音声入力による韻律制御機能を有する HMM 音声合成システムの改良 *

☆西垣 友理，高道 慎之介，戸田 智基，
Graham Neubig，Sakriani Sakti，中村 哲 (奈良先端大)

1 はじめに

動画コンテンツ等の創作活動支援のため，特定のキャラクタ性の音声を思い通りに合成するシステムが求められている．我々はこれまでに，通常のテキスト音声合成機能に加え，入力音声を用いた韻律制御を可能にするシステムを提案している [1]．本稿では，韻律制御による合成音声の品質改善の為，(1) 合成音声に反映させる継続長の単位，(2) アライメント用 HMM に対するモデル適応処理，(3) 入力音声の多様な発話様式への対応について調査する．合成音声の自然性と韻律の模倣性について実験的に評価することで，提案法による品質改善効果を明らかにする．

2 音声入力による韻律制御機能を有する HMM 音声合成システム [1]

本システムは，テキスト音声合成の出力音声に対して，与えられる入力音声の韻律を模倣するように，補正処理を施す機能を備える．処理手順を Fig. 1 に示す．入力テキストと入力音声を用いて，入力音声に対応した合成音声の継続長を決定し，合成音声のスペクトルパラメータと非周期成分を生成する．合成音声の F_0 は，入力音声の F_0 を変形して生成する．

2.1 継続長の制御

アライメント用 HMM から入力テキストに応じた文 HMM を構築し，入力音声の音声特徴量に対して HMM 状態アライメントを行う．得られた状態継続長に基づき，音声合成に用いる目標話者 HMM から構築された文 HMM の HMM 状態継続長を決定する．最終的に，従来の HMM 音声合成の枠組みと同様に，尤度最大化基準 [2] に基づいて音声パラメータを生成し，生成パラメータのうち，スペクトルパラメータと非周期成分を合成音声のパラメータとして使用する．

2.2 F_0 系列の制御

入力音声の F_0 と生成された F_0 を用いて，合成音声の F_0 を決定する．生成 F_0 系列の F_0 範囲と有声／無声情報に合うように，対数 F_0 の線形変換 [3] と，連続 F_0 系列 [5] を利用した有声／無声区間補正を用いて，入力音声の F_0 系列を変形する．最終的な合成音声は，生成されたスペクトルパラメータ及び非周期成分と，変形された F_0 系列を用いた波形生成処理により得られる．

2.3 検討すべき課題

本稿では，次の 3 点について検討する．

合成音声に反映させる継続長の単位：アライメント用 HMM と音声合成用 HMM における個々の HMM 状態は同一の特徴量セグメントをモデル化するとは限らない．そのため，HMM 状態レベルの継続長の反映は，合成音声の品質劣化を生じさせる可能性がある．そこで，音声パラメータ生成部に与える継続長の単位として，HMM 状態継続長，音素継続長，モーラ継続長の 3 つを検討する．音素／モーラ継続長が与えられた場合の合成音声の状態継続長は，音素／モーラ継続長を固定した下の状態継続長モデルの尤度最大化基準により決定される [6]．

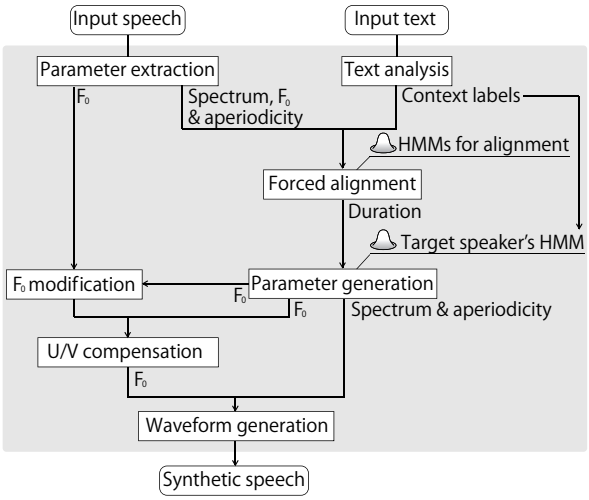


Fig. 1 音声入力による韻律制御機能を有する HMM 音声合成システムの処理手順

アライメント用 HMM の適応： 入力音声に対する高精度なアライメント結果を得るために，入力音声の音響特徴量を適切にモデル化したアライメント用 HMM が必要である．入力話者の音声を事前に十分に入手できる場合，入力話者依存 HMM の構築が可能である．一方で本稿では，少量のみの音声が得られる場合を想定し，目標話者 HMM のモデル適応 [4] で得られる適応 HMM の使用について検討する．

多様な入力発話様式への対応： 本システムの用途の一つとして，創作活動支援が想定される．その場合，入力音声の発話様式は多岐に渡ることが多く，必ずしもアライメント用 HMM を構築するために用いた音声の発話様式と一致するとは限らない．仮に，両発話様式が大きく異なる場合は，アライメント精度が低下し，合成音声の品質低下が生じると予想される．この問題に対して，本稿では，入力音声に対するアライメント HMM の適応処理を導入し，その有効性を検証する．

3 実験的評価

3.1 実験条件

音声合成に用いる目標話者 HMM の学習データは，女性話者による ATR 音素バランス文 [7] A-I セット 450 文とする．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．スペクトルパラメータは，STRAIGHT 分析 [8] 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，音源パラメータは，対数 F_0 および 5 周波数帯域における平均非周期成分 [9] を使用する．HMM は 5 状態 left-to-right 型とする．音声入力を行う話者（入力話者）は，目標話者とは異なる男女各 2 名とする．各入力話者による ATR 音素バランス文 A-I セット 450 文から，入力話者毎にアライメント用話者依存 HMM を学習する．

---

*Improvements to HMM-Based Speech Synthesis System with Prosody Modification Based on Speech Input. by NISHIGAKI, Yuri, TAKAMICHI, Shinnosuke, TODA, Tomoki, NEUBIG, Graham, SAKTI, Sakriani, NAKAMURA, Satoshi (Nara Institute of Science and Technology)

3.2 継続長単位とアライメント用 HMM に関する実験結果

合成音声に反映させる継続長単位と，入力話者の音声を用いたモデル適応によるアライメント用 HMM 構築の効果を調査する．継続長単位として，状態継続長 (“state”)，音素継続長 (“phone”)，モーラ継続長 (“mora”) の 3 単位を用いる．また，アライメント用 HMM として，各入力話者に対する話者依存 HMM (“Ref.”) のみでなく，目標話者 HMM に対して，各入力話者による 1 発話（“1 utt.”）と 56 発話（“56 utts.”）を適応文としてモデル適応処理を施して得られる適応モデルも用いる．これら 3 種類の HMM に，適応元である目標話者 HMM（“Target”）を加えた計 4 種類のアライメント用 HMM と，3 種類の継続長単位の各組み合わせを用いて，合成音声の自然性に関する MOS 評価および，入力音声をリファレンスとした際の韻律の模倣性に関する DMOS 評価を行う．評価データは，ATR 音素バランス文 J セット 53 文である．評価人数は，10 人（MOS 評価）と 14 人（DMOS 評価）である．

各評価結果を，Fig. 2 に示す．自然性と模倣性の両面において，適応 HMM（“1 utt.” と “56 utts.”）のスコアは，目標話者 HMM（“Target”）のスコアを上回っていることがわかる．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 utts.”) は，入力話者 HMM (“Ref.”) 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使用すると，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．

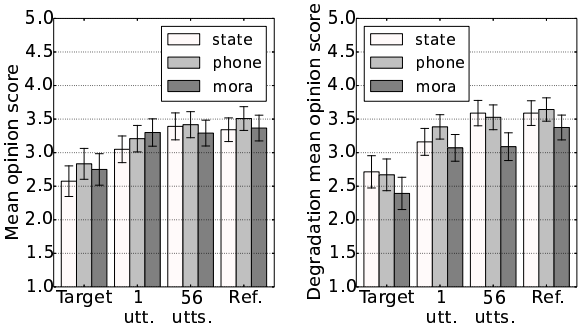
自然性と模倣性に関して，最高のスコアが得られている音素単位アライメントを用いた “Ref.” のスコアと，各継続長単位の “1 utt.” のスコアの間で t 検定を行った．その結果，自然性評価では状態単位とモーラ単位，模倣性評価ではモーラ単位を用いた場合に，それぞれ有意水準 1% にて有意差が認められた．故に，音素単位の使用が最も効果的であると判断する．

3.3 多様な入力発話様式への対応に関する実験結果

多様な発話様式を持つ入力音声に対する適応処理の影響を調査する．まず，テレビアニメーションとテレビドラマから有名なフレーズ・言い回しを持つ特徴的な音声 [1] を抽出し，次に，各入力話者に抽出音声を模倣させることで，入力音声を用意する．アライメント用 HMM には，“Ref.” と，各入力音声を適応データとして用いて “Ref.” 及び “Target” に対して適応処理を施した HMM を使用する．評価に用いる入力音声は 10 文である．継続長単位には音素単位を用いる．評価者は 8 名である．

韻律の模倣性に関する評価結果を，Fig. 3 に示す．入力音声を用いた適応 HMM（“1 utt. (Ref.)” と “1 utt. (Target)”）のスコアが，読み上げ音声で学習した HMM（“Ref.”）のスコアを上回っている．このことから，多様な発話様式を持つ入力音声に対しては，入力音声の発話毎に対する適応処理の有効であることが分かる．“1 utt. (Ref.)” と “1 utt. (Target)” では異なる HMM を適応に用いているが，有意なスコアの差は見られず，本適応処理は初期モデルに対して比較的頑健に動作する傾向が見られる．

[1]「海賊王に，オレはなる！」（出典： 作品名 ONE PIECE，原作者 尾田栄一郎，制作会社 東映アニメーション）や「見た目は子供，頭脳は大人，その名は，名探偵コナン！」（出典： 作品名 名探偵コナン，原作者 青山剛昌，制作会社 読売テレビ）など．



(a) 自然性に関する MOS 評価　(b) 韻律の模倣性に関する DMOS 評価

Fig. 2　モデル適応と継続長単位のための評価結果 (エラーバーは 95% 信頼区間)

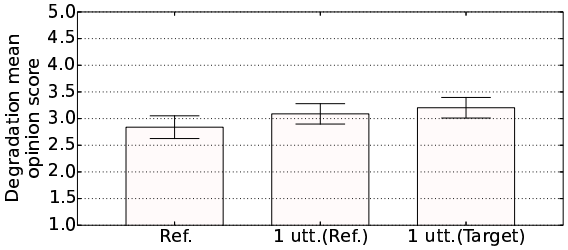


Fig. 3　多様な発話様式を持つ入力音声を用いた場合の，模倣性に関する DMOS 評価結果（エラーバーは 95% 信頼区間）

4　まとめ

本稿では，音声入力による韻律制御機能を有する HMM 音声合成システムの品質改善のため，(1) 合成音声に反映させる継続長の単位，(2) アライメント用 HMM に対するモデル適応処理，(3) 多様な発話様式を持つ入力音声へ対応について調査し，実験的評価で品質改善を確認した．今後は，多様な入力音声を用いた場合の模倣性を改善するため，入力音声のパワーや声質変化を合成音声に反映させる手法を検討する．

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．

参考文献

[1] 西垣 他，音講論（春），3-6-11，2014.
[2] K. Tokuda *et al.*, *Proc. ICASSP*, pp. 1315–1318, 2000.
[3] T. Toda *et al.*, *IEEE. Trans. on Audio, Speech and Language Processing*, Vol. 15, No. 8, pp. 2222–2235, 2007.
[4] J. Yamagishi *et al.*, IEICE Trans., Inf. and Syst., Vol. E90-D, No. 2, pp. 533–543, 2007.
[5] K. Yu *et al.*, *IEEE Trans. on Audio, Speech and Language*, Vol. 19, No. 5, pp. 1071–1079, 2011.
[6] 吉村 他，信学論 (D-II), Vol. J83-D-II, No. 11, pp. 2099–2107, 2000.
[7] 阿部 他，ATR テクニカルレポート，TR-I-0166, 1990.
[8] H. Kawahara *et al.*, *Speech Comm.*, Vol. 27, No. 3–4, pp. 187–207, 1999.
[9] Y. Ohtani *et al.*, *Proc. INTERSPEECH*, pp. 2266–2269, 2006.