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.

HMM音声合成における 分散共有フルコンテキストモデルによるF0パターン生成

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，合成音声の肉声感は劣化する．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おり，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．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．F_0 のモデル化に広く用いられる多空間確率分布 HMM （Multi-Space probability Distribution HMM： MSD-HMM）を用いて分散共有フルコンテキストモデルを構築し，F_0 パターンを生成する．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 F_0 パターン生成，MSD-HMM，分散共有フルコンテキストモデル，パラメータ生成法

F0 Contour Generation Using Rich Context Models in HMM-Based Speech Synthesis

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 over-smoothed. To alleviate this problem, we have proposed a spectral parameter generation method with rich context models and have showed its effectiveness. In this paper, we propose a F_0 contour generation method with the rich context models, which are successfully applied to Multi-Space probability Distribution HMM (MSD-HMM) for modeling F_0 contour. Experimental evaluations demonstrate that the proposed method yields significant improvements in the quality of synthetic speech.
Key words F_0 contour generation, MSD-HMM, rich context models, parameter generation method

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．計算機性能の向上に伴い導入されたコーパスベース方式 [1] には，図 1 に示すように，素片選択型合成法 [2], [3] 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 [4] に代表される統計的パラメトリック合成方式が存在する．入力テキストに対して最適な音声波形素片系列を選択・接続することで音声を合成する素片選択型合成法は，自然音声の波形素片の直接的な使用により高音質の音声を合成できる [5] 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，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を合成する HMM 音声合成は，合成音声の声質・発話様式制御 [6]〜[8] による柔軟な声質制御が可能である半面，統計処理により，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 [9].

素片選択型合成法における素片選択用コスト関数設計の困難さや，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，HMM 音声合成と素片選択型合成のハイブリッド法が提案されている [10]〜[12]．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 [10] が挙げられる．自動学習される HMM の尤度に基づくコスト計算により，合成音声の音質は大幅に改善する反面，波形素片の使用により，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 [11] が挙げられる．確率密度関数に基づくパラメータ生成法により，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，パラメータ生成時に用いる異なる音声パラメータ要素間の強い制約により，HMM 音声合成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 [13]．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，コンテキストクラスタリングによる初期パラメータ生成法 [14] により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．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有効性が，スペクトルパラメータにおいて示されているため，他の音声パラメータにおいても同様に有効性が期待される．

本稿では，分散共有フルコンテキストモデルにより F_0 パターンを生成する手法を提案する．F_0 のモデル化に広く用いられる多空間確率分布 HMM（MSD-HMM）[15] に対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適用し，F_0 パターンを生成する．提案法の有効性を示すために，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 節ではスペクトルパラメータ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を用いた F_0 パターン生成法を示し，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

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依存した HMM（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 [16].

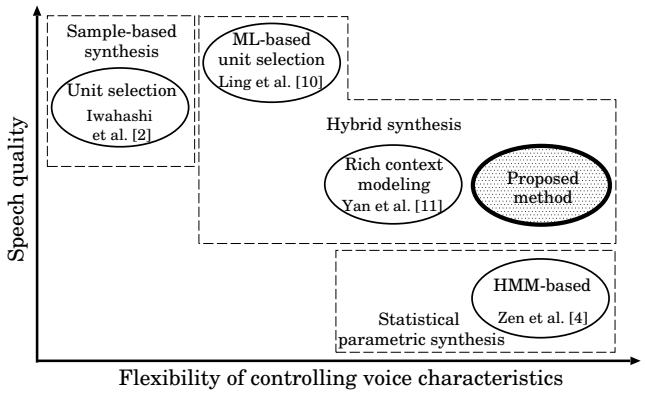

図 1　コーパスベース方式 [1] の関係図
Fig. 1　Overview of the corpus-based speech synthesis methods.

クラスタリング時には，次式で表される最小記述長（Minimum Description Length: MDL）基準を用いる [17].

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

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，$\boldsymbol{\Sigma}_c$ 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ノードにおける状態占有確率の総和を表す．コンテキストクラスタリングは HMM 状態毎及び音声パラメータ毎に行われ，クラス毎に出力確率密度関数（状態共有モデル）が計算される．

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．

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

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，$\mathcal{N}(\cdot; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ 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 MSD-HMM でモデル化される．

$$b_c(\boldsymbol{o}_t) = \begin{cases} w_c \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), & l_t = V \\ 1 - w_c, & l_t = U \end{cases} \tag{3}$$

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．

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

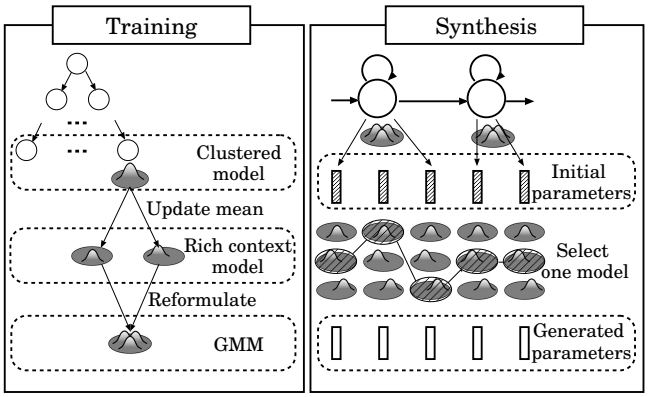

図 2　分散共有フルコンテキストモデルによる学習部及び生成部
Fig. 2　Training and synthesis processes with rich context models

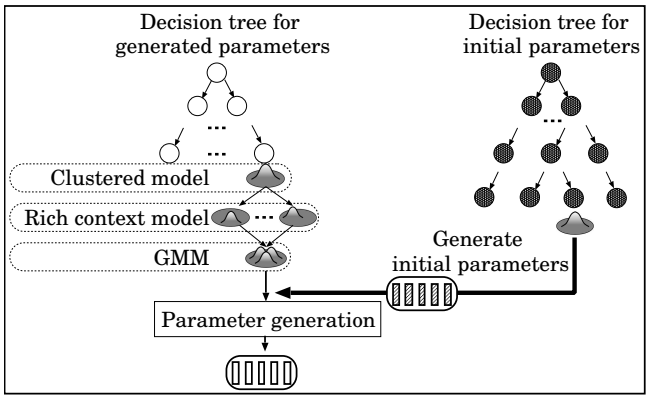

図 3　コンテキストクラスタリングによる初期パラメータ生成法の概観図
Fig. 3　Overview of the initialization method with tree-based context clustering.

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 [18]．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，合成音声品質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 [11]．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 $\mathcal{N}\left(\cdot; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c\right)$ により，次式で示される．

$$b_{c,m}\left(\boldsymbol{o}_t\right) = \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c\right) \tag{4}$$

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．

$$b_c\left(\boldsymbol{o}_t\right) = \sum_{m=1}^{M_c} \omega_m \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c\right) \tag{5}$$

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 [13]

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．

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

ただし，モデル系列を $\boldsymbol{m} = \{m_1, \cdots, m_T\}$ 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．

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

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

式 (7) 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等しいが，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．

初期パラメータの決定には，図 3 に示すように，コンテキストクラスタリングによる初期パラメータ生成法 [14] が有効である．初期パラメータを生成するために，MDL 基準によりサイズの大きな決定木を構築し，従来の HMM 音声合成のパラメータ生成法 [18] に従い，初期パラメータ系列を生成する．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るパラメータの GV 尤度 [19] を最大にするように，式 (1) のパラメータ a を設定することで，音質は著しく改善する．

4. 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を用いた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，有声空間の平均ベクトルを更新することで得られる．

$$b_{c,m}\left(\boldsymbol{o}_t\right) = \begin{cases} w_c \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c\right), & l_t = V \\ 1 - w_c, & l_t = U \end{cases} \tag{9}$$

ただし，空間重みは状態共有モデルと同じものを使用する．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．

$$b_c(\boldsymbol{o}_t) = \begin{cases} \sum_{m=1}^{M_c} w_{c,m}\mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c\right), & l_t = V \\ 1 - w_c, & l_t = U \end{cases} \quad (10)$$

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．

F_0 パラメータにおけるコンテキストクラスタリングによる初期パラメータ生成法では，まず，大きな決定木における状態共有モデルの有声空間重みが，閾値より大きいフレームを有声フレームとし，それ以外を無声フレームとすることで有声/無声区間を決定する．次に，有声フレームに対応する正規分布を連結し，有声/無声境界フレームにおける動的特徴量に対する精度行列（共分散行列の逆行列）を零行列とした後に，有声フレームのパラメータ系列を生成する．最後に，無声区間を除いたフレームに無声シンボルを挿入して初期パラメータ系列を生成する．初期パラメータ系列の有声/無声区間によって有声/無声判定を行うため，最終的に生成されるパラメータ系列は，大きな決定木における状態共有モデルの有声空間重みに依存する．

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 [20] 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 [21] 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 [22] 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）[23] の学習を行い，パラメータ生成時には系列内変動（Global Variance: GV）[19] を考慮しない．全評価において，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．

まず，提案法の有効性を評価するため，F_0 に対して分散共有フルコンテキストモデルを適用する．次に，スペクトル・F_0 に対して分散共有フルコンテキストモデルを適用し，両パラメータにおいて有効性を評価する．

5.2 F0 における評価

5.2.1 選択モデルと生成パラメータの評価

コンテキストクラスタリングによる初期パラメータ生成法を用いて，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，生成パラメータに対する GV の対数尤度 [19]，及び，有声/無声不一致率とする．ただし，HMM 状態系列 $\boldsymbol{q}$ は，従来の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．HMM の対数尤度は，自然音声のパラメータを用いて有声/無声区間を決定した後に計算する．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．有声/無声不一致率は，自然音声のパラメータに対する生成パラメータの有声/無声の不一致フレーム数を総フレーム数で正規化したものを表す．評価は，初期パラメータ生成法における決定木のパラメータ 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．図 4 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）と従来の状態共有モデルで生成した特徴量（Clus）を初期パラメータとした場合の尤度も計算する．

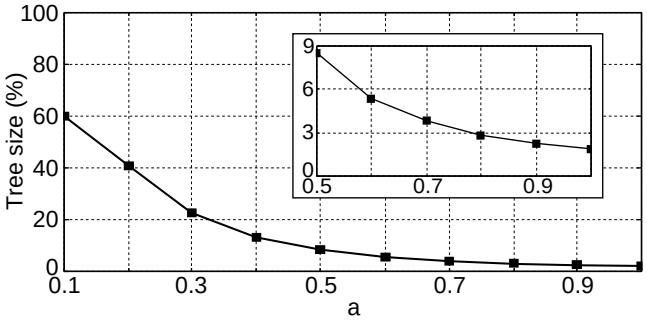

図 4　初期パラメータ生成法における決定木のサイズ
Fig. 4　Size of the decision tree for initial parameter generation.

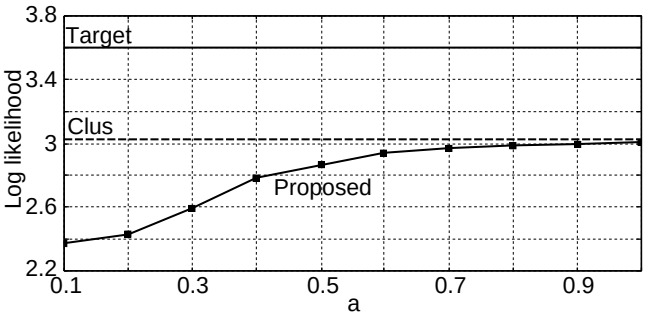

図 5　選択されたモデルの自然音声に対する HMM 尤度
Fig. 5　HMM likelihood of selected rich context models for natural parameters.

HMM の対数尤度を図 5 に，GV の対数尤度を図 6 に，有声/無声不一致率を図 7 に示す．図 5 から，a を小さくするに従い HMM 尤度は単調に減少することが分かる．また，全ての a における HMM 尤度は，自然音声を用いた場合の尤度に大きく及ばない事が分かる．一方，図 6 から，全ての a における GV 尤度は自然音声を用いた場合の尤度よりも大きく，a を小さくするに従い GV 尤度は上昇し，$a = 0.6$ で最大となり，その後減少する事が分かる．また，図 7 から，有声/無声不一致率は a を小さくするに従い緩やかに上昇することが分かる．これらの結果から，$a = 0.6$ の設定は，有声/無声不一致率が僅かに上昇するものの初期パラメータ生成法に有効であることが予想される．

5.2.2 提案法の有効性

次に，提案法の音質を評価する．従来の状態共有モデルから生成した特徴量（Conv），初期パラメータを，状態共有モデルから生成した特徴量（Proposed (Clus)），$a = 0.6$ の決定木による初期パラメータ生成法で生成した特徴量（Proposed ($a = 0.6$)），自然音声の特徴量（Target）とした，分散共有フ

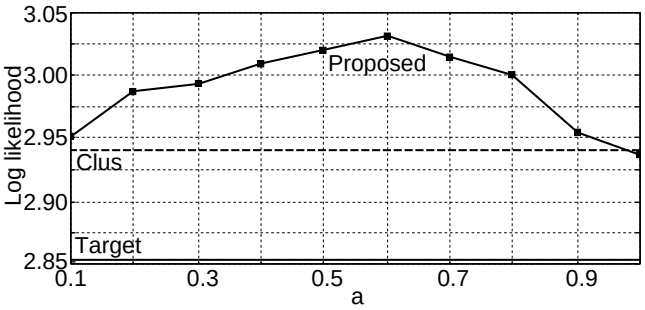

図 6　生成されたパラメータの GV 尤度

Fig. 6　GV likelihood for generated parameters.

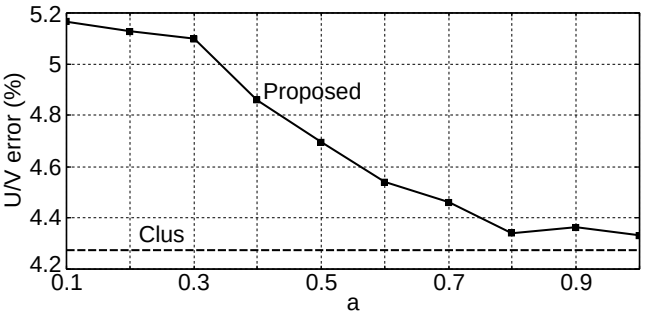

図 7　有声/無声不一致率

Fig. 7　Error rates of unvoiced/voiced decision.

ルコンテキストモデルを用いたパラメータ生成法で生成した特徴量の合成音声を用いて，音質に関するプリファレンステスト（AB テスト）を実施する．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．受聴者は男性 6 名とする．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．

音質の主観評価結果を図 8 に，各手法の F0 パターンを図 9 に示す．コンテキストクラスタリングによる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よりも高いことが分かる．この結果より，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，F_0 においても有効であることが示される．

5.2.3　モデル選択単位の比較

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，従来の状態共有モデルから生成した特徴量（Conv），フレーム単位で選択した分散共有フルコンテキストモデルから生成した特徴量（Proposed (Frame)），HMM 状態単位で選択した分散共有フルコンテキストモデルから生成した特徴量（Proposed (State)）を比較する．ただし，初期パラメータ系列は自然音声のパラメータとする．評価は，男性 7 名によるプリファレンススコアのスコアとする．

音質の主観評価結果を，図 10 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．この評価結果は，スペクトルパラメータにおける結果と同様である [13].

5.3　スペクトル・F0 における評価

スペクトル・F_0 の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（Conventional），分散共有

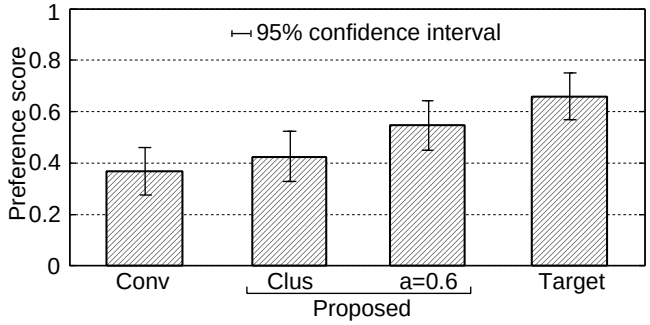

図 8　音質の主観評価結果（F_0）

Fig. 8　Preference scores on speech quality (F_0) .

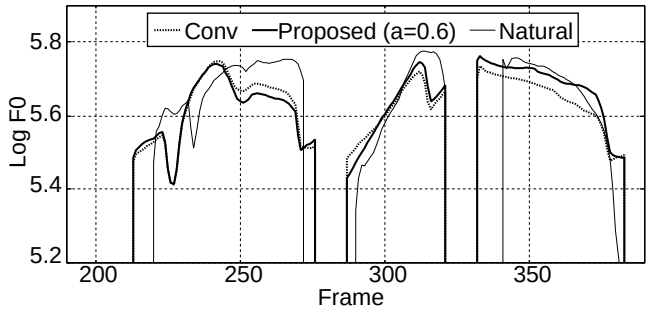

図 9　F_0 パターンの例．発話内容は ”それはたいてい”（”Natural” は自然音声の F_0 パターンを表す）

Fig. 9　An example of F_0 contours for a sentence fragment ”sorewa taitei” (”Natural” represents F_0 contour of natural speech) .

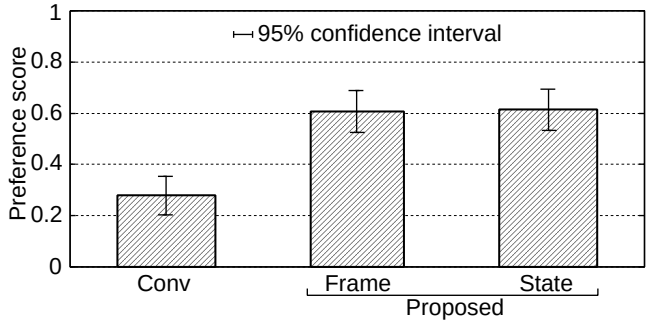

図 10　音質の主観評価結果（モデル選択単位の比較）

Fig. 10　Preference scores on speech quality (comparison of selection unit) .

フルコンテキストモデル（Proposed），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．ただし，ターゲットは，自然音声のパラメータ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手法を表す．初期パラメータ生成法には，コンテキストクラスタリングによる初期パラメータ生成法を使用し，生成パラメータの GV 尤度を最大とするようにパラメータ a を設定する．評価は，男性 8 名によるプリファレンススコアのスコアとする．

音質の主観評価結果を図 11 に示す．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

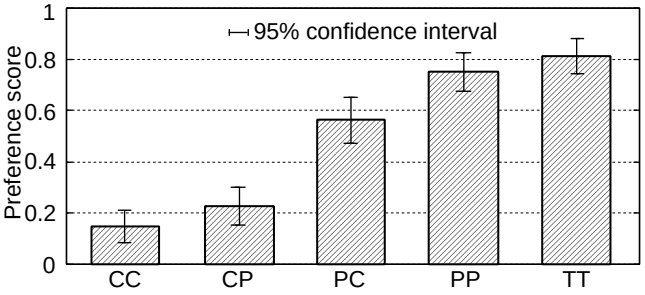

図 11　音質の主観評価結果（スペクトル・F_0 への適用の効果）

Fig. 11　Preference scores on speech quality (Effectiveness of proposed method for spectrum and F_0) .

表 1　評価に用いる手法

Table 1　Synthetic speech samples used for evaluation.

Method	Spectrum	F_0
CC	Conventional	Conventional
CP	Conventional	Proposed ($a = 0.6$)
PC	Proposed ($a = 0.1$)	Conventional
PP	Proposed ($a = 0.1$)	Proposed ($a = 0.6$)
TT	Target	Target

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．

6. ま と め

本稿では，HMM 音声合成において，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る合成音声の音質を改善させるために，本手法を更に F_0 パターン生成へ適用した．実験的評価結果から，スペクトル・F_0 に対して分散共有フルコンテキストモデルを適用することで，従来の HMM 音声合成，及び，スペクトルのみに対して分散共有フルコンテキストモデルを用いる場合と比較して，著しい音質改善が得られることが明らかになった．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．

文　献

[1] Y. Sagisaka, "Speech Synthesis by Rule Using an Optimal Selection of Non-Uniform Synthesis Units," Proc. ICASSP, pp. 679–682, New York, U.S.A, Apr. 1988.

[2] N. Iwahashi, N. Kaiki, and 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] A. J. Hunt and A.W. Black, "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database," Proc. ICASSP, pp. 373–376, Atlanta, U.S.A., May 1996.

[4] H. Zen, K. Tokuda, and A. Black, "Statistical Parametric Speech Synthesis," Speech Commun., Vol. 51, No. 11, pp. 1039–1064, 2009.

[5] A. K. Syrdal, C. W. Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Strom, K-S. Lee, and M.J. Makashay, "Corpus-based techniques in the AT&T NextGen synthesis system," Proc. ICSLP, Vol. 3, pp. 410–415, Beijing, China, Oct. 2000.

[6] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura, "Speaker Interpolation for HMM-based Speech Synthesis System", J. Acoust. Soc. Jpn. (E), Vol. 21, No. 4, pp. 199–206, 2000.

[7] J. Yamagishi, and T. Kobayashi, "Average-voice-based speech Synthesis Using HSMM-based Speaker Adaptation and Adaptive Training," IEICE Trans., Inf. and Syst., Vol. E90-D, No. 2, pp. 533–543, 2007.

[8] T. Nose, J. Yamagishi, T. Masuko, and 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.

[9] S. king, V. Karaiskos, "The Blizzard Challenge 2011," Proc. Blizzard Challenge workshop, Turin, Italy, Sept. 2011.

[10] Z. Ling, L. Qin, H. Lu, Y. Gao, L. Dai, R. Wang, Y. Jiang, Z. Zhao, J. Yang, J. Chen, and G. Hu, "The USTC and iflytek Speech Synthesis Systems for Blizzard Challenge 2007," Proc. Blizzard Challenge workshop, Bonn, Germany, Aug. 2007.

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

[12] Y.Qian, Z. Yan, Y. Wu, and F. K. Soong, "An HMM Trajectory Tiling (HTT) Approach to High Quality TTS," Proc. INTERSPEECH, pp. 422–425, Chiba, Japan, Sept. 2010.

[13] 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," Proc. INTERSPEECH, Portland, USA, Sept. 2012.

[14] 高道 慎之介, 戸田 智基, 志賀 芳則, Sakriani Sakti, Graham Neubig, 中村 哲, "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," 電子情報通信学会技術研究報告, SP2012-78, pp.37–42, Nov. 2012.

[15] K. Tokuda, T. Masuko, B. Miyazaki, and T. Kobayashi, "Multi-Space Probability Distribution HMM," IEICE Trans., Inf. and Syst., Vol. E85-D, No. 3, pp. 455–464, 2002.

[16] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-based speech synthesis," Proc. EUROSPEECH, pp. 2347–2350, Budapest, Hungary, 1999.

[17] 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.

[18] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech Parameter Generation Algorithms for HMM-based Speech Synthesis," Proc. ICASSP, pp. 1315–1318, Istanbul, Turkey, June 2000.

[19] T. Toda, and 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.

[20] M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara,“ ATR technical report, ” No. TR-I-0166, 1990.

[21] H. Kawahara, I. Masuda-Katsuse, and 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.

[22] 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, pp. 1–6, Firentze, Italy, Sept. 2001.

[23] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden Semi-Markov Model Based Speech Synthesis System," IEICE Trans., Inf. and Syst., E90-D, No. 5, pp. 825–834, 2007.