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有効性に関する検討*

☆高道 慎之介, 戸田 智基 (奈良先端大), 志賀 芳則, 河井 恒 (NICT), サクテイ サクリアニ, 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，素片選択型合成とのハイブリッド方式が提案されている [1]．波形素片の利用で音質改善が得られる一方，HMM 音声合成の利点であるモデル適応処理などの実現が困難となる．本稿では，HMM 音声合成の利点を生かしたハイブリッド方式として，分散共有フルコンテキストモデル [2] を用いた尤度に基づくパラメータ生成法を提案し，実験的評価で有効性を示す．

2 コンテキストクラスタリングによる汎化

HMM 音声合成において，考慮する文脈情報 (フルコンテキスト) は膨大であり，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，フルコンテキストモデルをクラスタリングして分布を共有する [3]．ここで，クラス c の出力確率密度関数は次式でモデル化される．

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

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，$\mathcal{N}\left(\cdot; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right)$ 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．

合成時には，合成音声のフルコンテキストに対するクラスを決定し，得られる分布系列からパラメータ系列 $\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \cdots, \boldsymbol{c}_T^\top\right]^\top$ を尤度基準で生成する [4]．クラスタリングにより，多数の素片を一つの分布でモデル化するため，高い汎化性能が得られる半面，生成されるパラメータは過剰に平滑化される．

3 分散共有フルコンテキストモデルに基づく合成法

3.1 分散共有フルコンテキストモデル

過剰な平滑化の影響を緩和しつつ未知音声に対する頑健性を向上させる方法として，分散共有フルコンテキストモデルが提案されている [2]．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 $b_{c,m}$ 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 $\boldsymbol{\Sigma}_c$ を持つ正規分布により，次式で示される．

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

学習時には，頑健なアライメント処理を行うために，コンテキストクラスタリングに基づくモデルにより十分統計量計算を行った後，フルコンテキスト毎の $\boldsymbol{\mu}_{c,m}$ を推定する．

3.2 パラメータ生成法

合成するフルコンテキストに対応するクラス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．本稿では，生成パラメータ系列に対する尤度に基づく手法として，混合正規分布 (Gaussian Mixture Model： GMM) モデルを用いるパラメータ生成法と，単一分布近似を用いるパラメータ生成法を提案する．

3.2.1 GMM を用いる手法

クラス c に属する M 個の分散共有フルコンテキストモデルから，次式の GMM を計算する．

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

ただし，ω_m は重みであり，$\omega_m = 1/M$ とする．状態継続長分布により決定された HMM 状態系列 $\boldsymbol{q} = [q_1, \cdots, q_T]^\top$ を用いて，尤度関数は次式で表わされる．

$$P\left(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\lambda}\right) = \sum_{all\ \boldsymbol{m}} P\left(\boldsymbol{o}, \boldsymbol{m}|\boldsymbol{q}, \boldsymbol{\lambda}\right) \tag{4}$$

ただし，分布系列を $\boldsymbol{m} = [m_1, \cdots, m_T]^\top$，特徴量系列を $\boldsymbol{o} = \left[\boldsymbol{o}_1^\top, \cdots, \boldsymbol{o}_T^\top\right]^\top$，HMM のパラメータセットを $\boldsymbol{\lambda}$ とする．静的・動的特徴量間の制約（$\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$）の下で尤度関数を最大化することで，パラメータ系列 $\hat{\boldsymbol{c}}$ を生成する．

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmax}} \sum_{all\ \boldsymbol{m}} P\left(\boldsymbol{o}, \boldsymbol{m}|\boldsymbol{q}, \boldsymbol{\lambda}\right) \tag{5}$$

ここで，$\boldsymbol{W}$ は，動的特徴量の計算に用いる重み係数によって決まる [5]．初期特徴量系列を決定し，EM アルゴリズムにより $\hat{\boldsymbol{c}}$ を更新することで最尤パラメータ系列を生成する [4]．

3.2.2 単一分布近似を用いる手法

式 (4) で示される尤度関数を単一分布系列 $\boldsymbol{m}$ により次式で近似する．

$$\sum_{all\ \boldsymbol{m}} P\left(\boldsymbol{o}, \boldsymbol{m}|\boldsymbol{q}, \boldsymbol{\lambda}\right) \simeq P\left(\boldsymbol{o}|\boldsymbol{m}, \boldsymbol{q}, \boldsymbol{\lambda}\right) \tag{6}$$

初期特徴量系列を決定した後に，次式に示す通り，単一分布系列の決定とパラメータの生成を繰り返すことで最尤パラメータ系列を生成する．

$$\hat{\boldsymbol{m}} = \underset{\boldsymbol{m}}{\operatorname{argmax}}\, P\left(\boldsymbol{o}|\boldsymbol{m}, \boldsymbol{q}, \boldsymbol{\lambda}\right) \tag{7}$$

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmax}}\, P\left(\boldsymbol{o}|\hat{\boldsymbol{m}}, \boldsymbol{q}, \boldsymbol{\lambda}\right) \tag{8}$$

提案法は，各素片を分散共有フルコンテキストモデルで表わし，フレーム毎にモデルを選択する処理とみなせる．静的・動的特徴量の両者を考慮することで，素片選択処理と同様に連続性も考慮した選択処理が行われる．また，各音声パラメータ（スペクトルや F_0 など）に対して独立に選択処理が可能なため，HMM 音声合成の利点であるモデル適応などの実現が容易である．GMM を用いた手法は，複数のモデルを選択・融合する処理とみなせる．なお，分布系列には HMM 状態内で同一の分布を用いる制約を入れることも容易である．

---

* A Study on the Effectiveness of Full-context Models with Tied-covariance Matrices in HMM-based Speech Synthesis. by TAKAMICHI, Shinnosuke, TODA, Tomoki (NAIST), SHIGA, Yoshinori, KAWAI, Hisashi (NICT), SAKTI, Sakriani, NAKAMURA, Satoshi (NAIST)

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．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．また，スペクトル特徴量の違いのみに着目するために，合成時に用いる状態系列として，自然音声の特徴量系列に対して，コンテキストクラスタリングによる従来モデルを用いて状態アライメントした結果を用いる．

4.2 初期特徴量に対する依存性

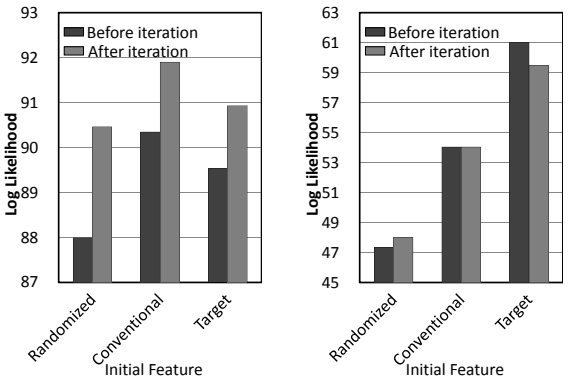
提案法における初期特徴量系列の依存性を調査する．単一分布近似を用いる手法において，初期特徴量系列により選択された分布系列（Before iteration）と，反復処理により最終的に選択された分布系列（After iteration）を，生成された特徴量系列と自然音声の特徴量系列に対する対数尤度により評価する．初期特徴量として，各クラスの分散共有フルコンテキストモデルから無作為抽出した分布から生成した特徴量（Randomized），従来のコンテキストクラスタリングによるモデルで生成した特徴量（Conventional），自然音声の特徴量（Target）の 3 種類を用いる．

実験結果を Fig.1 に示す．Fig.1(a) より，いずれの初期特徴量系列を用いた場合においても，反復処理により生成特徴量に対する尤度は上昇し，いずれも比較的近い値になる．一方，Fig.1(b) より，自然音声の特徴量に対する尤度は必ずしも上昇せず，初期特徴量に大きく依存する．従来モデルで生成した特徴量を用いて分布選択を行うことで，無作為に分布選択を行う場合と比較し，良好な結果が得られるが，生成される特徴量に対する尤度基準では最適な分布選択は困難であることが分かる．

4.3 提案法の有効性

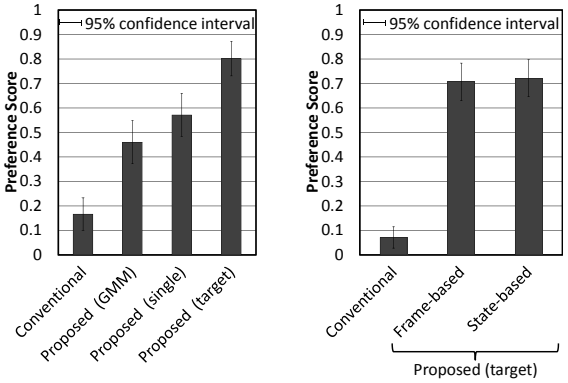
提案法の有効性の評価と分布系列決定単位の比較を行う．提案法の有効性の評価では，従来モデル（Clustered），提案法における GMM（Proposed（GMM）），単一分布近似（Proposed（single））により得られる合成音声の品質を主観評価する．その際に提案法の理想的な場合として，自然音声の特徴量に対して選択された分布（Proposed（target））により得られる合成音声についても評価する．提案法の初期特徴量として，従来モデルで生成した特徴量を用いる．分布系列決定単位の比較では，従来モデルと，提案法において自然音声の特徴量に対してフレーム単位，及び状態単位で選択された分布から得られる合成音声の品質を主観評価する．主観評価は，音質についてのプリファレンステストを行い，ランダムに提示された 2 種類の音声から被験者に音質の高い方を選択させて評価する．被験者は，各実験で男女 7 人とする．

実験結果を Fig.2(a), Fig.2(b) に示す．Fig.2(a) より，提案法により合成音声の品質が改善されることが分かる．また，GMM と単一分布近似に有意差があることから，単一分布近似の有効性が示される．一方で，自然音声の特徴量を用いて選択した分布を用いたスコアには及ばず，最適な分布選択は困難であることが分かる．Fig.2(b) より，決定単位の違いによる有意差はないため，必ずしもフレーム単位で分布を選択する必要性はないことが分かる．



(a) 生成された特徴量　(b) 自然音声の特徴量

Fig. 1　各特徴量に対する出力分布系列の尤度



(a) 手法間の比較　(b) 分布系列決定単位の比較

Fig. 2　音質の主観評価結果

5 まとめ

本稿では，HMM 音声合成の利点を生かしたハイブリッド方式として，分散共有フルコンテキストモデルと尤度に基づくパラメータ生成法を提案し，実験的評価で提案法の有効性を示した．その結果，生成される特徴量に対する尤度基準では最適な分布選択は困難であることが明らかになった．今後は，最適な分布選択を行う基準について検討する．

謝辞　本研究は，（独）情報通信研究機構の委託研究「知識・言語グリッドに基づくアジア医療交流支援システムの研究開発」の一環として実施した．

参考文献

[1] Z.Ling *et al.*, in Proc. of *Blizzard Challenge workshop*, 2007,
[2] Z.Yan *et al.*, *Interspeech* 2009, pp. 1755-1758, 2009.
[3] 吉村 他, 信学論 (D-2), Vol. J83-D-2, pp. 2099-2107, 2000.
[4] K.Tokuda *et al.*, *ICASSP* 2000, Vol. 3, pp. 1315-1318, 2000.
[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.