論文 / 著書情報
Article / Book Information

論題(和文)	重回帰HSMMに基づくスタイル推定とスタイル音声合成の検討
Title(English)	Style estimation and synthesis of speech based on multiple regression HSMM
著者(和文)	加藤陽一, 河野明文, 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. 267-268
Citation(English)	Proceedings of the ASJ 2007 Spring Meeting, Vol. , No. , pp. 267-268
発行日 / Pub. date	2007, 3

重回帰HSMMに基づくスタイル推定とスタイル音声合成の検討*

◎加藤陽一, 河野明文, 能勢隆, 小林隆夫 (東工大)

1 はじめに

我々はこれまで，音声合成における発話様式・感情表現 (以下スタイル) の表出・強調度合の制御を目的として，隠れセミマルコフモデル (HSMM) に基づく音声合成 [1] を拡張した重回帰 HSMM を用いた合成音声のスタイル制御法を提案した [2, 3]．重回帰 HSMM では，低次元のスタイル空間上のベクトル (スタイルベクトル) を用いて合成音声のスタイルの表出・強調度合を直観的に制御することが可能である．また，あらかじめ学習した重回帰 HSMM を用いて，入力音声のスタイルベクトルを最尤推定するスタイル推定法を提案し [4]，音声に含まれるスタイルの表出度合を定量的に推定できることを示した．しかし，従来のスタイル制御・スタイル推定法では，モデル学習時に与えるスタイルベクトルは各スタイル毎に一定としており，学習データ内での表出度合の変動は考慮されていなかった．そこで，本研究では，より学習データに即した重回帰 HSMM のモデル化を目的として，スタイル推定を用いたモデル学習法を提案し，スタイル推定およびスタイル音声合成において提案法の効果を検討する．

2 重回帰HSMMに基づくスタイルのモデル化とスタイル推定法

2.1 重回帰 HSMM によるスタイルのモデル化

重回帰 HSMM では，HSMM の各状態における出力分布 $\boldsymbol{\mu}_i$ および状態継続長分布の 平均 m_i がそれぞれ次のような重回帰で表されると仮定する．

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

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

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

ここで，$\boldsymbol{H}_{b_i}$ および $\boldsymbol{H}_{p_i}$ は $M \times (L+1)$ および $1 \times (L+1)$ 次元の回帰行列であり，M は $\boldsymbol{\mu}_i$ の次元である．また，$\boldsymbol{v}$ は各スタイルの表出・強調度合を表現する低次元のベクトル (スタイルベクトル) であり，この空間をスタイル空間と呼ぶ．モデル学習時には，学習データ $\{\boldsymbol{O}^{(1)}, \cdots, \boldsymbol{O}^{(K)}\}$ および対応するスタイルベクトル $\{\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(K)}\}$ を与えることで，EM アルゴリズムに基づく最尤推定により重回帰 HSMM のパラメータ $\boldsymbol{H}_{b_i}, \boldsymbol{H}_{p_i}$ を求めることが可能である [3]．

2.2 重回帰 HSMM に基づくスタイル推定

スタイル推定 [4] では，あらかじめ学習された重回帰 HSMM が与えられ，モデルパラメータ $\boldsymbol{H}_{b_i}, \boldsymbol{H}_{p_i}$ が固定されているとき，入力観測系列 $\boldsymbol{O}$ に対してスタイルベクトル $\overline{\boldsymbol{v}}$ を最尤推定する．

$$\overline{\boldsymbol{v}} = \underset{\boldsymbol{v}}{\operatorname{argmax}}\, P(\boldsymbol{O}|\lambda, \boldsymbol{v}) \tag{4}$$

本研究では，入力観測系列 $\boldsymbol{O}$ として 1 文章を与え，文章単位でスタイルベクトル $\overline{\boldsymbol{v}}$ を推定している．

3 スタイル推定を用いた重回帰HSMMの学習

従来の重回帰 HSMM のモデル学習では，学習データのスタイルベクトルの値は各スタイル毎に固定であると仮定していた．しかし，実際には同一のスタイル内でも，その表出度合は文章毎，文節毎などで異なっていると考えられる．そのため，それぞれの学習データに対して適切なスタイルベクトルを与えることによって，より学習データに即した重回帰 HSMM のモデル化が可能であると考えられる．

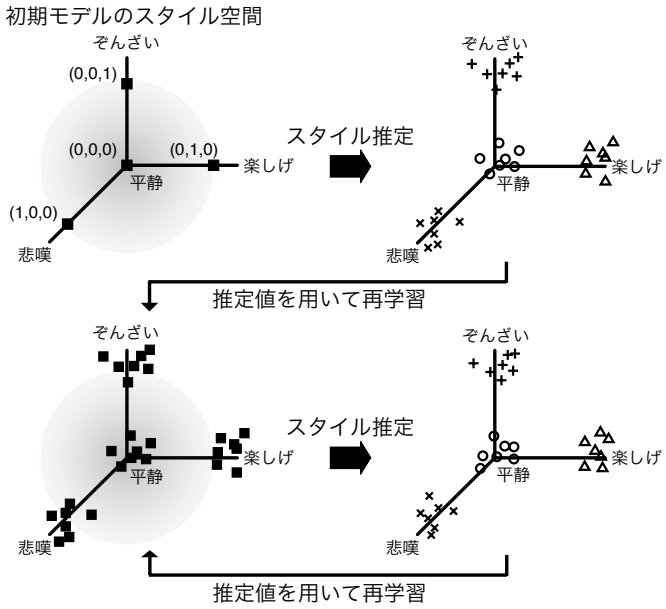

Fig. 1 スタイル推定を用いた重回帰 HSMM の学習

ここでは，まず初期モデルとしてスタイルベクトルをスタイル毎に一定として学習を行う．次に，このモデルを用いて全学習データに対して文章毎にスタイル推定を行い，得られたスタイルベクトルの値を用いて再びモデルの学習を行う (Fig. 1)．さらに，この推定と学習を繰り返すことにより，モデルの改善を図る．

4 実験

4.1 実験条件

実験では，プロの男性ナレーター MMI が ATR 日本語音韻バランス文 503 文章を「読上げ」「悲嘆」「楽しげ」「ぞんざい」の 4 つのスタイルで発声した音声を使用した．モデル学習には各スタイル 450 文章，計 1800 文章を用いた．重回帰 HSMM は，文献 [5] と同様のコンテキスト情報を用いたコンテキスト依存モデルである．

スタイル空間は原点を「読上げ」スタイルとし，他のすべてのスタイルを互いに独立であると仮定した 3 次元空間とした．初期モデルの学習時には，従来法と同様に,「読上げ」(0,0,0),「悲嘆」(1,0,0),「楽しげ」(0,1,0),「ぞんざい」(0,0,1) のスタイルベクトルを与えた．続いて，**3** で述べた推定と学習の繰返しを 10 回行った．

4.2 スタイルベクトルの推定結果の変化

Fig. 2 に，学習の繰返し回数と全学習データの推定結果の変化量の平均と標準偏差を実線で，各繰返しにおける最大の変化量を点線でそれぞれ示す．横軸

*Style estimation and synthesis of speech based on multiple regression HSMM, by KATO, Yoichi, KOUNO, Akifumi, NOSE, Takashi, and KOBAYASHI, Takao (Tokyo Institute of Technology).

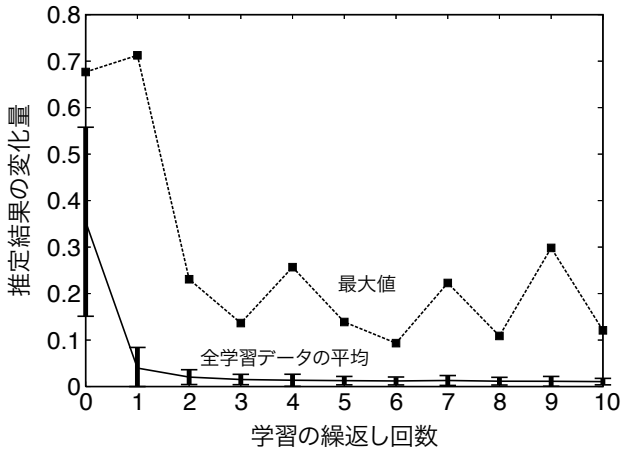

Fig. 2　スタイルベクトルの推定結果の変化

は学習の繰返し回数を表し，縦軸は前段階からの推定値の変化量を表す．変化量は，繰返し前後で推定したスタイルベクトル間のユークリッド距離を用いた．また，繰返し回数が 0 のときの値は，初期モデル学習時に設定したスタイルベクトルの値と，そのモデルを用いて学習データを推定した値との変化量である．

この結果より，学習データの変化量の平均は，繰返し 1 回でほぼ一定値となっており，スタイルベクトルの初期値との大きなずれは収まっている．しかし，最大値は繰返し 3 回目以降も振動しており，必ずしも収束していない．この原因の一つとして，スタイル推定を文章単位で行っているため，文章内におけるスタイルの変動が考慮されていないことが考えられ，より詳細な検討が必要である．

4.3　スタイル推定を用いた合成音声の主観評価

次に，スタイル推定が適切に行われているかを確認するため，スタイルベクトルを初期モデル学習時の値に設定した合成音声と，スタイル推定により得られたスタイルベクトルの値を用いた合成音声との比較評価を行った．合成時のモデルは，初期モデル (繰返しなし) を使用した．評価音声は，学習データに含まれる文章 (close) と含まれない文章 (open) の中から，それぞれスタイル推定結果において各スタイルで表出度合の最も大きい 4 文章と最も小さい 4 文章の計 8 文章を使用した．被験者は 7 名である．各被験者に参照音声として実際の発話の分析合成音を与え，それぞれの評価音声に対してどちらがより参照音声に近いかを選択させた．

結果を Fig. 3 に示す．この結果から，close 文章の「悲嘆」「楽しげ」の合成音声については推定値を用いた場合の方がややスコアが高くなっており，これらのスタイルに対しては適切なスタイル推定が行われていることがわかる．ただし，表出度合の小さい評価音声については，合成時のスタイルベクトルの値を小さくすることで，そのスタイルの特徴の表出が弱くなるため，参照音声に近いと判断されないことがあった．

一方，open の文章の合成音声においては，スタイルの再現性の改善は見られなかった．これは本実験での open の文章数が 53 文章と close の文章数に比べて少量であり，close の評価音声のスタイル表出度合の幅が「悲嘆」約 0.5〜1.4,「楽しげ」約 0.6〜1.2,「ぞんざい」約 0.6〜1.3 であるのに対し，open の場合はそれぞれ約 0.9〜1.3，0.7〜1.1，0.5〜1.1 と，スタイル表出度合の大きく異なるものを選ぶことができなかったことが一因であると考えられる．

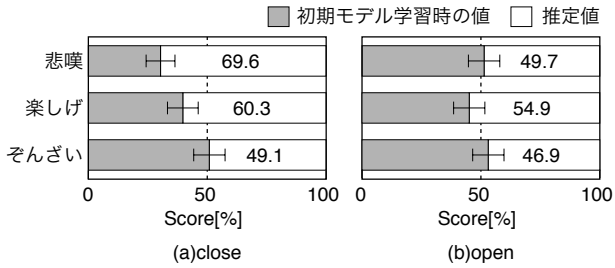

Fig. 3　スタイル推定を用いた合成音声の主観評価

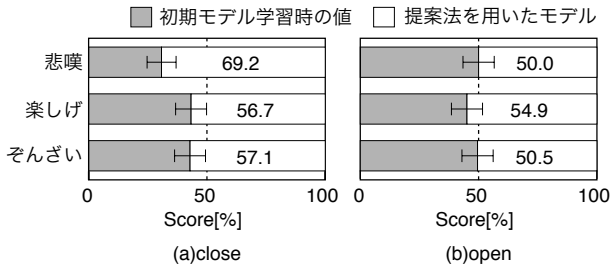

Fig. 4　提案法における合成音声の主観評価

4.4　推定を用いた学習モデルによる合成音声の主観評価

最後に，初期モデル (繰返しなし) と，提案法を用いて学習したモデルの合成音声の比較評価を行った．提案法のモデルは，**4.2** の結果から 1 回目以降の繰返し学習によるモデルの変化は小さいと考えられるため，推定と学習を 1 回繰り返したモデルを使用した．実験方法は **4.3** と同様である．

結果を Fig. 4 に示す．「ぞんざい」の close の文章に対するスタイル再現性の向上が若干見られた他は，**4.3** とほぼ同様の結果となった．

5　おわりに

本研究では，重回帰 HSMM においてより学習データに即したモデル化を実現するために，スタイル推定法を用いたモデル学習法を提案し，スタイル推定および合成音声の評価実験を行った．その結果，学習に用いたスタイル音声に対してスタイル再現性の向上が見られ，学習データに即したモデル化が行えることがわかった．一方，学習外のスタイル音声の再現性については有意な差異が見られなかったが，見方を変えると，スタイル制御に基づく音声合成におけるモデル学習では，スタイルベクトルをスタイル毎に固定して学習しても，その影響は小さいとも言うこともできる．

今後の課題としては，プロのナレーターによる模擬的なスタイル音声とは異なるデータを用いた場合における提案法の効果の検討が挙げられる．また，学習データの文節単位でのスタイル推定を用いたモデル学習法も今後の課題である．

参考文献

[1] Zen *et al.*, “Hidden semi-Markov model based speech synthesis,” Proc. ICSLP 2004, 1397-1400, 2004.
[2] 能勢他, “重回帰 HSMM を用いた音声のスタイル制御法の検討,” 音講論 (秋), 287-288, 2005.
[3] 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, 2006.
[4] 能勢他, “重回帰 HSMM に基づく音声の発話様式・感情表現の推定,” 音講論 (春), 219-220, 2006.
[5] Yamagishi *et al.*, “Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis,” IEICE Trans. Inf. & Syst., E88-D (3), 503-509, 2005.