社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

分散共有フルコンテキストモデルによるHMM音声合成の改善

高道慎之介† 戸田 智基† 志賀 芳則†† Sakriani Sakti† Graham Neubig†
中村 哲†

† 奈良先端科学技術大学院大学　〒 630–0192 奈良県生駒市高山町 8916–5
†† 情報通信研究機構　〒 619–0289 京都府相楽郡精華町光台 3–5
E-mail: †{shinnosuke-t,tomoki}@is.naist.jp

あらまし　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．これに対して我々は，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パラメータ生成法を提案している．最尤基準による反復的なパラメータ生成処理により，時間的な不連続性を緩和できる一方で，生成されるパラメータ及び合成音声の音質は，反復処理における初期パラメータに大きく依存する．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生成法を提案する．提案法では，過剰な平滑化の影響が小さい初期パラメータを生成するために大きなサイズの決定木を構築し，従来の HMM 音声合成のパラメータ生成法に従い初期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, 分散共有フルコンテキストモデル, パラメータ生成, コンテキストクラスタリング

Improvements of HMM-based Speech Synthesis Using Rich Context Models

Shinnosuke TAKAMICHI†, Tomoki TODA†, Yoshinori SHIGA††, Sakriani SAKTI†, Graham NEUBIG†, and Satoshi NAKAMURA†

† Nara Institute of Science and Technology,　Tatayama-cho 8916–5, Ikoma, Nara, 630–0192 Japan
†† National Institute of Information and Communications Technology,　Hikari-dai 3–5, Seika-cho, Soraku-gun, Kyoto, 619–0289, Japan
E-mail: †{shinnosuke-t,tomoki}@is.naist.jp

Abstract　In the traditional HMM-based speech synthesis, generated speech parameters tend to be excessively smoothed. To alleviate this problem, we have proposed a parameter generation method with rich context models in our previous work. This method improves speech quality while keeping the flexibility of HMM-based speech synthesis. However, synthetic speech still sounds muffled because the generated parameters strongly depend on over-smoothed initial parameters in iterative parameter generation procedure. In this paper, we propose an initialization method for generating less-smoothed initial parameters using context-clustered HMMs based on a large-sized decision tree. Experimental evaluations of the proposed method demonstrate that the proposed method yields significant improvements in the quality of synthetic speech.
Key words　HMM-based speech synthesis, rich context model, parameter generation, tree-based context clustering

1. ま え が き

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ベース方式 [1] の発展により，TTS の高品質化が急速に進み，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，図 1 に示すように，素片選択型合成法 [2] に代表されるサンプルベース方式，隠れマ

ルコフモデル（Hidden Markov Model：HMM）による音声合成 [3] に代表される統計的パラメトリック合成方式が挙げられる．素片選択型合成法では，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．自然音声の波形素片の直接的な使用により，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半面，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，HMM 音声合成では，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．合成音声の声質・発話様式制御 [4]〜[6] に代表されるような高い柔軟性を備えている反面，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，合成音声の肉声感が損なわれる傾向にある．

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．素片選択型合成法から派生したハイブリッド法の代表的な例として，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] が挙げられる．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反面，音声パラメータの音響モデリングの柔軟性は失われる．一方，HMM 音声合成から派生したハイブリッド法の代表的な例として，波形素片毎の音声パラメータ要素（スペクトル，F_0，継続長）を確率密度関数として保持する分散共有フルコンテキストモデルを，全音声パラメータ要素を考慮して選択する方式 [8] が挙げられる．確率密度関数の使用により，比較的柔軟性を保持して合成音声の音質は改善する反面，異なる音声パラメータ要素間の強い制約により，音響モデリングの柔軟性は失われる．

これに対して我々は，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している [9]．分散共有フルコンテキストモデルは，混合正規分布モデル（Gaussian Mixture Model：GMM）として表現され，合成時には，初期パラメータ系列を決定したのち，尤度最大化基準による反復的なパラメータ生成を行う．反復処理により生成されるパラメータ系列は初期パラメータ系列に大きく依存ものの，HMM 音声合成により生成したパラメータ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確認されている [9]．しかし，過剰な平滑化の影響は未だ大きく，その音質は十分ではない．この原因の一つとして，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．従って，過剰な平滑化の影響が少ない初期パラメータ系列を使用することで，更なる音質改善が得られると思われる．

本稿では，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おける初期パラメータ決定法として，コンテキストクラスタリングを用いた手法を提案する．コンテキストクラスタリングによる大きなサイズの決定木を構築し，従来の HMM 音声合成のパラメータ生成法に従って初期パラメータを生成する．この決定木から生成されるパラメータは過剰な平滑化の影響が少ないため，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法により最終的に生成されるパラメータも同様の影響を受ける．提案法の有効性を示すために，実験的評価を行う．

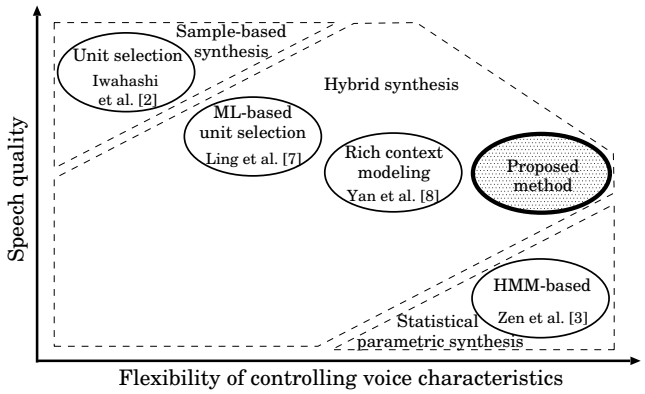

図 1　コーパスベース方式 [1] の関係図
Fig. 1　Overview of the corpus-based speech synthesis methods

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を用いたパラメータ生成法について解説し，4 節ではコンテキストクラスタリングによる初期パラメータ決定法を示す．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. 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，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（韻律的特徴）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，クラス毎に出力確率密度関数を正規分布でモデル化・共有する [10]．

学習時には，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 [11]．

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

ただし，c はクラス番号，C は総クラス数，a は C を制御するパラメータ，D は特徴量の次元数，$\mathbf{\Sigma}_c$ 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ノードにおける状態占有確率の総和を表す．総クラス数の増加に従い式 (1) 右辺の第一項の値は減少し，第二項の値は増加する．記述長を最小にする総クラス数 C は a によって変化し，

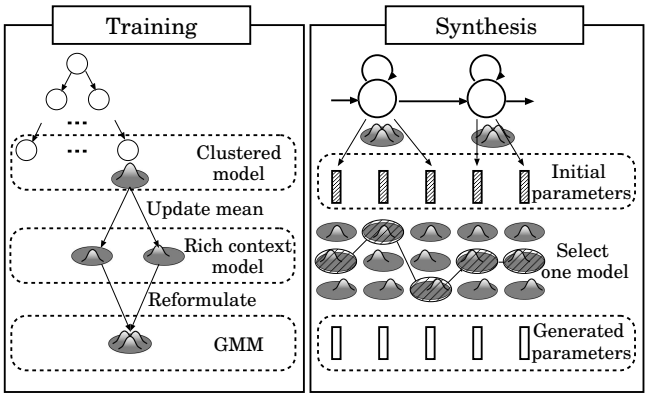

図 2　分散共有フルコンテキストモデルによる学習部及び生成部
Fig. 2　Training and synthesis processes with rich context models

a が小さいほど C が増加する．コンテキストクラスタリングで得られるクラス c の出力確率密度関数 b_c（状態共有モデル）は，次式でモデル化される．

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

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．コンテキストクラスタリングでは，HMM 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 [12]．ここで，特徴量系列を $\boldsymbol{o} = \left[\boldsymbol{o}_1^\top, \cdots, \boldsymbol{o}_T^\top\right]^\top$ 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を用いたパラメータ生成法

3.1 分散共有フルコンテキストモデルを用いた GMM の構築

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 [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 $\mathcal{N}\left(\cdot; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c\right)$ により，次式で示される．

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

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．

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．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．

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

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

3.2 パラメータ生成法

まず，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．

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

ただし，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．

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

$$\hat{\boldsymbol{c}}^{(i+1)} = \underset{\boldsymbol{c}}{\operatorname{argmax}}\, P\left(\boldsymbol{W}\boldsymbol{c}|\hat{\boldsymbol{m}}^{(i+1)}, \boldsymbol{q}, \boldsymbol{\lambda}\right) \tag{7}$$

3.3 初期パラメータ系列の影響

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，このパラメータ生成法の処理は素片選択型合成法に強く関係する．生成時において考慮される静的特徴量と動的特徴量に対する HMM 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 [13]，式 (6) は，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．一方で，反復処理により最終的に選択されるモデル及び生成パラメータ系列は，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 [9]，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．この原因の一つとして，過剰に平滑化された初期パラメータ系列の使用が影響していると考えられる．

4. 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 を小さくすることで，図 3 に示すようにクラス数（サイズ）を大きくした決定木を構築する．この決定木では，コンテキスト要因により決定した少数の音声素片から，クラス毎のモデルが計算されるため，状態共有モデルを用いた場

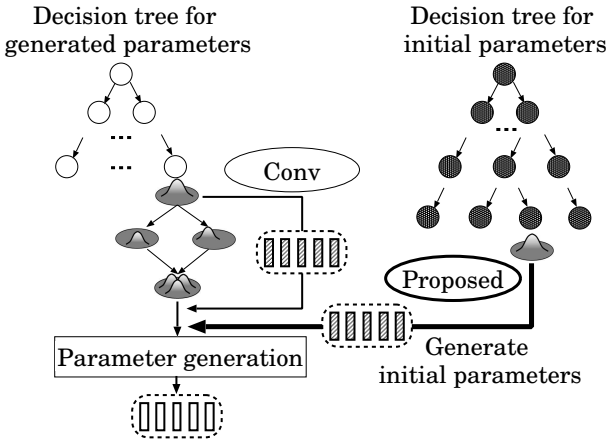

図 3　提案法の概要図

Fig. 3　Overview of the proposed method

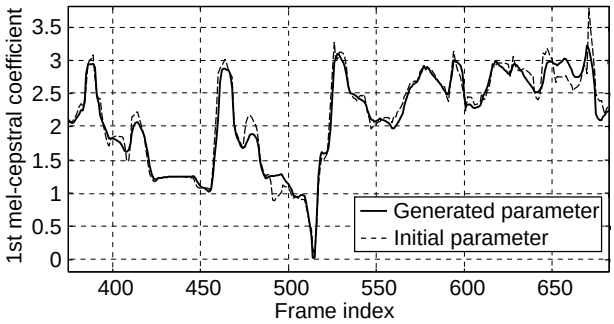

図 4　初期パラメータ系列と生成パラメータのメルケプストラム系列

Fig. 4　1st mel-cepstral coefficient sequences of initial parameter and generated parameter

合と比較して，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，頑健性に乏しいモデルの学習と，コンテキスト要因のみを考慮したモデル選択により，この決定木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，音響的要因である静的・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．大きな決定木から生成された初期メルケプストラム系列と，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生成されたメルケプストラム系列の例を図 4 に示す．初期パラメータ系列が持つ時間的な不連続性を緩和していることが確認される．

初期パラメータ生成用の決定木は，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．なお，この十分統計量は，分散共有フルコンテキストモデルの決定木を構築する際に用いるものとは異なる．

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 [14] 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 [15] 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 [16] 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）[17] の学習を行い，パラメータ生成時には系列内変動（Global Variance: GV）[18] を考慮しない．クラスの共分散行列 $\mathbf{\Sigma}_c$ は対角共分散行列を使用する．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．

まず予備実験として，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により，提案法で生成した初期パラメータ系列の時間的な不連続性が緩和されることを確認する．評価基準は生成パラメータに対する HMM の対数尤度であり，初期パラメータ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と反復処理により最終的に選択された系列において計算される．ただし，HMM の対数尤度は静的・動的特徴量毎に計算する．評価に用いる初期パラメータ系列は，提案法の $a = 0.1$ の決定木によるモデルを用いて生成した特徴量（Proposed），標準的な状態共有モデルで生成した特徴量（Conv），自然音声の特徴量（Target）の 3 種類とする．

次に，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 の対数尤度及び，生成パラメータに対する GV の対数尤度 [18] とする．ただし，HMM 状態系列 $\boldsymbol{q}$ は，標準的な状態共有モデルを用いて，自然音声に対して Viterbi アライメントを行うことで求める．また，GV の対数尤度は，自然音声のパラメータを用いて推定される確率密度関数より計算する．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（Proposed）において行う．表 1 に，各パラメータ a における決定木のサイズを示す．また，比較のため，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．

最後に，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，音源特徴量及び継続長は標準的な状態共有モデルを用いる．評価手法として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

表 1　決定木のサイズ（総クラス数とフルコンテキストモデルサイズに対する圧縮率）
Table 1　Decision tree size (total number of leaf nodes and compression rate for the size of full context models)

a	C	Tree size (%)
0.1	24212	89.3
0.2	8602	31.7
0.3	2297	8.47
0.4	868	3.20
0.5	366	1.35
0.6	264	0.97
0.7	207	0.76
0.8	178	0.65
0.9	160	0.59
1.0	151	0.55

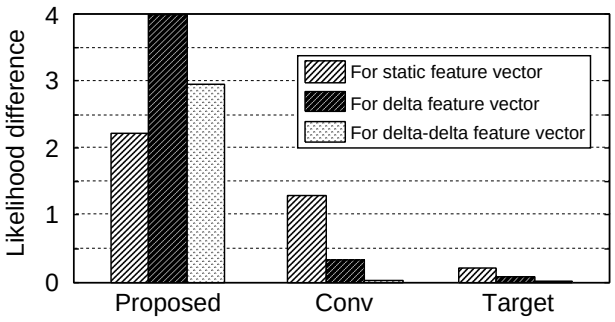

図 5　反復処理による尤度の差
Fig. 5　Likelihood differences through iteration

質の良い方を選択させる．受聴者は男女 7 人とする．

5.2 実験結果

反復処理による，生成パラメータに対する HMM の対数尤度の差を図 5 に示す．図 5 から，他手法と比較して提案法では，反復処理により動的特徴量に対する尤度が大きく上昇していることから，時間的な不連続性の緩和が確認される．また，それぞれの特徴量に対する尤度において，他手法では静的特徴量に対する尤度が最も上昇している一方，提案法では静的特徴量に対する尤度の上昇は最も小さい事が確認される．

HMM の対数尤度を図 6 に，GV の対数尤度を図 7 に示す．図 6 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．また，$a = 0.5$ の場合においても，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，図 7 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．

音質の主観評価結果を図 8 に，各手法による合成音声及び自然音声のスペクトログラムを図 9 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．この結果と図 7 から，提案法における合成音声の音質と GV 尤度との間には，比較的高い相関があることが分かる．

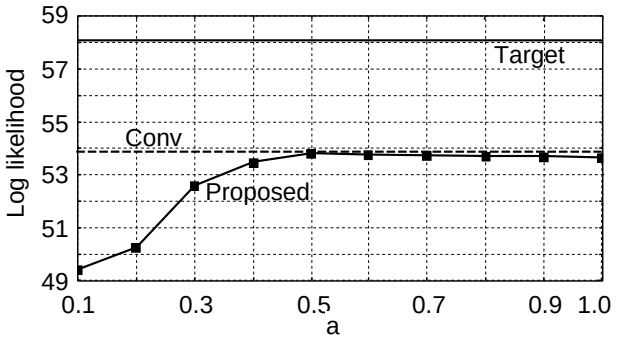

図 6　自然音声の特徴量に対する HMM 対数尤度
Fig. 6　Log-scaled HMM likelihood for natural speech parameter

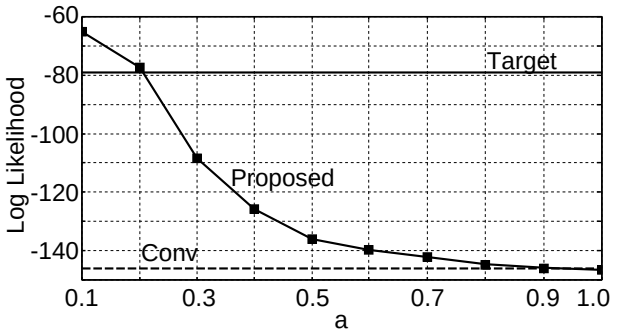

図 7　生成パラメータに対する GV 尤度
Fig. 7　Log-scaled GV likelihood for generated parameter

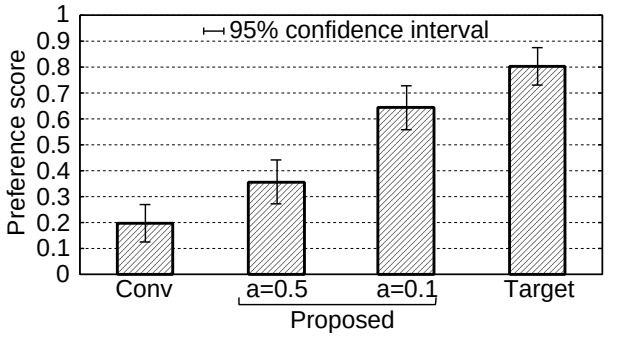

図 8　音質に関する主観評価結果
Fig. 8　Preference score for speech quality

6. まとめ

本稿では，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の音質改善を目的として，コンテキストクラスタリングによ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．

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

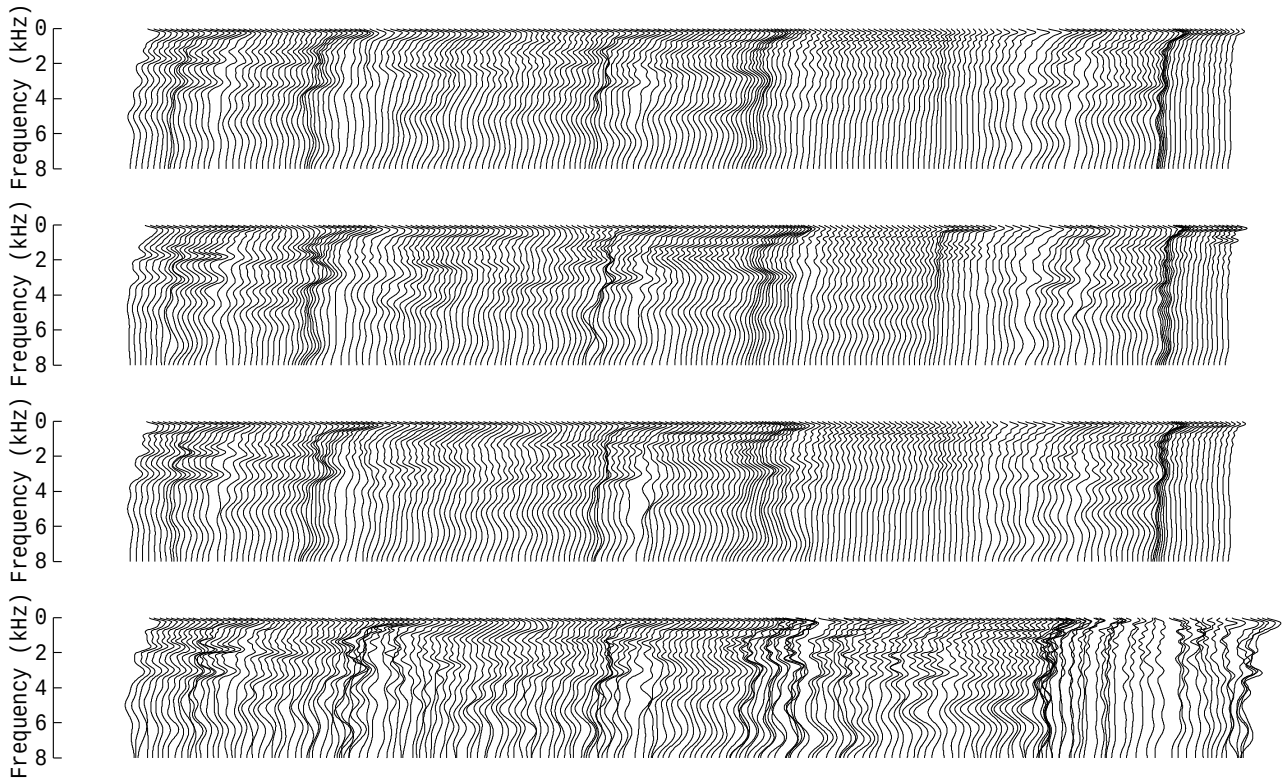

図 9　スペクトログラム（上から，Conv, Proposed ($a = 0.1$), Target, 自然音声を表す）
Fig. 9　Spectrogram (representing Conv, Proposed ($a = 0.1$), Target, and natural speech from top down.)

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] S. Takamichi, T. Toda, Y. Shiga, H. Kawai, S. Sakti, and S. Nakamura, "An Evaluation of Parameter Generation Methods with Rich Context Models in HMM-Based Speech Synthesis," INTERSPEECH2012, 2012.

[10] 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.

[11] 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.

[12] 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.

[13] S. Kataoka, N. Mizutani, K. Tokuda, T. Kitamura, " Decision tree backing-off in HMM-based speech synthesis, " INTERSPEECH2004, 2004.

[14] 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.

[15] 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 sounds," Speech Commun., Vol.27, No.3–4, pp.187–207, 1999.

[16] H. Kawahara, Jo Estill and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT ", MAVEBA 2001, 2001.

[17] S. E. Levinson, "Continuously variable duration hidden markov models for automatic speech recognition," Computer Speech and Language, pp.29–45, 1986.

[18] 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.