いて計算される十分統計量に基づき推定する．

2.3 分散共有フルコンテキストモデルを用いたパラメータ生成法 [1]

合成するフルコンテキストに対応するクラス c は決定木から求められるが，クラスに属する分散共有フルコンテキストモデルは多数存在するため，使用するモデルを選択してパラメータ生成を行う必要がある．これに対して，尤度基準によるモデル選択を実現するために，クラス c に属する M_c 個の分散共有フルコンテキストモデルから，次式の GMM を構築する．

$$b_c(\boldsymbol{o}_t) = \sum_{m=1}^{M_c} \omega_m \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c) \tag{4}$$

ただし，ω_m は重みであり，$\omega_m = 1/M_c$ とする．

合成時にはまず，状態継続長モデルにより HMM 状態系列 $\boldsymbol{q} = [q_1, \cdots, q_T]^\top$ を与える．次に，初期パラメータ系列 $\boldsymbol{c}^{(0)}$ を決定した後，静的・動的特徴量の制約（$\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$）の下で尤度を最大にするように，単一モデル系列 $\boldsymbol{m} = [m_1, \cdots, m_T]$ 及び，静的パラメータ系列 $\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \cdots, \boldsymbol{c}_T^\top\right]^\top$ を次式にて反復的に更新する．

$$\hat{\boldsymbol{m}}^{(i+1)} = \underset{\boldsymbol{m}}{\mathrm{argmax}}\, P\left(\boldsymbol{m} | \boldsymbol{W}\boldsymbol{c}^{(i)}, \boldsymbol{q}, \boldsymbol{\lambda}\right) \tag{5}$$

$$\hat{\boldsymbol{c}}^{(i+1)} = \underset{\boldsymbol{c}}{\mathrm{argmax}}\, P\left(\boldsymbol{W}\boldsymbol{c} | \boldsymbol{m}^{(i+1)}, \boldsymbol{q}, \boldsymbol{\lambda}\right) \tag{6}$$

ただし，$\boldsymbol{\lambda}$ は HMM のパラメータセットであり，$\boldsymbol{W}$ は，動的特徴量の計算に用いる重み係数によって決まる行列である [5]．

2.4 初期パラメータ生成法 [6]

分散共有フルコンテキストモデルを用いたパラメータ生成法では，反復処理による生成パラメータ系列が初期パラメータ系列に強く依存する [1]．従来の状態共有モデルを用いて初期パラメータを生成する場合，過剰な平滑化の影響を強く受けるため，音質の改善は僅かである．これに対し我々は，大きなサイズの決定木を用いた初期パラメータ生成法を提案している [6]．MDL 基準 [7] における決定木のサイズを決定するパラメータを，生成パラメータの系列内変動（Global Variance: GV）尤度 [8] が最大となるように設定することで音質を改善することが可能である．

3 分散共有フルコンテキストモデルを用いた F0 パターン生成

MSD-HMM に対して，分散共有フルコンテキストモデルによるパラメータ生成法を適用する．MSD-HMM における分散共有フルコンテキストモデルは，

---

*Quality Improvements with Rich Context Models for Spectral and F0 Components in HMM-based Speech Synthesis, by TAKAMICHI, Shinnosuke, TODA, Tomoki (NAIST), SHIGA, Yoshinori (NICT), SAKTI, Sakriani, NEUBIG, Graham, NAKAMURA, Satoshi (NAIST)

有声空間の平均ベクトルを更新することで得られる．

$$b_{c,m}(\boldsymbol{o}_t) = \begin{cases} w_c \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c\right), & l_t = \mathrm{V} \\ 1 - w_c, & l_t = \mathrm{U} \end{cases} \quad (7)$$

ただし，空間重みは状態共有モデルと同じ重みを使用する．次に，次式に示すように有声空間の正規分布を用いて GMM を構築する．

$$b_c(\boldsymbol{o}_t) = \begin{cases} \sum_{m=1}^{M_c} w_{c,m} \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c\right), & l_t = \mathrm{V} \\ 1 - w_c, & l_t = \mathrm{U} \end{cases} \quad (8)$$

ただし，$w_{c,m}$ は，クラス c に属する要素番号 m の分散共有フルコンテキストモデルの有声空間重みを表す．空間重みは最尤推定により計算可能だが，スペクトルパラメータにおいて等重みの有効性が示されているため，$\omega_{c,m} = \omega_c / M_c$ とする．

パラメータ生成時には，初期パラメータ生成法により有声/無声判定と初期パラメータ生成を行った後，分散共有フルコンテキストモデルを用いたパラメータ生成法によりパラメータ系列を生成する．

4 実験的評価

4.1 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 [9] A-I セット 450 文，評価データは同 J セット 53 文を使用する．スペクトル特徴量は，STRAIGHT 分析 [10] 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 及び 5 周波数帯域における平均非周期成分を使用する．HMM は 5 状態 left-to-right 型とし，パラメータ生成時には GV[8] を考慮しない．平均非周期成分と状態継続長には状態共有モデルを使用する．大きなサイズの決定木を用いた初期パラメータ生成法では，最終的に生成されるパラメータの GV 尤度が最大となるように決定木のサイズを設定する．

まず，提案法の有効性を評価する．初期パラメータとして，状態共有モデルから生成した特徴量 (Proposed (Clus))，大きなサイズの決定木を用いた初期パラメータ生成法により生成した特徴量 （Proposed (Tree)），自然音声の特徴量 （Target）を用いる際の，分散共有フルコンテキストモデルから生成したパラメータの合成音声の音質を評価する．また，従来法として，状態共有モデルから生成したパラメータ （Conv）を用いる．ただし，スペクトルパラメータは状態共有モデルを使用する．評価は 6 人の受聴者によるプリファレンススコアとする．

次に，スペクトル・F_0 の分散共有フルコンテキストモデルの有効性を評価する．合成音声には，表 1 に示すように，従来の状態共有モデル（Clustered），分散共有フルコンテキストモデル（Rich Context），ターゲット（Target）を組み合わせた手法により合成されたものを使用する．ターゲットは，自然音声の特徴量を初期パラメータとした分散共有フルコンテキストモデルにより生成する．”Rich Context” では，大きなサイズの決定木を用いた初期パラメータ生成法を使用する．評価は 8 人の受聴者によるプリファレンススコアとする．

4.2 実験結果

F_0 の分散共有フルコンテキストモデルにおける主観評価結果を，図 1(a) に示す．初期パラメータ生成法を用いた提案法のスコアが従来法のスコアより高いことから，提案法の有効性が示される．

スペクトル・F_0 の分散共有フルコンテキストモデルにおける音質の主観評価結果を図 1(b) に示す．分散共有フルコンテキストモデルのスペクトルパラメータへの適用によって著しく音質が改善し，更に，F_0 への適用により音質が改善する事が分かる．また，スペクトル・F_0 に適用した手法（”PP”）のスコアはターゲット（”TT”）のスコアに接近していることから，スペクトル・F_0 の分散共有フルコンテキストモデルを用いたパラメータ生成法は，音質改善に対して非常に有効であることが分かる．

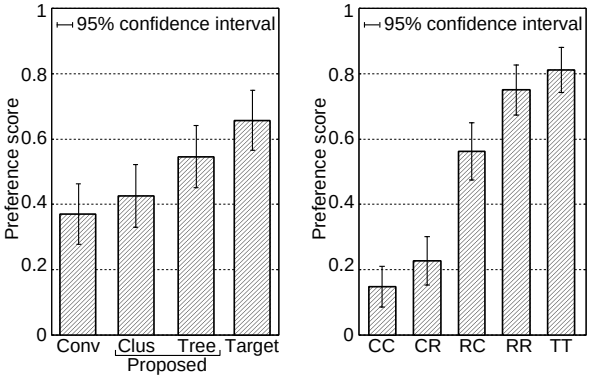

Fig. 1 分散共有フルコンテキストモデルによる提案法の主観評価結果 （(a) F_0 に適用，(b) スペクトル・F_0 に適用）

Table 1 評価に用いる手法

Method	Spectrum	F_0
CC	Clustered	Clustered
CR	Clustered	Rich Context
RC	Rich Context	Clustered
RR	Rich Context	Rich Context
TT	Target	Target

5 まとめ

本稿では，F_0 の分散共有フルコンテキストモデルを用いたパラメータ生成法を提案し，実験的評価でスペクトル・F_0 に対する分散共有フルコンテキストモデルの有効性を示した．その結果，従来の HMM 音声合成と比較して，著しい音質改善が明らかになった．今後は，分散共有フルコンテキストモデルを用いた話者適応法について検討する．

謝辞 本研究の一部は，JSPS 科研費 24240032 の助成を受け実施したものである．

参考文献

[1] S. Takamichi *et al.*, Proc. *INTERSPEECH*, 2012.
[2] 吉村 他, 信学論 (D–2), Vol. J83–D–2, pp. 2099–2107, 2000.
[3] K. Tokuda *et al.*, IEICE Trans., Inf. and Syst., Vol. E85–D, No. 3, pp. 455–464, 2002.
[4] Z. Yan *et al.*, Proc. *INTERSPEECH*, pp. 1755-1758, 2009.
[5] H. Zen *et al.*, *Speech Commun.*, 51(11), pp. 1039–1064, 2009.
[6] 高道 他, 信学技報, SP2012–78, pp. 37–42, 2012.
[7] K. Shinoda *et al.*, J. Acoust. Soc. Jpn.(E), Vol. 21, No. 2, pp.79–86, 2000.
[8] T. Toda *et al.*, *IEICE Transactions*, Vol. E90–D, No. 5, pp. 816–824, 2007.
[9] 阿部 他, ATR テクニカルレポート, TR–I–0166, 1990.
[10] H. Kawahara *et al.*, Speech Communication, Vol. 27, No. 3–4, pp. 187–207, 1999.