HMM 音声合成における 分散共有フルコンテキストモデルの選択法に関する検討*

☆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た [1]．この手法では，初期パラメータ系列を決定後，尤度最大化基準により，反復的にパラメータが生成されるが，得られる合成音声の音質は初期パラメータ系列に大きく依存する．本稿では，より音質の高い合成音声を得るために，コンテキストクラスタリングを用いた初期パラメータ決定法を提案し，実験的評価によりその有効性を示す．

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 分散共有フルコンテキストモデル

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，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を正規分布でモデル化する．クラスタリング基準として，次式で表される最小記述長（minimum description length: MDL）基準を用いる [3].

$$l^{(C)} = \frac{1}{2}\sum_{c=1}^{C}\Gamma(c)\log|\boldsymbol{\Sigma}_c| + aCD\log\Gamma(0) \qquad (1)$$

ただし，C は総クラス数，D は次元数，$\boldsymbol{\Sigma}_c$ はクラス c の共分散行列，$\Gamma(c)$ 及び $\Gamma(0)$ は，それぞれクラス c および決定木のルートにおける状態占有確率の総和を表す．a は総クラス数 C を制御するパラメータであり，a が小さいほど C が増加する．

標準的な設定（$a = 1.0$）によるコンテキストクラスタリングで得ら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，生成されるパラメータは過剰に平滑化される．これに対し，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]．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 $\mathcal{N}\left(\cdot;\boldsymbol{\mu}_{c,m},\boldsymbol{\Sigma}_c\right)$ により，次式で示される．

$$b_{c,m}(\boldsymbol{o}_t) = \mathcal{N}\left(\boldsymbol{o}_t;\boldsymbol{\mu}_{c,m},\boldsymbol{\Sigma}_c\right) \qquad (2)$$

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す．

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．

2.2 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 は決定木により求められるが，その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．まず，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，次式の混合正規分布モデル（GMM）を構築する．

$$b_c(\mathbf{o}_t) = \sum_{m=1}^{M_c}\omega_m\mathcal{N}\left(\boldsymbol{o}_t;\boldsymbol{\mu}_{c,m},\boldsymbol{\Sigma}_c\right) \qquad (3)$$

ただし，ω_m は重みであり，$\omega_m = 1/M_c$ とする．次に，状態継続長モデルにより与えられる HMM 状態系列 $\boldsymbol{q} = [q_1,\cdots,q_T]^\top$ を用いた際の尤度関数を，次式で近似する．

$$P(\boldsymbol{o}|\boldsymbol{q},\boldsymbol{\lambda}) = \sum_{\text{all }\boldsymbol{m}}P(\boldsymbol{o},\boldsymbol{m}|\boldsymbol{q},\boldsymbol{\lambda}) \simeq P(\boldsymbol{o},\boldsymbol{m}|\boldsymbol{q},\boldsymbol{\lambda}) \qquad (4)$$

ただし，特徴量系列を $\boldsymbol{o} = \left[\boldsymbol{o}_1^\top,\cdots,\boldsymbol{o}_T^\top\right]^\top$，モデル系列を $\boldsymbol{m} = [m_1,\cdots,m_T]^\top$ とし，HMM のパラメータセットを $\boldsymbol{\lambda}$ とする．

生成時には，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，単一モデル系列及び，静的パラメータ系列を次式にて反復的に更新する．

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

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

ただし，パラメータ系列を $\boldsymbol{c} = \left[\boldsymbol{c}_1^\top,\cdots,\boldsymbol{c}_T^\top\right]^\top$ とし，$\boldsymbol{W}$ 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].

2.3 初期パラメータ系列の影響

静的・動的特徴量に対する尤度を考慮することで，素片選択処理と同様に，時間的な連続性を考慮して各音声素片に対応する分散共有フルコンテキストモデルが選択される．一方で，反復処理により最終的に選択されるモデルは，初期パラメータ系列に大きく依存する．状態共有モデルから生成した特徴量を初期パラメータとすることで，合成音声の音質が改善されることが明らかになっているが [1]，過剰な平滑化の影響は未だに大きく，その音質は十分ではない．これには，過剰に平滑化された初期パラメータ系列が影響していると考えられる．

3 コンテキストクラスタリングを用いた初期パラメータ系列の生成

より適切な初期パラメータ系列を与えるため，サイズの大きな決定木に基づくコンテキストクラスタリングを用いる手法を提案する．コンテキストクラスタリングにおいて，MDL 基準のパラメータ a を小

*A Study on a Selection Method of Rich Context Models in HMM-based Speech Synthesis. by TAKAMICHI, Shinnosuke, TODA, Tomoki (NAIST), SHIGA, Yoshinori (NICT), SAKTI, Sakriani, NEUBIG, Graham, NAKAMURA, Satoshi (NAIST)

さくすることで，決定木のクラス数（サイズ）を大きくすることが可能である．この決定木におけるクラス毎のモデルはクラスに属する少数の音声素片から計算され，生成されるパラメータは過剰な平滑化の影響が少ない．そのため，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時には，この初期パラメータ系列の影響を大きく受け，平滑化の影響の少ないモデル系列が選択されると予想される．一方，サイズの大きな決定木におけるクラス毎のモデルから生成されるパラメータは，しばしば時間的に不連続な遷移を含む．これに対し，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は，標準的な状態共有モデルの共分散行列の使用により汎化性能を高め，さらに静的・動的特徴量を考慮したモデル選択を行う事で，時間的な不連続性を緩和する．

決定木は，分散共有フルコンテキストモデルの計算時と同様の十分統計量を用いて構築する．この決定木は，分散共有フルコンテキストモデルの決定木とは異なることに注意する．クラスタリング基準としてMDL基準を用い，MDL基準のパラメータaは，分散共有フルコンテキストモデルの決定木よりも小さく設定する．

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 [6] 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 [7] 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系列内変動（Global Variance: GV）[8] を考慮しない．分散共有フルコンテキストモデルの決定木構築時には，MDL 基準の a を 1.0 とする．本稿では，スペクトル特徴量に対してのみ，分散共有フルコンテキストモデルを適用する．

まず，提案法により生成した初期パラメータ系列を用いて，反復処理により最終的に選択された分散共有フルコンテキストモデル系列と生成されたパラメータを評価する．評価基準は，自然音声の特徴量に対する HMM の対数尤度 $\log P(\boldsymbol{o}|\boldsymbol{m},\boldsymbol{q},\boldsymbol{\lambda})$ 及び，生成パラメータに対する GV の対数尤度 [8] とする．評価は，提案法の決定木の a を 0.1 から 1.0 まで 0.1 刻みで変化させ，それぞれの決定木によるモデルから生成された初期パラメータにおいて行う．同時に，自然音声の特徴量（Target）と標準的な状態共有モデルで生成した特徴量（Conv）を初期パラメータとした場合の尤度も計算する．

次に，提案法の音質を評価する．標準的な状態共有モデルで生成した特徴量（Conv），提案法の $a = 0.5, 0.1$ の決定木によるモデルを用いて生成した特徴量（Proposed），自然音声の特徴量（Target）の 4 種類の初期パラメータを用いて得られる合成音声の音質を評価する．ただし，音源特徴量・継続長は標準的な状態共有モデルを用いる．評価は 7 人の受聴者に対するプリファレンススコアとする．

4.2 実験結果

HMM の対数尤度を Fig. 1 に，GV の対数尤度を Fig. 2 に示す．また，パラメータ a を変化させたときの決定木のサイズを Table 1 に示す．Fig. 1 から，a を小さくするに従い HMM 尤度はわずかに上昇し，$a = 0.5$ で最大となり，その後急激に減少する事が分かる．しかしながら，その尤度は自然音声の特徴量を初期パラメータとした場合の尤度に大きく及ばないことが分かる．一方，Fig. 2 から，a を減少させるに従い，GV 尤度は自然音声の特徴量を初期パラメータとした場合の尤度と同等以上まで上昇することが分かる．

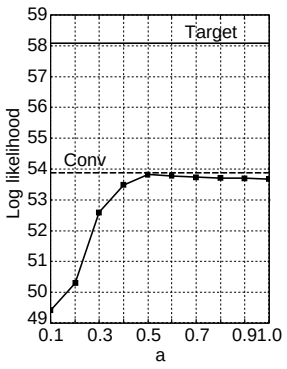

Fig. 1 自然音声の特徴量に対する HMM 尤度

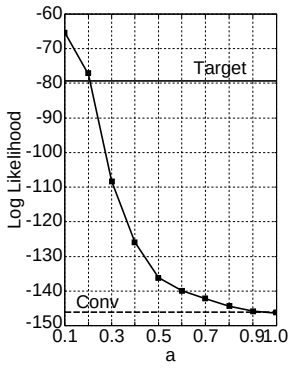

Fig. 2 生成パラメータに対する GV 尤度

Table 1 決定木のサイズ（フルコンテキストモデルサイズに対する圧縮率）

a	Tree size (%)
0.1	90.5
0.2	32.9
0.3	9.56
0.4	4.06
0.5	1.76
0.6	1.25
0.7	0.97
0.8	0.83
0.9	0.75
1.0	0.70

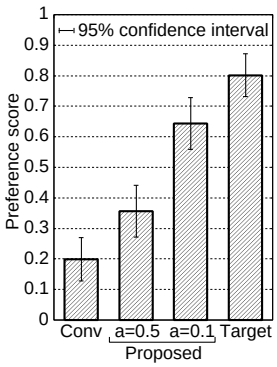

Fig. 3 主観評価結果

音質の主観評価結果を Fig. 3 に示す．提案法のうち $a = 0.1$ のスコアが最も高く，自然音声の特徴量を初期パラメータとした場合のスコアに接近していることから，提案法の有効性が示される．この結果と Fig. 2 から，合成音声の音質と GV 尤度との間には，比較的高い相関があることが分かる．

5 まとめ

本稿では，分散共有フルコンテキストモデルの選択法として，コンテキストクラスタリングを用いる初期パラメータ系列生成法を示し，実験的評価で有効性を示した．今後は，モデル系列選択尺度の検討，及び，音源パラメータに対する提案法の適用を行う．

参考文献

[1] 高道 他, 情処研報, Vol. 2012-SLP-92, No. 10, pp. 1–6, 2012.
[2] 吉村 他, 信学論 (D-2), Vol. J83-D-2, pp. 2099–2107, 2000.
[3] K. Shinoda *et al.*, J. Acoust. Soc. Jpn.(E), Vol. 21, No. 2, pp.79–86, 2000.
[4] Z. Yan *et al.*, *INTERSPEECH* 2009, pp. 1755–1758, 2009.
[5] H. Zen *et al.*, *Speech Commun.*, 51(11), pp. 1039–1064, 2009.
[6] 阿部 他, ATR テクニカルレポート, TR-I-0166, 1990.
[7] H. Kawahara *et al.*, Speech Communication, Vol. 27, No. 3–4, pp. 187–207, 1999.
[8] T. Toda *et al.*, *IEICE Transactions*, Vol. E90-D, No. 5, pp. 816–824, 2007.