Vol.2012-SLP-92 No.10
2012/7/20

分散共有フルコンテキストモデルによるHMM音声合成に関する検討

高道 慎之介[1,a)]　戸田 智基[1,b)]　志賀 芳則[2]　河井 恒[3]　Sakriani Sakti[1]　Graham Neubig[1]
中村 哲[1]

概要：隠れマルコフモデル (Hidden Markov Model：HMM) に基づく音声合成において，生成される音声パラメータは過剰に平滑化される傾向にあり，肉声感の低い音声が合成される．この問題を改善するために，HMM 音声合成と素片選択型合成のハイブリッド法がいくつか提案されている．波形素片の使用により，合成音声の音質は顕著に向上するが，同時に HMM 音声合成の利点である音響モデリングの柔軟性が失われる．本稿では，HMM 音声合成の利点を保持した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による音声パラメータ生成法を提案する．提案法では，素片毎の音声パラメータを持つ分散共有フルコンテキストモデルを混合正規分布としてモデル化し，最尤基準に基づいたパラメータ生成アルゴリズムにより音声パラメータを生成する．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．

キーワード：HMM 音声合成，過剰な平滑化，分散共有フルコンテキストモデル，パラメータ生成

A Study on HMM-Based Speech Synthesis Using Rich Context Models

Shinnosuke Takamichi[1,a)]　Toda Tomoki[1,b)]　Shiga Yoshinori[2]　Kawai Hisashi[3]
Sakriani Sakti[1]　Graham Neubig[1]　Nakamura Satoshi[1]

Abstract: In this paper, we propose parameter generation methods using rich context models in HMM-based speech synthesis as yet another hybrid method combining HMM-based speech synthesis and unit selection synthesis. In the traditional HMM-based speech synthesis, generated speech parameters tend to be excessively smoothed and they cause muffled sounds in synthetic speech. To alleviate this problem, several hybrid methods have been proposed. Although they significantly improve quality of synthetic speech by directly using natural waveform segments, they usually lose flexibility in converting synthetic voice characteristics. In the proposed methods, rich context models representing individual acoustic parameter segments are reformed as GMMs and a speech parameter sequence is generated from them using the parameter generation algorithm based on the maximum likelihood criterion. Since a basic framework of the proposed methods is still the same as the traditional framework, the capability of flexibly modeling acoustic features remains. We conduct several experimental evaluations of the proposed methods from various perspectives. The experimental results demonstrate that the proposed methods yield significant improvements in quality of synthetic speech.

Keywords: HMM-based speech synthesis, over-smoothing, rich context model, parameter generation

1. はじめに

テキスト音声合成（Text-To-Speech：TTS）は，任意のテキストから音声を合成する技術である．計算機性能の向上や音声データ規模の拡大に伴って導入されたコーパス

[1] 奈良先端科学技術大学院大学　情報科学研究科
[2] NICT
[3] KDDI
[a)] shinnosuke-t@is.naist.jp
[b)] tomoki@is.naist.jp

ベース方式 [1] の発展により，TTS の高品質化が急速に進んだ．ルールベース方式では必要不可欠であった極めて専門性の高い知識や経験に基づくシステムの最適化の必要性は大幅に下がり，アルゴリズムや実験的評価結果を異なるシステム間で共有することが可能となった．これにより，誰もが TTS を構築できるようになり，汎用性の高い方式が数多く確立されてきた．近年では，自然発話の様な高い音質や多様な発話様式を持つ音声を合成するために，多くの研究が盛んに行われている．

コーパスベース音声合成方式として大きく二つの手法が挙げられる．一つは，素片選択型合成法 [2] に代表されるサンプルベース方式であり，もう一つは統計量ベース方式である．サンプルベース方式では，入力テキストに対して最適な音声波形素片系列を選択し，それを接続することで音声を合成する．素片選択型合成法の最大の利点は，自然音声の波形素片を接続することにより，元の音声の特徴を保持しながら高音質の音声を合成できることである．しかしながら，波形素片の直接的な使用により，合成される音声の特徴は，元の音声の特徴に完全に依存する．

一方，隠れマルコフモデル（Hidden Markov Model：HMM）による音声合成 [3] に代表される統計量ベース方式では，音声コーパスから抽出される複数素片の音声パラメータの統計量を用いて音声波形の合成を行う．この方式では，スペクトル，F_0，継続長を，HMM に基づく統一的な枠組みで同時にモデル化する．合成時には，静的・動的特徴量間の明示的な制約条件の下で，最尤基準に基づいて HMM からこれらのパラメータを生成する．この方式の有用な点は，話者補間 [4] や話者適応 [5]，発話様式制御 [6] に代表されるように，合成音声の特徴を制御できる点である．しかしながら，統計処理により，音声パラメータの詳細な特徴が失われ，HMM から生成される音声パラメータ系列は過剰に平滑化される傾向にある．その結果，自然音声と比較して，肉声感の乏しいこもった音質の音声が合成される．

HMM 音声合成における過剰な平滑化の問題を回避するために，サンプルベース方式とのハイブリッド法が提案されている．例えば，HMM の尤度を最大化するように波形素片を選択する方式 [7] では，波形素片の使用により，HMM 音声合成と比較して合成音声の音質は大幅に改善する．しかし，音声パラメータのモデル化により得られる合成音声の特徴の柔軟な制御は困難となり，HMM 音声合成の利点が失われる．これに対して，比較的柔軟性を保ったハイブリッド方式として，波形素片毎の音声パラメータ（スペクトル，F_0，継続長）を確率分布として保持する分散共有フルコンテキストモデルが提案されている [8]．合成時には，一つの波形素片に対応する全てのパラメータの要素を考慮した確率分布を HMM の状態毎に選択し，選択された確率分布系列から音声パラメータを生成する．この手法でも音質は著しく改善するが，波形素片選択と同様に，確率分布選択時には異なる音声パラメータ間において強い制約を必要とするため，従来の HMM 音声合成が持つ音響モデリングの柔軟性は失われる．

本稿では，HMM 音声合成が持つ音響モデリングの柔軟性を保持した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する．HMM の状態毎に学習された分散共有フルコンテキストモデルを，混合正規分布モデル（Gaussian Mixture Model：GMM）として表現し，尤度最大化基準によるパラメータ生成を行う．従来の HMM 音声合成と同様に，各音声パラメータに対して独立に確率分布を選択することが出来る．提案法の有効性を示すために，様々な観点から実験的評価を行う．

2 節では従来の HMM 音声合成の基本的な枠組みについて触れる．3 節では分散共有フルコンテキストモデルについて解説し，4 節では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示す．5 節では実験的評価を行いその結果を示す．6 節では本稿のまとめについて述べる．

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，音素の様な分節的特徴や，文全体にわたるような超分節的特徴や韻律の特徴を捉えるために多様なコンテキスト要因が用いられる．それら多数のコンテキスト要因の組み合わせにより，各音声素片に対するコンテキスト（フルコンテキスト）が表現されるため，その数は指数的に増加し，天文学的な数字となる．全てのフルコンテキストをカバーする音声データの収集は不可能であり，各フルコンテキストはしばしば一つの音声素片にのみ対応することになる．そのため，各フルコンテキストに依存した HMM（フルコンテキストモデル）では，過学習の問題が生じ，また，未知のフルコンテキストへの対応が出来ない．そこで，各コンテキストに対する質問で構成される決定木でフルコンテキストモデルをクラスタリングして，各クラスで出力確率密度関数を共有する [9]．ここで，クラス c の出力確率密度関数 b_c は次式でモデル化される．

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

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．コンテキストクラスタリングでは，HMM の状態毎及び音声パラメータ毎にクラスを決定する．

合成時には，入力テキストのフルコンテキストに対するクラスを HMM 状態毎に決定し，文 HMM を形成するために各クラスに対応する出力確率密度関数が選択される．そして，静的・動的特徴量間の明示的な制約条件 ($\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$) の下で，HMM の尤度を最大化するようにパラメータ系列

$\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \cdots, \boldsymbol{c}_T^\top\right]^\top$ を生成する [10]．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定される行列である．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，合成音声の著しい劣化を生じさせる．

3. 分散共有フルコンテキストモデル

従来の方法では，決定木の同じクラスに属する複数の音声素片をモデル化するために，単一の正規分布が使われる．故に，その平均ベクトルは平滑化され，音声パラメータを過剰に平滑化させる要因となっている．一方，複数の音声素片を用いた正規分布の学習は，その共分散行列の頑健な推定に必要不可欠である．パラメータ推定の頑健性を保持しつつ過剰な平滑化を改善する方法として，各クラスに属する異なるフルコンテキストラベルで共分散行列を共有して，平均ベクトルのみフルコンテキストに依存させる分散共有フルコンテキストモデルが提案されている [8]．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 $\boldsymbol{\Sigma}_c$ を持つ正規分布により，次式で示される．

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

異なる平均ベクトルの総数は，学習データに現れるフルコンテキストの数と一致する．また，異なる共分散行列の数は，決定木のクラスの数と一致する．

学習時には，まず従来の方法により，コンテキストクラスタリングによるモデルの確率密度関数のパラメータを学習する．次に，そのモデルを用いて forward-backward アルゴリズムにより十分統計量を計算した後に，全てのフルコンテキストに対して平均ベクトルのみを更新し，共分散行列は各クラスのものを保持することで，分散共有フルコンテキストモデルを構築する．合成時には，コンテキストクラスタリングによるモデルとの Kullback-Leibler Divergence （KLD）を最小化する分散共有フルコンテキストモデルを選択する．その際には，異なる音声パラメータにおいて，同一の音声素片に対する確率密度関数が選択するという制約を設ける．最終的に，選択された確率密度関数から音声パラメータを生成する．

4. 分散共有フルコンテキストモデルによる合成法

本稿では，分散共有フルコンテキストモデルの尤度最大化基準に基づくパラメータ生成法を提案する．提案法は，各音声パラメータに対して，独立にモデル選択を行いパラメータ生成を行うといった従来の HMM 音声合成の枠組みを保持している．これにより，**3.** で示した従来法と比較し，より柔軟な音声合成処理が実現できると期待される．

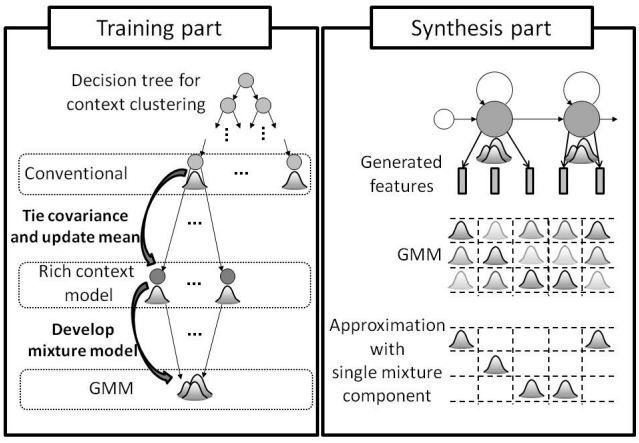


図 1 提案法における学習と生成処理
Fig. 1 Training and synthesis processes in proposed methods

学習と生成処理を図 1 に示す．

4.1 GMM の計算

従来と同じ方法で分散共有フルコンテキストモデルを学習した後，各クラスに属する全ての分散共有フルコンテキストモデルを用いて構築される GMM により出力確率密度関数をモデル化する．

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

ただし，ω_m は m 番目の混合要素の重みであり，各クラスの要素数は M_c とする．混合分布における重みは，forward-backward アルゴリズムにより得られる状態滞在確率に基づいて推定できる．しかしながら，予備実験の結果，等重み（$\omega_m = 1/M_c$）にすることで合成音声の音質が改善する傾向が得られたため，本稿では等重みを用いる．

4.2 パラメータ生成法

従来の HMM 音声合成と同様に，状態継続長モデルに基づき，HMM 状態系列 $\boldsymbol{q} = [q_1, \cdots, q_T]^\top$ を決定する．このとき，HMM の尤度は次式で示される．

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

ただし，混合要素系列（分散共有フルコンテキストモデル系列）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．静的特徴量系列は，$\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$ の制約の下で HMM の尤度を最大にするように決定する．これは，従来の HMM 音声合成のパラメータ生成処理と同様である [10]．

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

以下では，EM アルゴリズムと単一分布近似を用いる方法について述べる．

4.2.1 EM アルゴリズム

$\boldsymbol{c}$ の最尤推定値は，EM アルゴリズム（Expectation-Maximization algorithm）により決定される．まず，初期パラメータ系列 $\boldsymbol{c}^{(0)}$ を決定する．そして，E ステップで，現在の推定値 $\boldsymbol{c}^{(i)}$ から与えられる事後確率 $P\left(\boldsymbol{m}|\boldsymbol{W}\boldsymbol{c}^{(i)},\boldsymbol{q},\boldsymbol{\lambda}\right)$ を計算し，M ステップで，次に示す補助関数を最大にするように新たな推定値 $\boldsymbol{c}^{(i+1)}$ を決定する．EM アルゴリズムでは，これらのステップを反復的に行う．

$$Q\left(\boldsymbol{c}^{(i)},\boldsymbol{c}^{(i+1)}\right) = \sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{m}|\boldsymbol{W}\boldsymbol{c}^{(i)},\boldsymbol{q},\boldsymbol{\lambda}\right) \ln P\left(\boldsymbol{W}\boldsymbol{c}^{(i+1)},\boldsymbol{m}|\boldsymbol{q},\boldsymbol{\lambda}\right) \quad (6)$$

4.2.2 単一分布近似

式 (4) で表された HMM の尤度を単一の混合要素系列で近似する．

$$\sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{o},\boldsymbol{m}|\boldsymbol{q},\boldsymbol{\lambda}\right) \simeq P\left(\boldsymbol{o},\boldsymbol{m}|\boldsymbol{q},\boldsymbol{\lambda}\right) \quad (7)$$

初期パラメータ系列 $\boldsymbol{c}^{(0)}$ を決定した後，単一混合要素系列及び，静的パラメータ系列は次式にて反復的に更新する．

$$\hat{\boldsymbol{m}}^{(i+1)} = \underset{\boldsymbol{m}}{\operatorname{argmax}} P\left(\boldsymbol{m}|\boldsymbol{W}\boldsymbol{c}^{(i)},\boldsymbol{q},\boldsymbol{\lambda}\right), \quad (8)$$

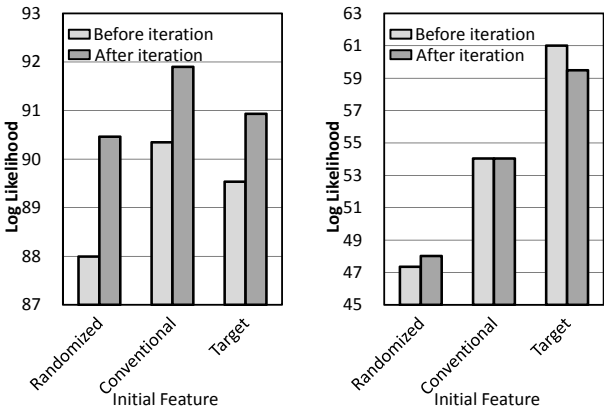
$$\hat{\boldsymbol{c}}^{(i+1)} = \underset{\boldsymbol{c}}{\operatorname{argmax}} P\left(\boldsymbol{W}\boldsymbol{c}|\hat{\boldsymbol{m}}^{(i+1)},\boldsymbol{q},\boldsymbol{\lambda}\right) \quad (9)$$

4.2.3 考察

各分散共有フルコンテキストモデルは，しばしば一つの音声素片のみに対応する．したがって，提案法の処理は素片選択型合成法に強く関係する．提案法において，静的特徴量と動的特徴量に対する HMM の尤度は，それぞれ素片選択におけるターゲットコストと接続コストとみなすことができ [11]，ターゲットコストと接続コストが最小になるように，音声素片を選択していることに相当する．EM アルゴリズムによる合成処理は，音声パラメータを生成するために，複数の音声素片を選択して融合する処理と類似している [12]．一方で，単一分布近似による合成処理は，音声パラメータ生成のために単一音声素片系列を選択する処理と類似している [2]．

提案法は，分散共有フルコンテキストモデルによる従来手法とは異なり，分布選択時において異なる音声パラメータ間に強い制約を使用しない．したがって，提案法は HMM 音声合成における音響モデリングの柔軟性を保持している．例えば，異なる音声パラメータの要素で別々の分散共有フルコンテキストモデルを選択することが可能である．また，音声パラメータの要素毎に別々のパラメータ生成法を適用することも可能である．

4.2 で述べた方法では，分散共有フルコンテキストモデルをフレーム毎に選択することに相当するが，同一の HMM 状態で同一の分散共有フルコンテキストモデルを選択する制約を加えることで，状態毎の選択処理も実現できる．な



(a) 生成された特徴量系列に対する対数尤度　(b) 自然音声の特徴量系列に対する対数尤度

図 2 初期特徴量系列に対する依存性
Fig. 2 Effect of the initial parameter sequence

お，提案法では初期特徴量系列を決定する必要がある．いくつかの方法が考えられるが，単純な方法の一つは，コンテキストクラスタリングによる従来法で生成されたパラメータ系列を使用することである．

5. 実験的評価

5.1 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 [13] A-I セット 450 文とする．評価データは同 J セット 53 文を使用する．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 [14] 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．5 状態 left-to-right 型の隠れセミマルコフモデル（Hidden Semi-Markov Model：HSMM）の学習を行い，パラメータ生成時には系列内変動（Global Variance: GV）[15] を考慮しない．コンテキストクラスタリングは最小記述長（Minimum Description Length：MDL）基準 [16] によって行う．

まず，自然音声の継続長が与えられた下で，様々な観点から提案法の有効性を明らかにする．自然音声の継続長は，コンテキストクラスタリングによるモデルを用いて，自然音声に対して Viterbi アライメントを行うことで求める．次に，継続長を含んだ全音声パラメータの合成処理において，提案法の有効性を評価する．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．音源特徴量と継続長に対してはコンテキストクラスタリングによる従来法を適用する．

5.2 自然音声の継続長における評価

5.2.1 初期特徴量系列に対する依存性

提案法の初期特徴量系列に対する依存性を調査するために，3 種類の初期特徴量系列として，1) 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された分布から生成した特徴量（Randomized），2) 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），3) 自然音声の特徴量（Target）を用いる．提案法は，単一分布近似を用いる．初期特徴量系列により選択された分散共有フルコンテキストモデル系列（Before iteration）と反復処理により最終的に選択された系列（After iteration）を，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．ここで，対数尤度は評価文の対数尤度をフレーム数で割ったものを表し，この値が大きいほど適切な系列が選択されていることを表す．また，提案法の目的関数である，生成された特徴量系列に対する尤度についても合わせて評価する．

実験結果を，図 2 に示す．図 2(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近い値まで上昇している．一方，図 2(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，初期特徴量系列に大きく依存する．これらの結果より，初期特徴量系列の設定は本質的に重要であり，生成される特徴量系列に対する尤度基準では，最適な分散共有フルコンテキストモデル系列を決定することは困難であることが分かる．なお，生成された特徴量系列に対する尤度と比較し，自然音声の特徴量系列に対する尤度は明らかに小さいことが分かる．

5.2.2 提案法の有効性

提案法を評価するために，コンテキストクラスタリングによる従来モデル（Conventional），EM アルゴリズムによる提案法（Proposed（GMM)），単一分布近似による提案法（Proposed（single)），理想値として，自然音声の特徴量系列から選択された単一分布系列（Proposed（target)）により合成された音声を比較する．提案法の初期パラメータ系列は，従来モデルで生成した特徴量を用いる．音質に関するプリファレンステスト（AB テスト）を実施し，受聴者には 4 手法により生成された音声の全ての組み合わせを受聴させ，音質の良い方を選択させる．受聴者は男女 7 人とする．

実験結果を，図 3(a) に示す．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．更に，単一分布近似の使用は，EM アルゴリズムの使用よりも音質が高い事が示される．また，**5.2.1** で示したように，最適な単一分布系列（Proposed(target)）を選択することは，尤度基準では困難であることが分かる．

5.2.3 単一分布系列選択単位の比較

提案法における分布系列の選択単位が合成音声に与える影響を調査するため，コンテキストクラスタリングによる従来法（Conventional），フレーム単位で選択した単一分布系列による提案法（Frame-based），HMM 状態単位で選択した単一分布系列による提案法（State-based）を比較する．提案法における選択処理では，ターゲットとして自然音声の特徴量を使用する．なお，選択単位の違いにより，異なる分布が選択されることを確認するために，各選択単位において選ばれた分布系列を図 4 に示す．ただし，状態毎に混合数（分散共有フルコンテキストモデル数）が異なるため，縦軸は混合数で正規化した要素番号を表している．

実験結果を，図 3(b) に示す．フレーム単位と状態単位の選択に大きな差はなく，状態単位の選択もまた，合成音声の音質改善に効果的であることが分かる．この結果と，図 2 から，提案法では適切な初期値に基づき状態毎に選択処理を行えば良いと予想される．

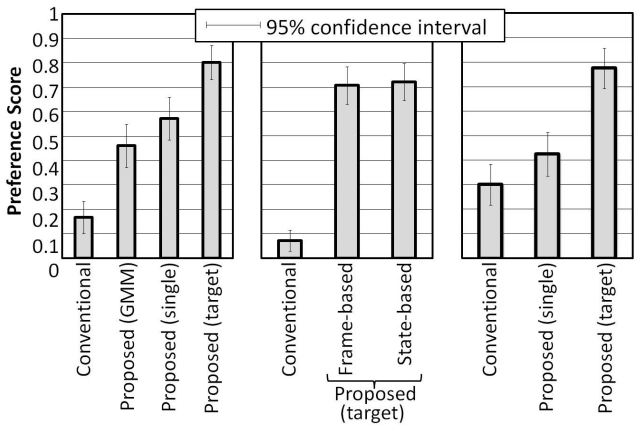


図 3 音質に関する主観評価結果

Fig. 3 Preference scores for speech quality

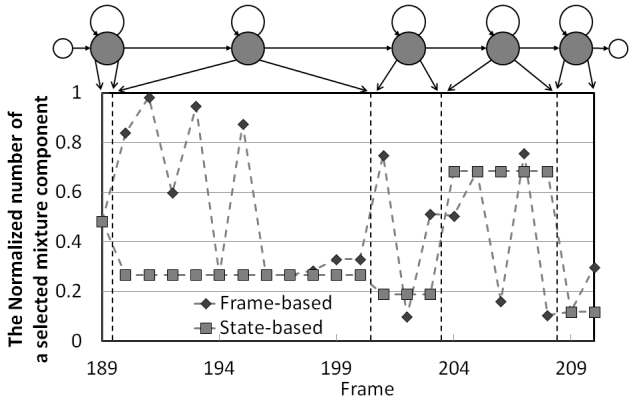


図 4 フレーム単位，状態単位で選択された単一分布系列の例

Fig. 4 An example of selected mixture component (Frame-based, State-based)

5.3 全パラメータを生成した場合の提案法の有効性

実際の TTS と同様に，継続長を含む全てのパラメータを合成する条件下での，提案法の有効性を確認するために，コンテキストクラスタリングによる従来法（Conventional），状態毎に選択した単一分布近似による提案法（Proposed(single)），理想値として，自然音声の特徴量を使って状態毎に選択した単一分布系列（Proposed(target)）

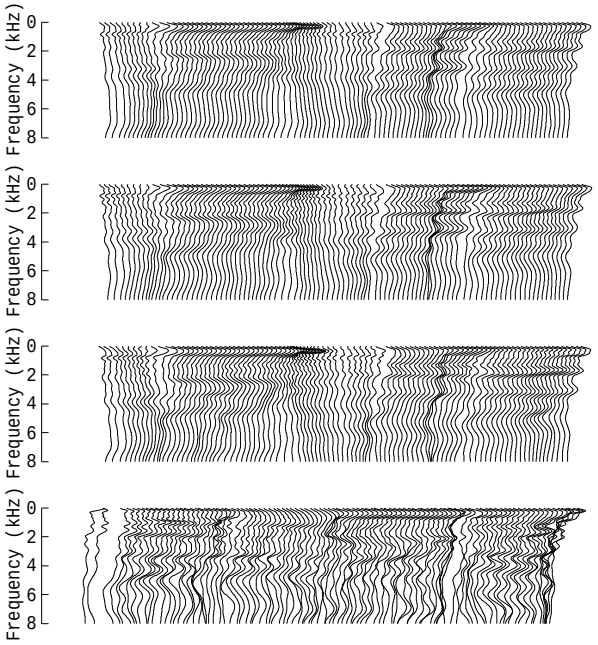


図 5 スペクトログラム（上から，Conventional, Proposed(single), Proposed(target), 自然音声を表す）

Fig. 5 Spectrogram(representing Conventional, Proposed(single), Proposed(target), and natural speech from above.)

を比較する．提案法の初期特徴量系列は，従来法で生成された特徴量系列を使用する．評価として，**5.2.2** と同様の条件でプリファレンステストを行う．

実験結果を，図 3(c) に示す。また，Conventional, Proposed(single), Proposed(target), 自然音声のスペクトログラムを，図 5 に示す．全音声パラメータを合成した場合においても，提案法により合成音声の音質は顕著に改善することが分かる．一方で，単一分布系列による提案法（Proposed(single)）と，自然音声の特徴量を使って選択した単一分布系列（Proposed(target)）は，図 3(a) と図 3(c) で差が大きくなっていることが分かる．このことから，最尤基準による分布選択では，継続長の影響を大きく受けることが分かる．

6. まとめ

本稿では，音声パラメータを柔軟にモデル化する 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た．提案法では，分散共有フルコンテキストモデルにより各音声素片における個々の音声パラメータを柔軟にモデル化するとともに，それらを GMM として表現することで，最尤基準に基づくパラメータ生成法を実現した．様々な観点から実験的評価を行うことで，提案法の有効性を示した．一方で，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] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," Proc.ICASSP, pp.679–682, 1988.

[2] N. Iwahashi, N. Kaiki, Y. Sagisaka, "Speech segment selection for concatenative synthesis based on spectral distortion minimization," IEICE Trans, Fundamentals, Vol.E76-A, No.11, pp.1942–1948, 1993.

[3] H. Zen, K. Tokuda, A. Black, "Statistical parametric speech synthesis," Speech Commun., Vol.51, No.11, pp.1039–1064, 2009.

[4] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, T. Kitamura, "Speaker interpolation for HMM-based speech synthesis system", J. Acoust. Soc. Jpn. (E), Vol.21, No.4, pp.199–206, 2000.

[5] J. Yamagishi, T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," IEICE Trans. on Inf. and Syst., Vol.E90-D, No.2, pp.533–543, 2007.

[6] T. Nose, J. Yamagishi, T. Masuko, T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," IEICE Trans. Inf. and Syst., Vol.E90-D, No.9, pp.1406–1413, 2007.

[7] Z. Ling, L. Qin, H. Lu, Y. Gao, L. Dai, R. Wang, Y. Jiang, Z. Zhao, J. Yang, J. Chen, G. Hu, "The USTC and iflytek speech synthesis systems for Blizzard Challenge 2007," Proc. of Blizzard Challenge workshop, 2007.

[8] Z. Yan, Q. Yao, S.K. Frank, "Rich Context Modeling for High Quality HMM-Based TTS," INTERSPEECH 2009, pp.1755–1758, 2009.

[9] 吉村 貴克, 徳田 恵一, 益子 貴史, 小林 隆夫, 北村 正, "HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化," 信学論（D-2）, Vol.J83-D-2, No.11, pp.2099–2107, 2000.

[10] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," Proc.ICASSP, pp.1315–1318, 2000.

[11] S. Kataoka, N. Mizutani, K. Tokuda, T. Kitamura, " Decision tree backing-off in HMM-based speech synthesis, " INTERSPEECH 2004, WeB1403p.12, 2004.

[12] T. Mizutani, T. Kagoshima, "Concatenative Speech Synthesis Based on the Plural Unit Selection and Fusion Method", IEICE Trans. on Inf. and Syst., Vol. E88-D, No.11, pp.2565–2572, 2005.

[13] 阿部 匡伸, 匂坂 芳典, 梅田 哲夫, 桑原 尚夫, "研究用日本語音声データベース利用解説書（連続音声データ編）," ATR テクニカルレポート, TR=1-0166, 1990.

[14] H. Kawahara, I. Masuda-Katsuse, A.D. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in souds," Speech Commun., Vol.27, No.3–4, pp.187–207, 1999.

[15] T. Toda, K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," IEICE Trans, Vol.E90-D, No.5, pp.816–824, 2007.

[16] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," J.Acoust.Soc.Jpn.(E), Vol.21, No.2, pp.79–86, 2000.