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.

音声入力に基づく韻律制御機能を有するHMM音声合成システム

西垣　友理† 　高道慎之介† 　戸田　智基† 　Graham Neubig† 　Sakriani Sakti†

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

あらまし　音声合成を用いた創作活動の活発化により，目標とする特定話者の音声の合成処理において，ユーザの思い通りの音声を合成できるインターフェースの開発が望まれている．これに対して本報告では，高品質かつ表現力豊かな音声合成の実現を目指して，HMM音声合成において，通常のテキスト音声合成機能を保持しつつ，ユーザによる入力音声を用いて目標話者の合成音声の韻律を制御する手法を提案する．入力音声に対して，HMM状態アライメントを行うことで，入力音声の継続長を抽出し，それを目標話者用HMMに反映させることで，入力音声の継続長に対応した目標話者の合成音声を生成する．さらに，入力音声のF_0パターンを合成音声に反映させることで，入力音声のF_0パターンおよび継続長に対応した目標話者の合成音声を実現する．入力音声と合成音声間における韻律パラメータの不一致による自然性劣化を抑えるために，反映させる継続長の単位に関する検討，モデル適応処理を用いたアライメント用HMMの構築に関する検討，および，有声 / 無声情報に対応する補正処理に関する検討を行う．複数の入力話者を対象とした実験的評価結果から，提案法の有効性を示す．

キーワード　HMM音声合成，韻律制御，継続長単位，モデル適応，有声 / 無声情報

HMM-Based Speech Synthesis System with Prosody Modification Based on Speech Input

Yuri NISHIGAKI†, Shinnosuke TAKAMICHI†, Tomoki TODA†, Graham NEUBIG†, Sakriani SAKTI†, and Satoshi NAKAMURA†

† Nara Institute of Science and Technology,　Tatayama-cho 8916–5, Ikoma, Nara, 630–0192 Japan

Abstract　As a creative activity using speech synthesis technologies has been grown rapidly, it is desired to develop an interface to synthesize speech of a specific target speaker as users want. In this report, we propose a prosody modification method using user's speech inputs in HMM-based speech synthesis system in order to achieve high--quality and expressive speech synthesis. The propose method allows users to guide prosody of synthetic speech of the target speaker by using their own voices while preserving original functions of the HMM-based speech synthesis system as a text-to-speech synthesis system. Both duration information of the input speech extracted by performing HMM state alignment and F_0 patterns of the input speech are effectively used to control the duration and F_0 patterns of synthetic speech of the target speaker. To alleviate the degradation of naturalness caused by prosodic mismatches between the input speech and the synthetic speech, we investigate an appropriate unit for the HMM state alignment, model adaptation for building an HMM used for the alignment, and correction of unvoiced/voiced information. Experimental evaluations are conducted for multiple input speakers, which demonstrates the effectiveness of the proposed method.

Key words　HMM-based speech synthesis, prosody modification, alignment unit, model adaptation, unvoiced/voiced information

1. まえがき

コーパスベース音声合成技術により，合成音声の明瞭性および自然性の改善のみでなく，話者性 (個人性) の再現精度に関しても，飛躍的な改善がもたらされた [1] . これに伴い，テキスト音声合成システム CeVIO [2] や歌声合成システム VOCALOID [3] に代表されるように，特定のキャラクタ性を有する音声 / 歌声の合成技術に対する需要が日々増しており，所望の音声 / 歌声を創作する活動において，その利用が大いに期待されている．

音声創作活動においては，テキストから音声を合成するテキスト音声合成技術は欠かせない基盤技術である．その中でも，HMM 音声合成 [4] , [5] は，合成音声の声質や韻律的特徴を柔軟に制御することが可能であるため ，その有効性は極めて高い．音声パラメータを直接手動で操作するシステム [6] と比較し，抽象化されたモデルパラメータの手動操作による合成音声の制御が可能であるため [7] ，より高い操作性が得られる．一方で，ユーザの思い通りの音声を合成することは未だ容易ではなく，より使い勝手の良いユーザインタフェースの構築が望まれる．

ユーザが意図した合成音声を作成する上で，手動操作を補完する入力情報として，ユーザが発声する音声情報を用いる枠組みが考えられる．例えば，歌声合成の分野においては，所望の表情を持つ歌声を合成するために，ユーザによる歌声を参照データとして入力し，VOCALOID の入力パラメータを自動で最適化する手法が提案されている [8] , [9] . ユーザが求める表情を自身の歌声により表現することで，VOCALOID で合成される特定歌手の歌声に反映させることができ，より直感的かつ容易に所望の歌声が合成可能となる．話し声を対象とした音声合成の分野においても，類似した枠組みとして，ユーザの入力音声の声質を所望の目標話者の声質へと変換する声質変換技術 [10] , [11] の利用が考えられる．入力音声の韻律特徴を保持することが可能であるため，所望の韻律による目標話者の音声が作成可能となる．しかしながら，声質変換は HMM 音声合成と比較して，合成音声の品質および話者性再現精度が劣化する傾向にある．これに対して，HMM 音声合成の枠組みにおいて，入力された音声の韻律特徴を合成音声に反映させる枠組みが提案されている [12] . この枠組みでは，HMM 音声合成による品質を保持したまま，入力音声による韻律操作が可能となる．一方で，入力音声の韻律を反映させる上で独自のコンテキストを用いる必要があるため，従来のテキスト音声合成システムとしての性能を保持できる保証はない．仮に，テキスト音声合成システムとしての機能を保持しつつ，入力音声による韻律制御を可能とする枠組みを実現することができれば，手動操作と入力音声による操作を相補的に使用することも可能となり，より操作性に優れたインタフェースを実現できると考えられる．

本報告では，HMM 音声合成において，通常のテキスト音声合成機能を保持し，かつ，音声を用いて合成音声の韻律制御を可能とする手法を提案する．入力音声に対して，HMM 状態アライメントを行うことで，入力音声の継続長を抽出する．それを合成時に用いる目標話者用 HMM に反映させることで，入力音声の継続長に対応した目標話者の合成音声を生成する．さらに，入力音声の F_0 パターンを目標話者の合成音声へと転写する．本処理を行う上で，ユーザによる入力音声と目標音声との間には，韻律パラメータの不一致が生じるため，合成音声の品質が劣化する可能性がある．この問題に対応するため，反映させる継続長の単位の選定，モデル適応処理を用いたアライメント用 HMM の性能改善，および，有声 / 無声情報の不一致に対応する補正処理に関する検討を行う．個々の処理の効果を評価するために，複数の入力話者を対象とした実験的評価を行う．

2. HMM 音声合成

HMM 音声合成における学習部では，自然音声のパラメータ (スペクトル，F_0，非周期成分) 系列をコンテキスト依存 HMM によりモデル化する．また，状態継続長モデルの導入により，各種音声パラメータと継続長を統一的な枠組みの下でモデル化することが可能である [5] . 生成時には，入力テキストを解析することで得られるコンテキストに基づき，文 HMM を構築し，状態継続長モデルから HMM 状態系列 $\boldsymbol{q} = [q_1, \cdots, q_T]$ を次式により決定する．

$$\hat{\boldsymbol{q}} = \underset{\boldsymbol{q}}{\operatorname{argmax}} P\left(\boldsymbol{q}|\boldsymbol{\lambda}\right) \tag{1}$$

ただし，T はフレーム数，q_t はフレーム t における HMM 状態インデックス，$\boldsymbol{\lambda}$ は HMM のパラメータセットである．音声パラメータ系列は，静的・動的特徴量間の明示的な制約の下で，HMM 尤度を最大化するように生成される [13] .

$$\boldsymbol{c} = \underset{\boldsymbol{c}}{\operatorname{argmax}} P\left(\boldsymbol{W}\boldsymbol{c}|\hat{\boldsymbol{q}}, \boldsymbol{\lambda}\right) \tag{2}$$

ただし，$\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \cdots, \boldsymbol{c}_t^\top, \cdots, \boldsymbol{c}_T^\top\right]^\top$ は音声パラメータ系列，$\boldsymbol{c}_t = \left[c\left(1\right), \cdots, c\left(D\right)\right]^\top$ はフレーム t における音声パラメータベクトル，W は動的特徴量の計算に用いる重みによって構成される行列 [4] である．出力音声は，生成パラメータからボコーダーに基づく波形生成処理を経て合成される．HMM 音声合成の合成音声の音質は，自然音声と比較して劣化する傾向にあるが，パラメータ生成時に系列内変動を考慮することで，音質の改善が可能である [14].

3. 音声による韻律制御を有する HMM 音声合成

目標話者の HMM を用いて，ユーザが入力した音声の韻律を模倣した合成音声を生成するために，音声による韻律制御法を提案する．提案法における処理の流れを図 1 に示す．入力テキストおよび入力音声を用いて，所望の韻律を持つ目標話者の音声を合成する．なお，目標話者の HMM はテキスト音声合成で用いられるものと同一であるため，音声が入力されない際には，通常の HMM 音声合成処理により音声を合成できる．

3.1 継続長の制御

入力音声の継続長情報を抽出するために，HMM 状態アライメントを行う．まず，入力音声の音響特徴量に対応したアライメント用の HMM を用意し，入力テキストに応じた文 HMM を構成する．次に，入力音声に対して HMM 状態アライメントを

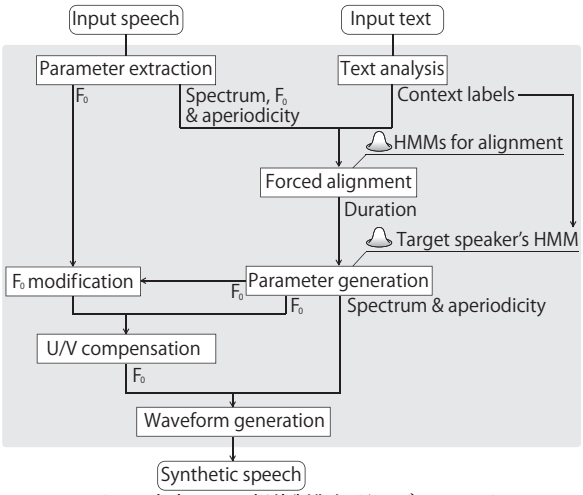


図 1　音声による韻律制御処理のブロック図

Fig. 1　Diagram of prosody modification based on speech inputs.

行うことで，入力音声の状態継続長を決定する．合成音声の生成に用いる目標話者 HMM を用いて，入力テキストに応じた文 HMM を構成し，入力音声から得られた継続長の情報に基づき，HMM 状態系列を決定する．最終的に，従来の枠組みと同様に，尤度最大化基準に基づき音声パラメータを生成することで，入力音声の継続長を持つ目標話者の合成音声パラメータ系列が得られる．

本報告において，主に以下の 2 点について注意深く検討する必要がある．

1：　入力音声に対して十分な精度で HMM 状態アライメントを行うために，入力音声の音響特徴量を適切にモデル化しているアライメント用 HMM を用いる必要がある．

2：　アライメント用 HMM と音声合成用 HMM は異なる話者の音声をモデル化するものであり，両 HMM における個々の HMM 状態は同一の音響特徴量セグメントをモデル化する保証はない (例えば，ある音素に対する 5 状態 HMM において，アライメント用 HMM では 2 状態目で音素中心部の定常箇所をモデル化するのに対し，音声合成用 HMM では 4 状態目で音素中心部の定常箇所をモデル化するということが起こり得る) ため，知覚的に悪影響が生じない範囲で継続長を反映させる必要がある．

1 つ目の点に関しては，本報告では，アライメント用 HMM として，1) 入力話者 (音声入力を行うユーザ) の音声が事前に十分な量入手できる場合を想定した入力話者 HMM と，2) 入力話者の少量の音声を用いて目標話者 HMM をモデル適応 [15] することで得られた適応 HMM の使用を検討する．なお，モデル適応の際には，不特定話者モデルや平均声モデル [15] を用いることも可能である．

2 つ目の点に関しては，本報告では，入力音声から合成音声へと反映させる継続長の単位として，状態継続長，音素継続長，モーラ継続長の使用を検討する．音素継続長およびモーラ継続長を用いる際には，合成音声パラメータ生成時において，音素内およびモーラ内の HMM 状態系列は，音素 / モーラ継続長が与えられている条件の下での状態継続長モデルの尤度最大化基準により決定する [5].

3.2　F_0 系列の制御

入力音声の継続長に基づき生成された合成音声パラメータに対して，さらに入力音声の F_0 系列を反映させる．最終的に得られた合成音声パラメータから，入力音声の継続長および F_0 系列を持つ目標話者の合成音声を生成する．

3.2.1　F_0 系列の生成と音高変換

合成音声の F_0 系列を，入力音声の F_0 系列と入れ替えることで，入力音声の継続長および F_0 系列を持つ合成音声パラメータ系列を構築する．その際に，入力話者と目標話者の F_0 範囲の差を補正するために，入力音声の F_0 に対して，以下の線形変換を行う．

$$\hat{x_t} = \frac{\sigma_y}{\sigma_x}\left(x_t - \mu_x\right) + \mu_y \tag{3}$$

ただし x_t はフレーム t における入力音声の対数 F_0，μ_x と σ_x はそれぞれ x_t の平均と標準偏差，μ_y と σ_y は合成する目標話者の対数 F_0 の平均と標準偏差である．得られた合成音声パラメータから，入力音声の継続長および F_0 系列を持つ目標話者の合成音声を生成する．

3.2.2　有声/無声区間の補正

入力音声の F_0 系列は，入力音声に依存した有声 / 無声情報を持つ．一方で，合成時に使用するスペクトルパラメータ系列は目標音声に対応しているため，スペクトルと有声 / 無声情報の不一致が生じる可能性がある．そこで，目標話者 HMM から生成される F_0 系列の有声/無声情報を用いて，入力音声の F_0 系列の有声 / 無声情報を補正する．

まず，入力音声の F_0 系列に対してスプライン補間処理を行うことで，無声区間の F_0 を推定し，連続的な F_0 系列を得る [16], [17] ．この際に，マイクロプロソディ [18] の除去も行う．得られた連続的な F_0 系列に対して，目標話者 HMM により決定される無声情報を付与することで，合成音声の F_0 系列を生成する．

上記の処理によって得られる F_0 系列の一例を図 2 に示す．図 2 より，入力音声の F_0 系列の音高および有声 / 無声区間が補正されていることがわかる．

4.　実験的評価

4.1　実 験 条 件

目標話者 HMM の学習データとして，日本人女性話者による ATR 音素バランス文 [19] 中の A - I セットの計 450 文を用いる．一方で，音声入力を行う話者（入力話者）は，目標話者とは異なる日本人男性 / 女性話者各 2 名の計 4 名とする．各入力話者による ATR 音素バランス文中の A - I セットの計 450 文を用いて，入力話者依存アライメント用 HMM を学習する．また，各入力話者の音声データを用いて目標話者 HMM を適応することで，適応処理に基づく入力話者依存アライメント用 HMM も構築する．適応データとして，各入力話者による 450 文を使用し，適応文数を変化させることで，適応処理に必要なデータ量について検討する．各入力話者に対して，ATR 音素バランス文中の J セット 53 文を評価データとして使用する．

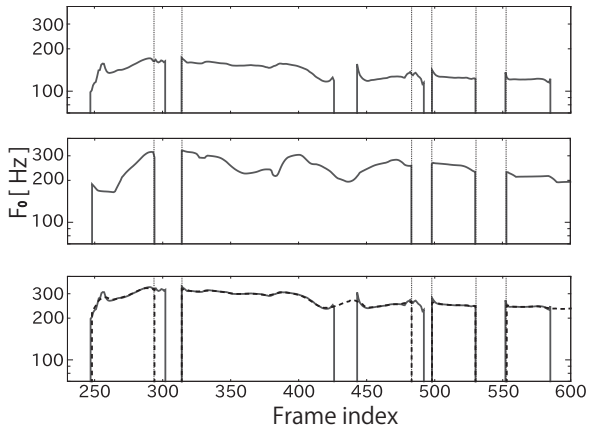


図 2　各 F_0 系列の例（上 : 入力音声の F_0 系列，中 : 目標話者 HMM から入力音声の継続長に基づき生成される F_0 系列，下 : 音高変換を施した入力話者の F_0 系列（点線）と，さらに有声 / 無声情報の補正を施した F_0 系列（実線））

Fig. 2　Example of F_0 contours (top: F_0 contour of input speech, middle: F_0 contour generated from target speaker's HMM, bottom: F_0 contour of input speech after range transformation (dotted line) and that after unvoiced/voiced compensation (solid line)).

音声データのサンプリング周波数は 16 kHz とする．スペクトルパラメータとして，STRAIGHT 分析 [20] で得られるスペクトル包絡をモデル化した 0 次から 24 次のメルケプストラム係数を使用する．音源パラメータとして，対数 F_0 および 5 周波数帯域における平均非周期成分 [21] を使用する．フレームシフトは 5 ms とする．HMM は 5 状態 left-to-right 型とし，通常の学習処理 [5] に基づき，コンテキスト依存音素 HMM を学習する．

継続長制御の評価においては，入力音声の継続長を反映させる単位として，状態継続長（State），音素継続長（Phone），モーラ継続長（Mora）を使用する．アライメント用 HMM として，入力話者 HMM（450 文にて学習），適応 HMM（1 文から 450 文まで適応文数を変化），および目標話者 HMM（適応前のモデルに対応）を用いる．客観評価として，入力話者 HMM で継続長を決定した際の合成音声のメルケプストラムをリファレンスとして，他のアライメント用 HMM で継続長を決定した際の合成音声に対するメルケプストラム歪みを測定する．アライメント単位毎にメルケプストラム歪みを測定することで，各継続長単位間の類似性や適応文に応じた適応処理の効果について調査する．また，主観評価として，3 種類の継続長単位に対して，目標話者 HMM（Target），1 文適応による適応 HMM（1 utterance），56 文適応による適応 HMM（56 utterances），入力話者 HMM（Reference）をアライメント用 HMM として使用した際の合成音声計 12 種類（継続長単位が 3 種類とアライメント用 HMM が 4 種類の組み合わせ）を比較する．合成音声の自然性に関する MOS 評価実験と，入力音声をリファレンスとした際の韻律の摸倣性に関する DMOS 評価を行う．

F_0 系列制御の評価においては，有声 / 無声情報補正処理の効果を調査する．補正あり（w/ mod）と補正なし（w/o mod）の 2 手法に対して，合成音声の自然性に関する対比較実験を行う．アライメント用 HMM には，入力話者 HMM を用い，継続長単位は音素継続長とする．また，各入力話者に対する有声 / 無声情報不一致率（U/V 不一致率）についても調査する．

4.2　継続長に関する評価実験

4.2.1　客観的評価実験結果

図 3 (a) ～ (c)，適応 HMM および目標話者 HMM により継続長を決定した際のメルケプストラム歪みを示す．また，参考として，目標話者 HMM により継続長を決定したメルケプストラムに対する歪みも計算した．ただし，各図におけるリファレンスの生成時に用いた継続長単位は，それぞれ，(a) 状態継続長，(b) 音素継続長，(c) モーラ継続長単位である．全ての継続長単位において，1 文適応の適応 HMM を用いることで，目標話者 HMM を用いた場合と比較して，メルケプストラム歪みが大きく減少する．また，適応文数を増加させることで，緩やかではあるが，さらにメルケプストラム歪みが減少する傾向が見られる．このことから，アライメント用 HMM の構築において，モデル適応は有効であり，1 文という限られた適応データ量においても，その効果が得られることが分かる．

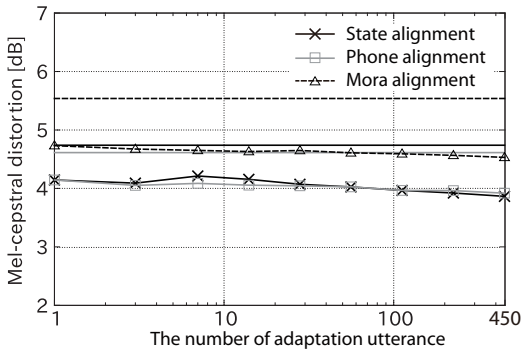
次に，継続長単位に着目すると，図 3 (a) ～ (c), から，リファレンス生成時に用いた継続長単位と同じ単位を用いた際に，最もメルケプストラム歪みが小さくなる傾向が見られる．また，音素継続長単位に関しては，リファレンス生成時に状態継続長 (図 3 (a)) またはモーラ継続長 (図 3 (c)) を用いた際においても，比較的メルケプストラム歪みが小さいことが分かる．特に，リファレンス生成時に状態継続長を用いた際 (図 3 (a)) においては，リファレンスと同じ状態継続長を用いた場合と比較しても，同程度に小さなメルケプストラム歪みが得られることが分かる．

4.2.2　主観的評価実験結果

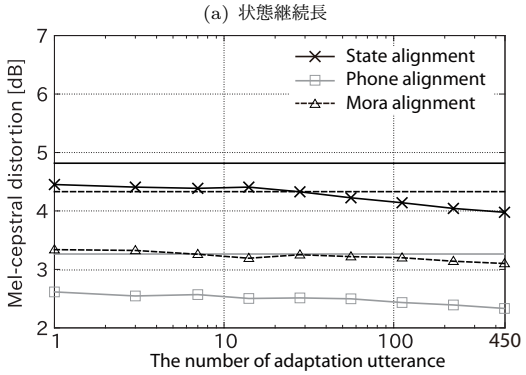
自然性に関する MOS 評価結果を図 4 に，韻律の模倣性に関する DMOS 評価を図 5 に示す．図 4 と図 5 から，自然性と模倣性の両面において，適応 HMM (“1 utterance” と “56 utterance”) のスコアは，目標話者 HMM (“target”) のスコアを上回ることが分かる．また，モーラ継続長単位を使用した場合の模倣性のスコアを除いて，56 文による適応 HMM (“56 utterance”) は，入力話者 HMM (“Reference”) と同等の自然性および模倣性が得られることが分かる．以上の結果から，適応処理によるアライメント用 HMM の構築は，自然性と模倣性を改善する上で効果的な手法であるといえる．

モーラ継続長単位を使用した際には，他の継続長単位を使用した場合と比較して，模倣性が劣化する傾向が見られる．モーラ継続長単位を利用した際には，他の継続長単位を使用した場合と比較して，合成時における HMM 状態系列は目標話者 HMM の状態継続長モデルの影響をより強く受ける．その結果，模倣性の劣化が生じると考えられる．

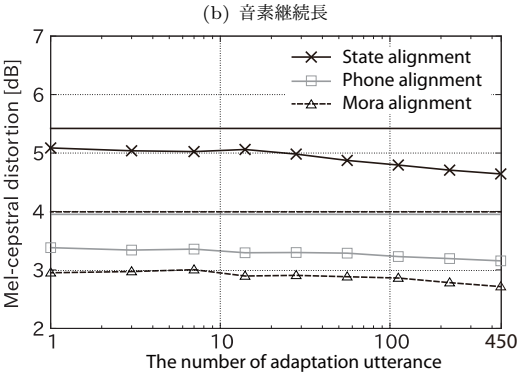
自然性と模倣性に関して，最も高いスコアが得られている音素単位アライメントを用いた “Reference” のスコアと，他手法のスコアとの間にて，t 検定を行った結果を表 1（自然性に関する MOS に対する結果）と表 2（ 模倣性に関する DMOS に対する結果 ） に示す．1 文を用いた適応 HMM に着目すると，



(a) 状態継続長



(b) 音素継続長



(c) モーラ継続長

図 3　各継続長単位におけるメルケプストラム歪み (実線 (黒色)，実線 (灰色)，破線はそれぞれ，目標話者 HMM を用いて状態，音素，モーラの継続長単位を決定した場合の歪みを表す．)

Fig. 3　Mel-cepstral distortion in each alignment unit (Black line, Gray line, and dotted line indicate the distortion with the state-level, phoneme-level, and mora-level alignments determined using the target speaker's HMM.) .

自然性に関しては，状態単位を用いた際のスコアとの間に，有意水準 1 % にて有意差が認められる．また，模倣性に関しては，状態単位およびモーラ単位を用いた際のスコアとの間に，有意水準 1 % にて有意差が認められる．音素単位を用いた際には，自然性および模倣性の両面において有意差が認められない．このことから，音素単位の使用が最も効果的であると判断する．

4.3　有声/無声補正処理の評価結果

図 6 に U/V 不一致率を，図 7 に自然性に関するプリファレンススコアを示す．図 6 より，全ての入力話者において，U/V 不一致が生じていることが分かる．また，図 7 より，4 名中 3 名の入力話者において，U/V 補正処理により自然性が改善されたことが分かる．なお，Speaker 2 では，他の話者と比較し，U/V 不一致率は中程度であるものの，自然性の改善効果は顕著に大きい．通常，有声区間におけるパワーは，無声区間のパワーと比較して大きい傾向にある．そのため，有声区間におけるスペクトルと無声音源を用いて音声を合成した場合，自然性が大きく劣化する．このことから，Speaker 2 のように，目標話者 HMM による有声区間が入力音声の無声区間に多く割り当たる際には，U/V 補正処理による自然性の改善効果が上昇すると考えられる．

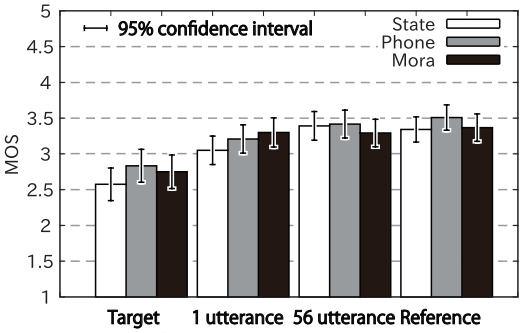


図 4　自然性に関する MOS 評価結果

Fig. 4　Mean opinion score on speech quality.

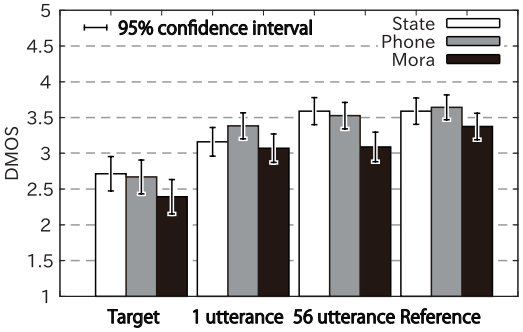


図 5　韻律の模倣性の DMOS 評価結果

Fig. 5　Degradation mean opinion score on prosodical mimic ability.

表 1　自然性に関する MOS スコアにおける，音素アライメントを用いた “Reference” に対する各手法の p 値

Table 1　P-values compared to “Reference” with the phoneme-level alignment in term of the MOS scores for speech quality.

	State	Phone	Mora
Target	$8.77e^{-10}$	$6.89e^{-6}$	$6.96e^{-7}$
1 utterance	$7.58e^{-4}$	0.0259	0.128
56 utterance	0.389	0.490	0.102
Reference	0.187	\	0.283

表 2　模倣性に関する DMOS スコアにおける，音素アライメントを用いた “Reference” に対する各手法の p 値

Table 2　P-values compared to “Reference” with the phoneme-level alignment in term of the DMOS scores for mimic ability.

	State	Phone	Mora
Target	$3.07e^{-9}$	$4.03e^{-10}$	$1.05e^{-14}$
1 utterance	$3.91e^{-4}$	0.0425	$2.73e^{-5}$
56 utterance	0.680	0.365	$6.73e^{-5}$
Reference	0.675	\	0.0376

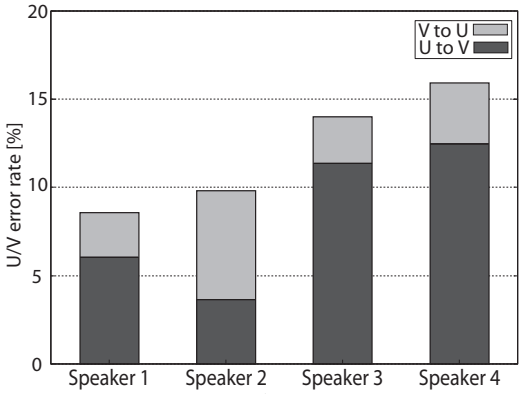


図 6　有声/無声不一致率

Fig. 6　U/V Error rate.

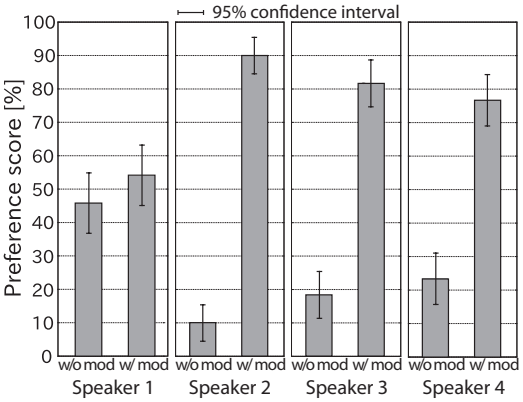


図 7　自然性に関する主観評価結果

Fig. 7　Preference score for speech quality.

5. ま と め

本報告では，目標とする特定話者の音声合成処理において，所望の韻律を持つ合成音声を実現する枠組みを目指し，音声入力による韻律制御機能を有するHMM音声合成法を提案した．提案法において，入力音声と合成音声間における韻律特徴の不一致が最終的な合成音声に与える影響を調査するために，継続長制御時に用いる継続長単位とアライメント用HMMの構築法に関する検討と，F_0系列制御における有声 / 無声情報の補正処理に関する検討を行った．実験結果から，1) 音素単位による継続長制御が有効であること，2) モデル適応処理によるアライメント用HMMの構築は有効であり，適応文数が1文の場合においても，自然性および模倣性の両面において大きな改善効果が得られること，3) 有声 / 無声情報の補正処理により自然性が改善されることを示した．今後は，音声入力によるパワー系列制御や声質制御の導入に取り組む．

謝辞　本研究の一部は，JSPS 科研費 26280060 および 24300073 の助成を受け実施したものである．

文　献

[1] 徳田 恵一, ブラック アラン, “音声合成研究も協調と競争の時代に : The Blizzard Challenge,” 日本音響学会誌, Vol. 62, No. 6, pp. 466–472, Jun., 2006.

[2] CeVIO Creative Studio `http://cevio.jp`

[3] H. Kenmochi, and H. Ohshita, “VOCALOID - Commercial singing synthesizer based on sample concatenation,” *Proc. INTERSPEECH*, pp. 4011–4012, Antwerp, Belgium, Aug., 2007.

[4] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” *Speech Commun.*, Vol. 51, No. 11, pp. 1039–1064, Apr., 2009.

[5] 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,” 電子情報通信学会論文誌, Vol. J83-D-II, No. 11, pp. 2099–2107, Nov., 2000.

[6] 阿部 匡伸, 水野 秀之, 水野 理, 野田 喜昭, 高橋 敏, 中嶌 信弥, “音声デザインツール Sesign,” 電子情報通信学会論文誌, Vol. J84-D-II, No. 6, pp. 927–935, Jun., 2000.

[7] 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, Sept., 2007.

[8] 中野 倫靖, 後藤 真孝, “VocaListener : ユーザ歌唱の音高および音量を真似る歌声合成システム,” 情報処理学会論文誌, Vol. 52, No. 12, pp. 3853–3867, Dec., 2011.

[9] 中野 倫靖, 後藤真孝, “VocaListener2 : ユーザ歌唱の音高・音量に加えて声色変化も真似る歌声合成システム,” 情報処理学会論文誌, Vol. 54, No. 6, pp. 1771–1783, Jun., 2013.

[10] Y. Stylianou, O. Cappe, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 2, pp. 131–142, Mar., 1998.

[11] T. Toda, A. W. Black, and K. Tokuda, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” *IEEE Trans. Audio, Speech, and Language*, Vol. 15, No. 8, pp. 2222–2235, Nov., 2007.

[12] T. Nose, Y. Ota, and T. Kobayashi, “HMM-based voice conversion using quantized F0 context,” *IEICE Trans. Inf. and Syst.*, Vol. E93-D, No. 9, pp. 2483–2490, Sept., 2010.

[13] 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, Jun., 2000.

[14] 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, May, 2007.

[15] 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, Feb., 2007.

[16] K. Yu and and S. Young, “Continuous F0 modeling for HMM based statistical parametric speech synthesis,” *IEEE Trans. Audio, Speech and Language.*, Vol. 19, No. 5, pp. 1071–1079, Jul., 2011.

[17] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, “A hybrid approach to electrolaryngeal speech enhancement based on spectral subtraction and statistical voice conversion,” *Proc. INTERSPEECH*, pp. 3067–3071, Lyon, France, Sept., 2013.

[18] P. Taylor, *Text-To-Speech synthesis.*, Cambridge Univ. Press, 2009.

[19] 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, Sept., 1990.

[20] H. Kawahara, I. Matsuda-Katsuse, and A. D. Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” *Speech Commun.*, Vol. 27, No. 3–4, pp. 187–207, Apr., 1999.

[21] 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,” In *MAVEBA 2001*, pp. 1–6, Firenze, Italy, Sept., 2001.