論文 / 著書情報
Article / Book Information

論題(和文)	平均声と話者・スタイル適応を用いたスタイル制御法の検討
Title(English)	A study on style control technique for average-voice-based speech synthesis using speaker and style adap-
著者(和文)	橘 誠, 井澤信介, 能勢隆, 小林隆夫
Authors(English)	Makoto Tachibana, Shinsuke Izawa, Takashi Nose, Takao Kobayashi
出典(和文)	日本音響学会2007年春季研究発表会講演論文集, Vol. , No. , pp. 195-196
Citation(English)	Proceedings of the ASJ 2007 Spring Meeting, Vol. , No. , pp. 195-196
発行日 / Pub. date	2007, 3

平均声と話者・スタイル適応を用いたスタイル制御法の検討*

◎橘　誠, 井澤信介, 能勢隆, 小林隆夫 (東工大)

1　はじめに

我々は多様な発話様式・感情表現（スタイル）を合成音声で実現する手法の一つとして重回帰隠れセミマルコフモデル (HSMM) によるスタイル制御手法 [1] を提案し，合成音声のスタイルの表出・強調度合を直観的に制御できることを示した．しかしこの手法は制御対象とするスタイルをそれぞれ数百文章のデータで学習する必要があり，容易に任意の話者の声での音声合成を実現することや新たなスタイルを追加することが難しい．そこで，本研究ではスタイル制御を容易に任意の話者・スタイルで行うことを目的に，平均声と話者適応に基づく音声合成方式 [2] を用いた重回帰 HSMM の学習法を提案し，その有効性を検討する．

2　平均声と話者・スタイル適応を用いた重回帰 HSMM の学習

2.1　平均声モデルからのスタイル依存 HSMM の作成

Fig. 1 に提案手法のブロック図を示す．提案手法では，まず，複数の話者の「読上げ」スタイルの音声で学習した平均声モデルに話者・スタイル適応を行うことでスタイル毎のコンテキスト依存 HSMM（以下スタイル依存 HSMM）を作成する．話者・スタイル適応では目標話者・目標スタイルの少量の適応データを用いて平均声モデルのパラメータを変換する．このときモデルパラメータの共有構造は更新されないため，目標話者のスタイル依存モデルは，すべて平均声モデルと同じ共有構造となっている．

2.2　重回帰 HSMM によるスタイルのモデル化

重回帰 HSMM[1] では単一ガウス分布を仮定した状態 i の出力分布 $b_i(\boldsymbol{o})$ 及び状態継続長分布 $p_i(d)$ の平均 $\boldsymbol{\mu}_i$, m_i を次のような重回帰式で表す．

$$\boldsymbol{\mu}_i = \boldsymbol{H}_{b_i}\boldsymbol{\xi} \tag{1}$$

$$m_i = \boldsymbol{H}_{p_i}\boldsymbol{\xi} \tag{2}$$

$$\boldsymbol{\xi} = [1, v_1, v_2, \cdots, v_L]^\top = [1, \boldsymbol{v}^\top]^\top \tag{3}$$

ここで $\boldsymbol{v}$ はスタイルの表出・強調度合を表した低次元のベクトル (スタイルベクトル) であり，$\boldsymbol{H}_{b_i}$ および $\boldsymbol{H}_{p_i}$ はそれぞれ出力分布及び状態継続長分布のための回帰行列である．

2.3　回帰行列の初期値計算

各スタイル依存 HSMM のリーフノードの分布と適応データに対して設定したスタイルベクトルから，最小二乗基準により回帰行列の初期値を求める．なお，スタイル依存モデルの全てのリーフノードの分布が話者・スタイル適応の過程でスタイル毎に適切に変換されている場合には，適応データに含まれない分布についても計算が可能である．

2.4　回帰行列の MAP 推定

従来の重回帰 HSMM の学習 [1] では，回帰行列を EM アルゴリズムを用い最尤推定するが，本手法のように目標話者の各スタイルのデータが少量である場合には，回帰行列の推定精度が低下してしまう．そこで前節で求めた回帰行列の初期値 $\overline{\boldsymbol{H}}_{b_i}$ と MAP 推定 [3] を用いて，十分な適応データが得られる回帰行列のみを補正することとする．式 (1) および平均ベクトル $\boldsymbol{\mu}_i$ の MAP 推定量と ML 推定量の関係から MAP 推定量 $\boldsymbol{H}_{b_i}^{MAP}$ は，

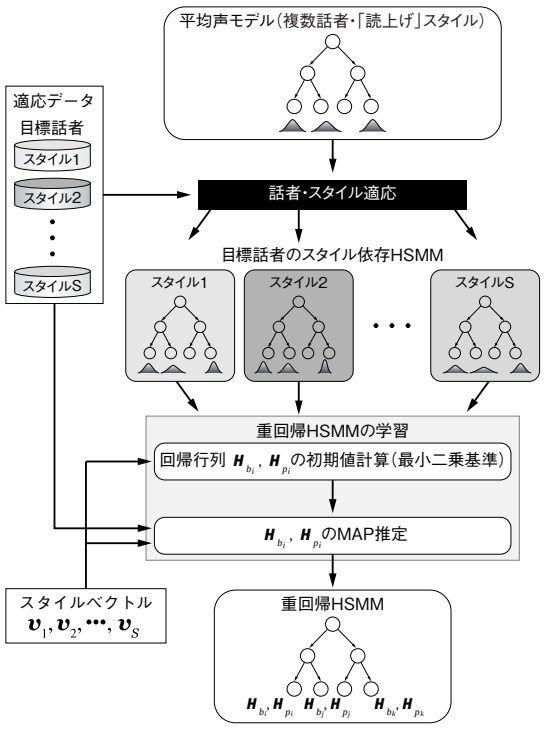

Fig. 1　平均声モデルを用いた重回帰 HSMM の学習

$$\boldsymbol{H}_{b_i}^{MAP} = \frac{\tau_{out}\overline{\boldsymbol{H}}_{b_i} + \Gamma_{out}(i)\boldsymbol{H}_{b_i}^{ML}}{\tau_{out} + \Gamma_{out}(i)} \tag{4}$$

$$\Gamma_{out}(i) = \sum_{n=1}^{K}\sum_{t=1}^{T^{(n)}}\sum_{d=1}^{t}\gamma_t^d(i)\cdot d \tag{5}$$

と表すことができる．ここで τ_{out} は MAP 推定におけるハイパーパラメータ，$\boldsymbol{H}_{b_i}^{ML}$ は最尤推定によって求めた回帰行列である．また，K は観測系列の総数，$T^{(n)}$ は n 番目の観測系列 $\boldsymbol{O}^{(n)}$ の総フレーム数，$\boldsymbol{o}_s^{(n)}$ は $\boldsymbol{O}^{(n)}$ の時刻 s における観測ベクトル，$\gamma_t^d(i)$ は状態 i で観測系列 $\boldsymbol{o}_{t-d+1}^{(n)}, \cdots, \boldsymbol{o}_t^{(n)}$ を出力する確率を表す．これにより，適応データが十分に得られる回帰行列に対しては最尤推定に近い結果を得ることができる．なお $\boldsymbol{H}_{p_i}$ についても同様に求めることができる．

3　実験

3.1　実験条件

音声の特徴ベクトルは，サンプリングレート 16kHz の音声信号を，フレーム長 25ms，フレーム周期 5ms のブラックマン窓を用いてメルケプストラム分析した

*A study on style control technique for average-voice-based speech synthesis using speaker and style adaptation. by TACHIBANA, Makoto, IZAWA, Shinsuke NOSE, Takashi and KOBAYASHI, Takao (Tokyo Institute of Technology)

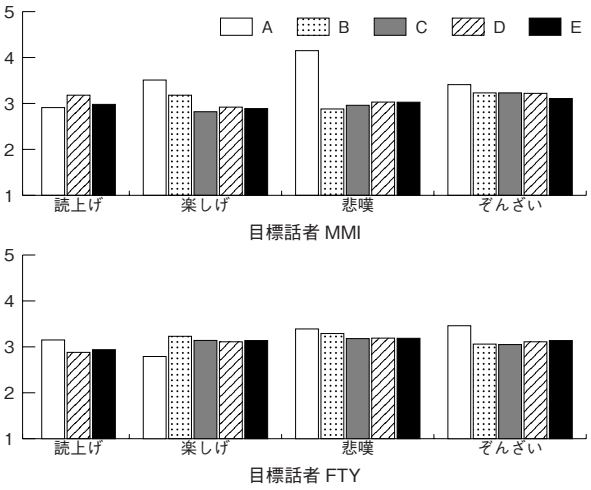

Fig. 2　話者性・スタイルの再現性の評価

0 次から 24 次のメルケプストラムと対数基本周波数，及びこれらの Δ および Δ^2 パラメータからなる 78 次元のベクトルとした．平均声モデルは ATR 日本語音声データベースセット B に含まれる男性 5 名，女性 4 名の各話者 450 文章，計 4050 文章を用いて STC[4] によるコンテキストクラスタリングを行い，SAT[2] を用いて学習した．目標話者・目標スタイルの音声データとして男性話者 MMI と女性話者 FTY が ATR 音韻バランス文をそれぞれ「読上げ」,「悲嘆」,「楽しげ」,「ぞんざい」のスタイルで発声したデータ [5] を用い，適応データとして，各スタイル 50 文章を用いた．話者・スタイル適応アルゴリズムは，話者適応における評価実験で良好な結果が得られた線形変換と MAP の組合せ手法 [6] を用い，線形変換には CSMAPLR[7] を使用した．スタイル空間は，文献 [1] と同様に「読上げ」を原点とし，その他のスタイルを独立な軸とした三次元空間を用い，適応データのスタイルベクトルは，スタイル毎に「読上げ」(0,0,0),「楽しげ」(1,0,0),「悲嘆」(0,1,0),「ぞんざい」(0,0,1) と設定した．主観評価の被験者は 8 名で，評価文章は学習・適応データに含まれない 53 文章とし，被験者毎にランダムな 8 文章を評価した．

3.2　話者性・スタイルの再現性の評価

話者・スタイル適応の効果および提案手法で学習した重回帰 HSMM の話者・スタイルの再現性を評価した．評価は A:スタイル毎に 450 文章で学習したスタイル依存モデル [5]，B:目標話者の「読上げ」450 文章で学習したモデルにスタイルのみ 50 文章でスタイル適応 [8]（ただし適応アルゴリズムは **3.1** と同様），C:平均声モデルを目標話者の「読上げ」50 文章で話者適応したモデルに，さらに各目標スタイルの 50 文章でスタイル適応したスタイル依存 HSMM，D:平均声モデルから目標話者・目標スタイルに話者性とスタイルを同時適応したスタイル依存 HSMM，E:D のモデルを用いて提案手法により学習した重回帰 HSMM の 5 種類のモデルから合成した音声について，目標スタイルの実際の発話の分析合成音を基準に「1：全く似ていない」から「5：リファレンスと同等」の 5 段階で評価した．E の合成時のスタイルベクトルは，適応データのスタイルベクトルと同じ値をスタイル毎に与えている．

Fig. 2 に結果を示す．平均声モデルに適応を行ったモデルにおいても，スタイル適応と同程度のスタイルの再現性が得られた．また C, D の結果から適応時の話者性とスタイルの区別はあまり必要ないと考えられる．さらに提案手法によって学習した重回帰 HSMM においても話者・スタイル適応を行った HSMM と同程度の再現性が得られた．

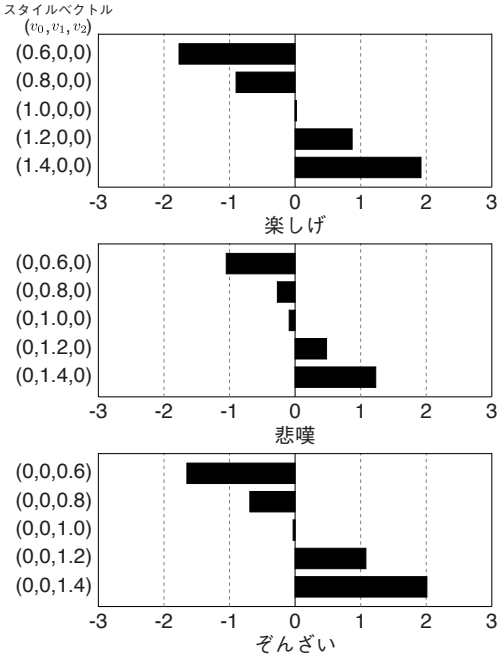

Fig. 3　スタイルの表出・強調度合の評価スコア

3.3　スタイルの表出度合の制御の評価

提案手法で学習した重回帰 HSMM において，合成時のスタイルベクトルをそれぞれ単独で 0.6〜1.4 の間を 0.2 間隔で 5 段階に変化させ，変化させたスタイルベクトルの軸に対応するスタイルの表出・強調度合の印象を D の合成音声を基準として -3 から 3 の 7 段階で評価した．Fig. 3 に話者 MMI において各スタイルベクトルを変化させた際の評価結果を示す．この結果から提案手法においても各スタイルベクトルの変化に従って，対応するスタイルの表出・強調度合の印象を変化可能であることが確認できた．

4　おわりに

本研究では，平均声と話者・スタイル適応を用いた重回帰 HSMM の学習法を提案し，主観評価から提案手法においても合成音声のスタイルを制御可能であることが示された．今後の課題としては，同様に少量の目標話者の音声データからスタイル制御が可能な話者適応手法 [9] との比較が挙げられる．

参考文献

[1] 能勢他, “重回帰 HSMM を用いた合成音声のスタイル制御,” 信学技報, SP2005-160, 61-66, 2006.
[2] Yamagishi *et al.*, “A training method of average voice model for HMM-based speech synthesis,” IEICE Trans. Fundamentals, E86-A(8), 1956-1963, 2003.
[3] Gauvain, Lee, “Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains,” IEEE Trans. Speech Audio Process., 2(2), 291-298, 1994.
[4] Yamagishi *et al.*, “A context clustering technique for average voice models,” IEICE Trans. Inf. & Syst., E86-D(3), 534-542, 2003.
[5] Yamagishi *et al.*, “Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis,” IEICE Trans. Inf. & Syst., E88-D(3), 502-509, 2005.
[6] 緒方他, “平均声に基づく音声合成における線形変換と MAP に基づく音響モデル学習法,” 信学技報, 106(333), 49-54, 2006.
[7] 中野他, “平均声に基づく音声合成のための話者適応アルゴリズムの評価,” 音講論 (春), 385-386, 2006.
[8] Tachibana *et al.*, “A style adaptation technique for speech synthesis using HSMM and suprasegmental features,” IEICE Trans. Inf. & Syst., E89-D(3), 1092-1099, 2006.
[9] 井澤他, “合成音声のスタイル制御における話者適応の検討,” 音講論 (秋), 255-256, 2006.