# 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*

☆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

コーパスベース音声合成技術の発達により，特定のキャラクタ性を有する音声合成技術が構築され，所望の音声を創作する活動においてその利用が期待されている．特に，テキストから音声を合成する技術の一つであるHMM音声合成 [1] は，合成音声の特徴を柔軟に制御することが可能であるため，注目を集めている．音声特徴量やモデルパラメータの操作によって，合成音声を手動制御する処理が実現されている一方で，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．

歌声合成の分野においては，所望の歌声を歌声合成システムで作成するために，ユーザの歌声を参照して歌声合成システムの操作パラメータを最適化する枠組みが提案されている [2]．これと類似した枠組みとして，音声合成の分野においては，テキストと音声を入力として，HMMを用いて声質変換を行う手法 [3] が提案されている．しかしながら，この手法では，入力音声を利用する上での専用のコンテキスト要因を使用するため，従来のテキスト音声合成処理の精度を保持できる保証はない．

本稿では，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を制御する手法を提案する．入力される音声の韻律を合成音声へと反映させる際には，有声/無声区間補正処理を行う．補正処理に着目した実験的評価結果から，提案法の有効性を示す．

## 2 韻律制御機能を有する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Fig. 1に示す．入力テキストおよび入力音声を用いて，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．

### 2.1 システムの処理の流れ

音声合成用として目標話者のHMMを用い，アライメント用として音声入力を行うユーザのHMMを用いる．入力テキストに応じたアライメント用HMMにより，入力音声に対して状態アライメントを行うことで，入力音声の継続長を決定する．目標話者HMMに対して，入力テキストおよび入力音声の継続長を与えることで，入力音声の継続長を持つ合成音声パラメータ系列を生成する．なお，本稿では，入力音声の継続長は音素継続長で表し，音素内の状態継続長については尤度最大化基準により決定する [4]．

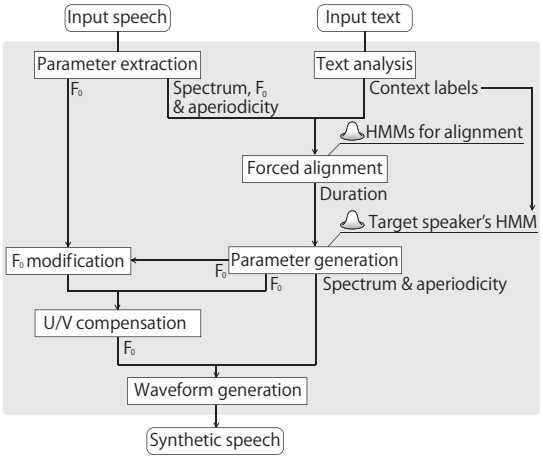


Fig. 1 音声による韻律制御処理のブロック図

次に，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，ユーザと目標話者の$F_0$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．

$$\hat{x_t} = \frac{\sigma_y}{\sigma_x}(x_t - \mu_x) + \mu_y \quad (1)$$

ただし$x_t$はフレーム$t$における入力音声の対数$F_0$，$\mu_x$と$\sigma_x$はそれぞれ$x_t$の平均と標準偏差，$\mu_y$と$\sigma_y$はテキスト音声合成部で生成した対数$F_0$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．

### 2.2 有声/無声区間の補正

入力音声の$F_0$系列に対して式（1）に示す補正処理のみを行った場合，有声/無声情報に関しては依然として入力音声に依存したものとなる．一方で，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 [5]．この際に，マイクロプロソディの除去も行う．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．本処理によって得られる$F_0$系列の一例をFig. 2に示す．

*HMM-Based Speech Synthesis System with Speech-driven Prosody Modification. by NISHIGAKI, Yuri, TAKAMICHI, Shinnosuke, TODA, Tomoki, NEUBIG, Graham, SAKTI, Sakriani, NAKAMURA, Satoshi (NAIST)

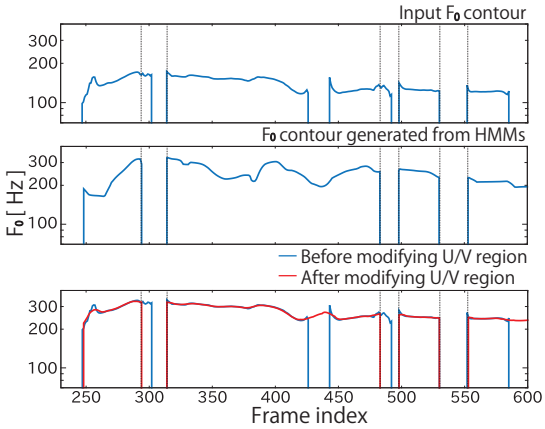


Fig. 2　各 $F_0$ 系列の図（上：入力音声の $F_0$ 系列，中：目標話者 HMM から入力音声の継続長に基づき生成される $F_0$ 系列，下：有声/無声区間の補正をする前の入力話者の $F_0$ 系列（青線）と補正した後の $F_0$ 系列（赤線））

Fig. 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．

## 3　実験的評価

### 3.1　実験条件

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 [6] 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 [7] 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 HMM を学習する．各入力話者による J セット 53 文を評価データとする．

有声/無声補正処理の効果を評価するために，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．また，各入力話者に対する有声/無声不一致率（U/V 不一致率）についても調査する．

### 3.2　実験結果

Fig. 3 に U/V 不一致率を，Fig. 4 にプリファレンススコアを示す．Fig. 3 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，Fig. 4 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．以上より，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．

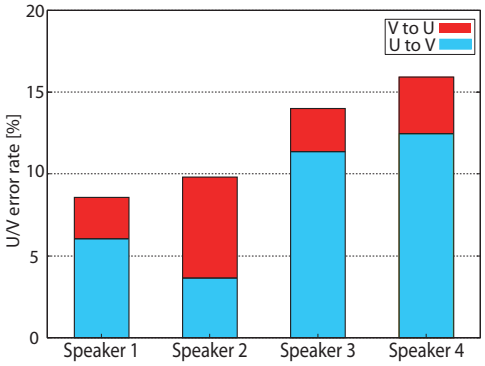


Fig. 3　有声/無声不一致率

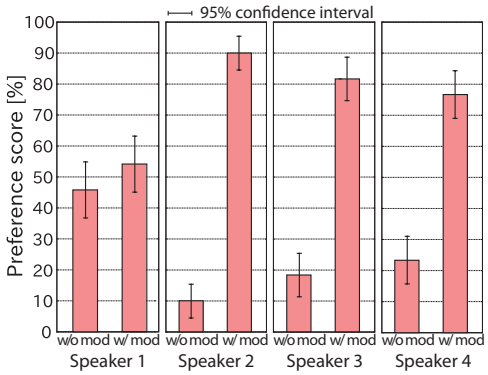


Fig. 4　自然性に関する主観評価結果

## 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 音声合成法を提案した．主に，韻律制御部における有声/無声区間の補正に着目し，実験的評価結果からその有効性を示した．今後は，テキスト音声合成部における継続長の単位についての検討，任意の入力話者への対応を行う．

**謝辞**　本研究の一部は，JSPS 科研費 22680016 の助成を受け実施したものである．

## 参考文献

[1] H. Zen *et al.*, *Speech Commun.*, 51(11), pp. 1039-1064, 2009.
[2] 中野 他，情処学論, Vol. 52, No. 12, pp. 3853-3867, December 2011.
[3] T. Nose *et al.*, *IEICE Trans. Inf. and Syst.*, Vol. E93-D, No. 9, pp. 2483–2490, Sept. 2010.
[4] 吉村 他, 信学論 (D-2), Vol. J83-D-2, pp. 2099–2107, 2000.
[5] K. Yu, *et al.*, *IEEE Trans. Audio, Speech and Language*, Vol. 19, No. 5, pp. 1071–1079, 2011.
[6] 阿部 他，ATR テクニカルレポート，TR-I-0166, 1990.
[7] H. Kawahara *et al.*, *Speech Commun.*, Vol. 27, No. 3–4, pp. 187–207, 1999.