社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－SLP－74（7）
2008 / 12 / 9

重回帰HMMに基づくスタイル推定を用いた音声認識における音響モデル学習法

井島 勇祐† 橘 誠† 能勢 隆† 小林 隆夫†

† 東京工業大学 大学院総合理工学研究科
〒 226-8502 横浜市緑区長津田町 4259-G2-4
E-mail: †{yusuke.ijima,makoto.tachibana,takashi.nose,takao.kobayashi}@ip.titech.ac.jp

あらまし 本論文では，重回帰 HMM に基づくスタイル推定を用いた音声認識手法において，この手法を容易に任意の話者へ適用することを目的に，重回帰 HMM の学習に話者非依存モデルとモデル適応手法を導入する手法を提案する．提案法では，まず話者非依存モデルに目標話者の各スタイルの少量の適応データを用いて，話者性とスタイルの同時適応を行い，重回帰 HMM の学習に用いる目標話者のスタイル適応 HMM を作成する．得られたスタイル適応 HMM のモデルパラメータと発話様式・感情表現（スタイル）の表出度合を表すスタイルベクトルから，最小二乗法により重回帰 HMM の回帰行列を求め，最尤推定により補正を行う．プロのナレータと一般の発話者が発話した模擬感情音声に対して音素認識実験を行い，その結果から提案法の性能評価を行う．また，提案法ではスタイル推定の結果から，認識結果だけでなく入力音声のスタイルも得られることを示す．

キーワード 音声認識，重回帰 HMM，話者適応，スタイル適応，スタイル推定

Acoustic Model Training Technique for Speech Recognition Using Style Estimation with Multiple-Regression HMM

Yusuke IJIMA†, Makoto TACHIBANA†, Takashi NOSE†, and Takao KOBAYASHI†

† Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology,
4259-G2-4, Nagatsuta-cho, Midori-ku, Yokohama-shi, 226-8502, Japan
E-mail: †{yusuke.ijima,makoto.tachibana,takashi.nose,takao.kobayashi}@ip.titech.ac.jp

Abstract We propose a technique for emotional speech recognition based on multiple-regression HMM (MRHMM). To achieve emotional speech recognition for an arbitrary speaker with a small amount of training data, we incorporate a speaker and style adaptation technique into speaker-dependent MRHMM-based emotional speech recognition. In the proposed technique, we first adapt the speaker-independent model to target speaker's respective styles with a small amount of speech data. Then, using obtained speaker- and style-adapted HMMs and low-dimensional style control vector for each training style, the regression matrices of MRHMM are estimated based on least square method and maximum likelihood estimation. We assess the performance of the proposed technique on the recognition of acted emotional speech uttered by both professional narrators and non-professional speakers and show the effectiveness of the technique.

Key words speech recognition, multiple-regression HMM(MRHMM), speaker adaptation, style adaptation, style estimation

1. はじめに

音声には，言語情報だけでなく，発話様式や感情表現といったパラ言語情報も含まれている．現在の音声認識システムは，読上げ調で発話した音声に対しては非常に高い認識性能が得られているが，感情音声や自然発話などに対しては，音声の音響的特徴量が変動してしまうため，読上げ調の音声と同等の性能は得られていない．この問題を解決する手法として，それぞれの発話様式・感情表現の変動毎にモデルを用意することが考えられる．これは，発話様式・感情表現の変動の幅が限られてお

り，その変動が予測できる場合には有効であると考えられるが，実際には感情表現や発話様式の表出・強調度合は発話によって大きく変動してしまう．そのため，考えられる全ての変動に対して，モデルを用意することは現実的ではない．

この問題を解決するための手法として，モデル適応を利用する手法が考えられるが，感情表現や発話様式は発話毎や発話内でも変動するため，発話毎などの短い区間でオンラインのモデル適応を行うことが望ましい．そのために一文章，一発話といった非常に少量のデータを用いて，モデル適応を行う必要があると考えられる．そこで，我々は発話様式・感情表現を含んだ音声を認識することを目的に，少量のデータから推定可能な低次元のパラメータを用いた高速なモデル適応手法を提案し，入力音声の発話様式・感情表現とマッチしたモデルを使用した場合と同等の認識性能が得られることを示した[1]．この手法は，重回帰 HMM [2] を利用し，まず重回帰 HMM に基づくスタイル推定手法 [3] を用いて，入力音声の発話様式や感情表現の表出度合を表す低次元のベクトル（スタイルベクトル）を推定する．そして，推定したスタイルベクトルを用いて，HMM の出力確率密度関数の新しい平均ベクトルを計算し，得られた HMM を用いて，通常の音声認識を行う．この手法は，低次元のパラメータに基づいてモデル適応を行うという意味で固有声法 [4] と同様であるが，一連の音声認識の過程において，認識結果である言語情報だけでなく，パラ言語情報である入力音声のスタイルの表出度合も得ることができるという特長がある．

しかし，従来法 [1] では，重回帰 HMM の学習に目標話者の多量の音声データが必要であるため，任意の話者へ適用することは現実的ではなかった．この問題を解決するためには，不特定話者重回帰 HMM を用いることが考えられるが，発話様式や感情表現は話者によって変動してしまうため，十分な認識性能は得られない可能性がある．

そこで本論文では，この手法を容易に任意の話者へ適用することを目的として，重回帰 HMM の学習に話者非依存モデルとモデル適応手法を利用する手法を提案する．また，プロのナレータと一般の発話者が発話した模擬感情音声に対して，音素認識実験の結果から，提案法の性能を評価する．

2. 重回帰 HMM に基づく音声認識手法

2.1 重回帰 HMM に基づくスタイルのモデル化

重回帰 HMM に基づくスタイルのモデル化では，出力確率密度関数の平均ベクトルは発話様式・感情表現（スタイル）の表出度合を表す低次元の空間上のベクトル（スタイルベクトル）の関数により表される．

HMM の各状態における出力確率密度関数の平均ベクトルをそれぞれ $\boldsymbol{\mu}_i$ とすると，重回帰 HMM では平均ベクトルを次のような重回帰式で表す．

$$\boldsymbol{\mu}_i = \boldsymbol{h}_0^{(i)} + \boldsymbol{A}_i \boldsymbol{v} = \boldsymbol{H}_i \boldsymbol{\xi} \tag{1}$$

$$\boldsymbol{H}_i = [\boldsymbol{h}_0^{(i)}, \cdots, \boldsymbol{h}_L^{(i)}] \tag{2}$$

$$\boldsymbol{A}_i = [\boldsymbol{h}_1^{(i)}, \cdots, \boldsymbol{h}_L^{(i)}] \tag{3}$$

$$\boldsymbol{\xi} = [1, \boldsymbol{v}^\top]^\top \tag{4}$$

ここで，$\boldsymbol{v} = [v_1, \cdots, v_L]^\top$ はスタイルベクトルである．また，$\boldsymbol{H}_i$ は $M \times (L+1)$ 次元の回帰行列であり，M は $\boldsymbol{\mu}_i$ の次元数である．

十分な学習データが得られる場合，モデル学習時に学習データおよび対応するスタイルベクトルを与えることで，EM アルゴリズムに基づく最尤推定により重回帰 HMM のパラメータである回帰行列 $\boldsymbol{H}_i$ と共分散行列 $\boldsymbol{\Sigma}_i$ を推定することができる [5]．

2.2 スタイル推定に基づく音響モデルのオンライン適応

あらかじめ学習された重回帰 HMM が与えられ，モデルパラメータ $\boldsymbol{H}_i$，$\boldsymbol{\Sigma}_i$ が固定されているとき，入力音声に対して最適なスタイルベクトルを最尤推定することが可能である [3]．

スタイル推定では，重回帰 HMM を λ，入力観測系列を $\boldsymbol{O} = (\boldsymbol{o}_1, \cdots, \boldsymbol{o}_T)$ として，次式で定義される最適なスタイルベクトル $\overline{\boldsymbol{v}}$ を求める．

$$\overline{\boldsymbol{v}} = \underset{\boldsymbol{v}}{\operatorname{argmax}}\, P(\boldsymbol{O}|\lambda, \boldsymbol{v}) \tag{5}$$

$$\overline{\boldsymbol{v}} = [\overline{v_1}, \cdots, \overline{v_L}]^\top \tag{6}$$

このとき，スタイルベクトルの再推定式は次式で表される [3]．

$$\overline{\boldsymbol{v}} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \gamma_t(i) \boldsymbol{A}_i^\top \boldsymbol{\Sigma}_i^{-1} \boldsymbol{A}_i \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \gamma_t(i) \boldsymbol{A}_i^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{o}_t - \boldsymbol{h}_0^{(i)}) \right) \tag{7}$$

ここで，T は入力観測系列 $\boldsymbol{O}$ のフレーム数，N は状態数，$\gamma_t(i)$ は時刻 t において i 番目の状態に滞在する確率である．本研究では，入力観測系列 $\boldsymbol{O}$ として一文章を与え，文章単位でスタイルベクトル $\overline{\boldsymbol{v}}$ を推定している．

そして，重回帰 HMM に推定したスタイルベクトル $\overline{\boldsymbol{v}}$ を与えることで，式 (1) により新しい平均ベクトルをもつ HMM を得ることができる．この得られたモデルは，入力音声のスタイルに適応されたモデルであると考えることができる．

2.3 話者非依存モデルとモデル適応手法を用いた重回帰 HMM の学習法

重回帰 HMM の学習には，一般的に目標話者が発話した各スタイル毎に数十分程度の音声データを使うことが望ましい．しかし，不特定の話者に対し，個別に多量の音声データを用意することは現実的ではない．一方，我々は重回帰 HSMM に基づく音声のスタイル制御手法，スタイル推定手法において，平均声モデルと話者・スタイルの同時適応手法を利用する手法を提案し，目標話者の少量の音声データから，重回帰 HSMM を学習する手法を提案している [6, 7]．そこで本研究では，重回帰 HMM の学習に同様な手法を導入する．

図 1 に重回帰 HMM の学習部のブロック図を示す．まず，複数の話者が読上げスタイルで発話した十分な量の音声データを用いて，話者非依存モデル（SI モデル）を学習する．次に，事前に目標話者が発話した各スタイルの少量の音声データより，

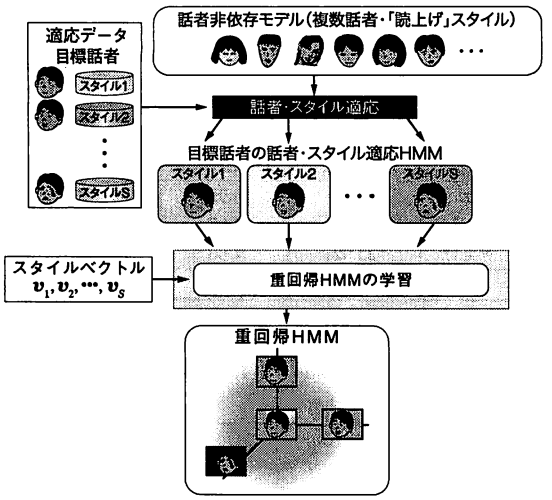


図 1 話者非依存モデルとモデル適応手法に基づく重回帰 HMM の学習法

Fig. 1 MR-HMM training based on speaker-independent model and model adaptation.

SI モデルから目標話者の各スタイルへ話者性とスタイルの適応を行う．そして，得られた話者・スタイル適応 HMM から最小二乗法と最尤推定により，目標話者の重回帰 HMM を得る．

最小二乗法においては，適応データのスタイルの総数を S とし，各スタイルに適応したモデルの出力分布の平均ベクトルを $\boldsymbol{\mu}_i^{(s)}$ $(1 \leqq s \leqq S)$，各スタイルに設定するスタイルベクトルを $\boldsymbol{\xi}^{(s)}$ $(1 \leqq s \leqq S)$ とするとき，次式で定義される二乗誤差

$$E = \sum_{s=1}^{S} \left\| \boldsymbol{\mu}_i^{(s)} - \boldsymbol{H}_i \boldsymbol{\xi}^{(s)} \right\|^2 \tag{8}$$

を最小化する $\boldsymbol{H}_i$ を求める [6, 7]．すなわち，E を $\boldsymbol{H}_i$ で微分して 0 とおき，$\boldsymbol{H}_i$ について解くことにより，回帰行列の初期値

$$\overline{\boldsymbol{H}}_i = \left(\sum_{s=1}^{S} \boldsymbol{\mu}_i^{(s)} \boldsymbol{\xi}^{(s)\top} \right) \left(\sum_{s=1}^{S} \boldsymbol{\xi}^{(s)} \boldsymbol{\xi}^{(s)\top} \right)^{-1} \tag{9}$$

を得る．

さらに，最小二乗法により得られた回帰行列の初期値を次式によって補正する [7]．

$$\boldsymbol{H}_i = \frac{\tau \overline{\boldsymbol{H}}_i + \Gamma_i \boldsymbol{H}_i^{ML}}{\tau + \Gamma_i} \tag{10}$$

ここで，$\overline{\boldsymbol{H}}_i$ は式 (9) により得られる回帰行列の初期値，$\boldsymbol{H}_i^{ML}$ は適応データから最尤推定により推定される回帰行列である．また，τ は補正の重みを調整するパラメータであり，Γ_i は次式により求められる．

$$\Gamma_i = \sum_t \gamma_t(i) \tag{11}$$

これにより，状態 i において適応データが十分に得られる場合，得られる回帰行列 $\boldsymbol{H}_i$ は最尤推定により得られる $\boldsymbol{H}_i^{ML}$ と近いものとなる．

2.4 重回帰 HMM に基づく音声認識手法

重回帰 HMM に基づく音声認識は次のように行う．まず，入力音声のスタイルベクトルを文章毎に推定する．次に，推定されたスタイルベクトルを用いて，重回帰 HMM から認識用の HMM を得る．そして，得られた HMM を用いて通常の音声認識を行う．なお，スタイル推定を行う際には，入力音声のラベル情報が必要となる [1, 3] ため，本研究では 2 パス方式で認識を行っている．

話者非依存モデルの学習：

Step 0 複数の話者が発話した読上げの音声データを用いて，話者非依存モデルを学習する．

重回帰 HMM の学習：

Step 1 話者・スタイル適応により，話者非依存モデルを目標話者の各スタイルのモデルへ適応する．

Step 2 式 (9) より，目標話者の重回帰 HMM の初期値を得る．

Step 3 式 (10) を用いて，重回帰 HMM のパラメータを補正する．

重回帰 HMM に基づく音声認識：

Step 4 学習した重回帰 HMM にスタイルベクトル $\mathbf{0}$ を与え，読上げスタイルの HMM を得る．

Step 5 得られた HMM を用いて，入力音声に対して音素認識を行う．

Step 6 **Step 5** で得られた音素列を用いて，入力音声のスタイルベクトル $\overline{\boldsymbol{v}}$ を推定する．

Step 7 重回帰 HMM に推定したスタイルベクトル $\overline{\boldsymbol{v}}$ を与え，新しい平均ベクトルをもつ HMM を得る．

Step 8 得られた HMM を用いて音声認識を行い，認識結果を得る．

3. 実　　験

3.1 実験条件

実験には，プロのナレータと一般の発話者が発話した音声データを用いた．プロのナレータの音声データベースは，文献 [1] で使用したものと同一であり，男性話者 MMI，MJI と女性話者 FTY が ATR 音韻バランス文 503 文章を「読上げ」,「悲嘆」,「楽しげ」の 3 スタイルで発話した音声データを使用した．なお，この音声は実際のスタイルを含んだ自発音声ではなく，模擬スタイルで読み上げた音声である．一般の発話者のデータは，男性 8 名，女性 1 名の計 9 名の学生が ATR 音韻バランス文 503 文章のうち，サブセット A，I の 100 文章を「読上げ」,「悲嘆」,「楽しげ」,「怒り」の 4 スタイルを発話した模擬感情音声を収録したものである．プロのナレータ，一般の発話者ともに音声は防音室で収録し，スタイルの表出度合は各スタイル内であまり変化することがないよう収録時に指示している．

音声の特徴ベクトルは，フレーム長 25 ms，フレームシフト

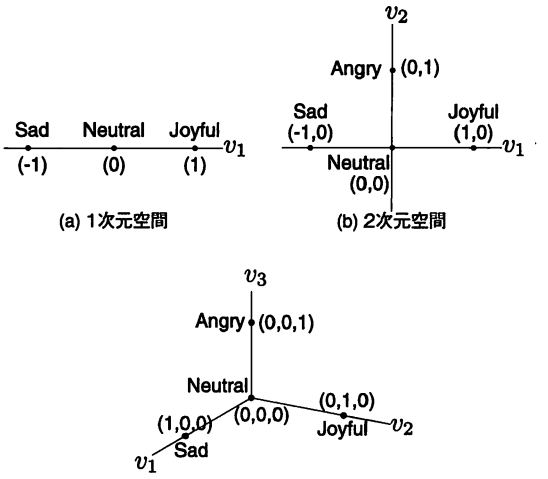


図 2 実験に使用したスタイル空間
Fig. 2 Style spaces for MRHMM.

10 ms のハミング窓で分析し，1 次から 12 次の MFCC と対数パワー，及びこれらの Δ パラメータからなる 26 次元とした．音声信号のサンプリングレート 16 kHz である．また，使用した音素数は無音，ポーズを含む 42 個である．モデル化には 1 混合 3 状態 left-to-right 型のトライフォン HMM を使用した．モデルのパラメータ共有は，決定木に基づくクラスタリングにより行い，決定木の構造は，**3.2** に示す特定話者重回帰 HMM を除き，全て同一のものとなっている．

話者非依存モデルの学習には，ATR 日本語音声データベースセット B に含まれる男性 6 名，女性 4 名が読上げスタイルで発話した各話者 450 文章，計 4500 文章の音声データを使用した．なお，この 10 名は，先に述べたプロのナレータと一般の発話者とは異なる話者である．

話者・スタイル適応には，目標話者が発話した各スタイル 5 文章（20 秒程度）を使用した．なお，適応データの選択による結果の偏りを軽減するために，適応データはランダムに選択した上で適応データを変更し，2 回実験を行っている．話者・スタイル適応におけるモデル適応手法は，最尤線形回帰（MLLR）と MAP 推定の組合せ手法 [8] を使用した．MLLR では，目標話者・スタイルの適応データが少量であることから，全ての分布で一つの回帰行列を共有している．また，重回帰 HMM のパラメータ補正の重み係数である τ は予備実験の結果から，$\tau = 100$ とした．

3.2 話者・スタイル適応の適応性能

まず，話者・スタイル適応の性能を評価するために，話者非依存モデル（SI-HMM），特定話者重回帰 HMM（SD-MRHMM），話者・スタイル適応重回帰 HMM（SA-MRHMM）の比較を行った．音声データは，プロのナレータが発話した「読上げ」，「悲嘆」，「楽しげ」の 3 スタイルのデータを使用し，重回帰 HMM のスタイル空間は図 2(a) に示す 1 次元空間を使用した．目標話者の SD-MRHMM は，文献 [1] と同一の手法で学習し，各スタイル 450 文章，計 1350 文章を使用し学習を行った．SA-MRHMM の学習の際には，各スタイル 5 文章を用いて，SI-HMM から目標話者の各スタイルに話者・スタイル適応を行っている．評価データには，モデル学習に用いない 50 文章を使用し，10-fold クロスバリデーションを行った．

表 1 SI モデル，SD-MRHMM，SA-MRHMM における音素誤り率の比較
Table 1 Comparison of phoneme error rates (%) for SI-HMM, SD-MRHMM, and SA-MRHMM.

入力スタイル	モデル		
	SI-HMM	SD-MRHMM	SA-MRHMM
読上げ	20.24	5.91	10.77
悲嘆	27.95	7.07	14.51
楽しげ	23.92	7.60	15.15
平均	24.03	6.87	13.48

表 2 一般の発話者における音素誤り率 (%)
Table 2 Phoneme error rates (%) for non-professional speakers' emotional speech.

	SI-HMM	2 次元空間	3 次元空間
読上げ	22.69	15.89	15.75
悲嘆	29.17	19.33	19.41
楽しげ	25.14	18.64	18.67
怒り	30.32	22.40	22.37
平均	26.83	19.06	19.05

表 1 に，3 名の話者の平均音素誤り率を示す．音素誤り率は次式により求めている．

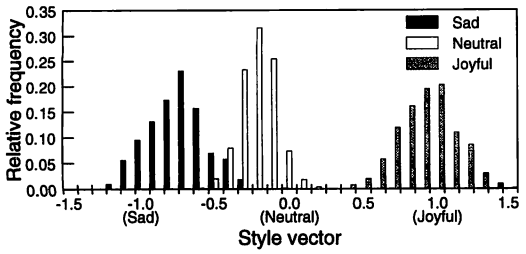
$$error(\%) = \left(1 - \frac{H}{H+D+S}\right) \times 100 \qquad (12)$$

ここで，H，S，D はそれぞれ正解音素数，置換誤り数，脱落誤り数である．SA-MRHMM の誤り率は，SD-MRHMM と比べ大きくなっているが，SI-HMM と比較すると大幅に誤り率が減少していることがわかる．

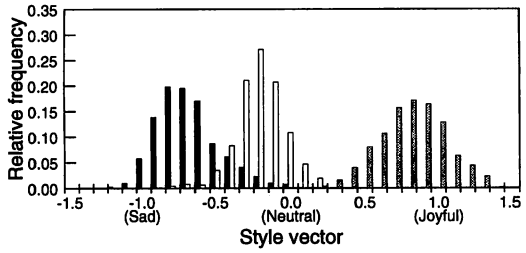
次に，SD-MRHMM，SA-MRHMM を用いた場合の，入力音声のスタイルベクトルの推定値の分布を求めた．図 3，図 4 に，それぞれ SD-MRHMM，SA-MRHMM を用いて推定されたスタイルベクトルのヒストグラムを示す．全ての結果において，スタイルごとにスタイルベクトルの推定値の分布が異なっており，学習時に各スタイルに与えたスタイルベクトルの値に近い位置に分布していることがわかる．また，SA-MRHMM は学習に各スタイル 5 文章しか使用していないにも関わらず，各スタイル 450 文章を用いて学習した SD-MRHMM と同様な分布が得られている．

3.3 一般の発話者における性能評価

次に，一般の発話者の音声データを使用し，提案法の性能評価を行った．使用した音声データは，「読上げ」，「悲嘆」，「楽しげ」，「怒り」の 4 スタイルで，重回帰 HMM のモデル化には，2 次元空間（図 2(b)）と 3 次元空間（図 2(c)）のスタイル空間を使用した．評価データには，SI モデルの学習と話者・スタイル適応に用いていない 50 文章を使用し，2-fold クロスバリ

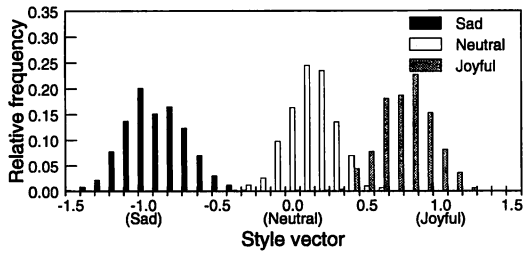


(a) SD-MRHMM

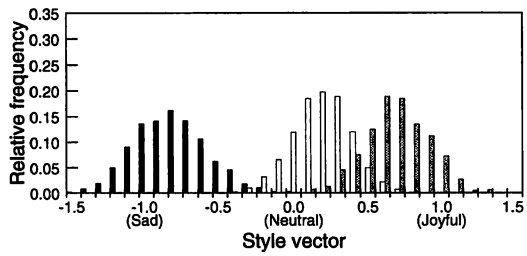


(b) SA-MRHMM

図 3 スタイルベクトルの推定値のヒストグラム（男性話者 MMI）
Fig. 3 Histograms of the estimated values of the style vector (male speaker MMI).



(a) SD-MRHMM



(b) SA-MRHMM

図 4 スタイルベクトルの推定値のヒストグラム（女性話者 FTY）
Fig. 4 Histograms of the estimated values of the style vector (female speaker FTY).

デーションを行った．

表 2 に，9 名の話者の平均音素誤り率を示す．全ての場合において，2 次元空間，3 次元空間の重回帰 HMM は，SI-HMM よりも音素誤り率が大幅に低いことがわかる．また，2 次元空間と 3 次元空間の重回帰 HMM の音素誤り率は，ほぼ同等であることもわかる．

表 3 一般の発話者における正解識別率 (%)
Table 3 Correct classification rates (%) for non-professional speakers' emotional speech.

	2 次元空間	3 次元空間
読上げ	98.61	99.17
悲嘆	89.39	97.61
楽しげ	58.50	65.33
怒り	86.28	86.28
平均	83.20	87.10

次に，2 次元空間，3 次元空間で学習した重回帰 HMM を用いて，入力音声のスタイルベクトルの推定値の分布を求めた．図 5 に，9 名の話者から選んだ 2 名の男性話者，1 名の女性話者について，求めたスタイルベクトルの分布を示す．結果から，スタイルベクトルの分布はスタイル毎に異なっていることがわかる．また，推定されたスタイルベクトルと重回帰 HMM の学習時に各スタイルに与えたスタイルベクトルとのユークリッド距離を求め，距離が最小のスタイルを識別結果として，スタイル識別を行った．表 3 に，9 名の平均の正解識別率を示す．楽しげ以外のスタイルでは，2 次元，3 次元空間ともに 85 %以上と高い識別率が得られた．楽しげの識別率が低い理由は，図 5(a), (c) のように，楽しげと怒りのスタイルの間に強い相関がある話者が存在するためだと考えられる．そのため，推定結果が怒り方向へずれてしまい，正解識別率の低下につながったと考えられる．また，2 次元空間と 3 次元空間の結果を比較すると，3 次元空間は 2 次元空間より平均で約 4 %高い識別率が得られている．これは，3 次元空間ではそれぞれのスタイルを独立の軸として学習しているため，2 次元空間に比べ，モデル学習時に他のスタイルの影響を受けにくいためであると考えられる．

3.4 重回帰 HMM と混合数を増加させた HMM との比較

最後に，通常の HMM と重回帰 HMM の認識性能の比較を行った．評価を行ったモデルは，重回帰 HMM，話者・スタイル適応 HMM（SA-HMM），1 混合不特定スタイルモデル（1-M HMM），4 混合不特定スタイルモデル（4-M HMM）の 4 つである．ここで，重回帰 HMM は，**3.3** 節で使用した 3 次元空間のモデルと同じである．SA-HMM は，**2.4** 節の **Step 1** において得られるモデルで，目標話者の各スタイル 5 文章を用いて，SI モデルから各スタイルへ適応した 4 つのモデルを用いた．1-M HMM は，目標話者の各スタイル 5 文章，計 20 文章を用いて，SI モデルから適応したモデルであり，4-M HMM は，4 つの SA-HMM の各リーフノードに存在する分布を混合して得た 1 状態当り 4 混合分布のモデルで，各分布の混合重みは全て同一である．なお，1-M HMM，4-M HMM は MAP 推定に基づくパラメータ補正を行っている．また，SA-HMM の結果は，入力音声のスタイルにマッチしたモデルを使用した場合の認識結果である．

表 4 に，各モデルの音素誤り率を示す．通常の HMM は，混合数を増加させると認識性能の向上が見られ，4-M HMM は，重回帰 HMM とほぼ同程度の性能が得られた．しかし，重回帰 HMM は 4-M HMM と比べ，モデルのパラメータ数が少なく，

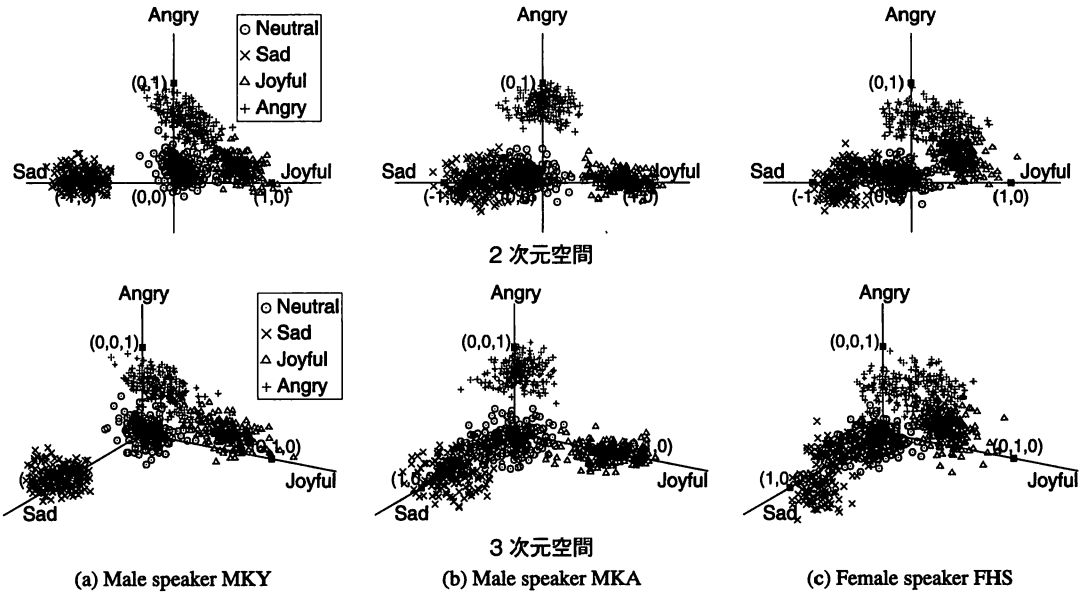


図 5　一般の発話者のスタイルベクトルの推定値の分布

Fig. 5　Distributions of estimated values of the style vector for non-professional speakers' test samples.

表 4　一般の発話者の各モデルによる音素誤り率の比較 (%)

Table 4　Comparison of phoneme error rates (%) of respective models for non-professional speakers' emotional speech.

入力	モデル			
スタイル	SA-HMM	1-M HMM	4-M HMM	重回帰 HMM
読上げ	16.11	16.09	15.98	15.75
悲嘆	20.14	20.58	19.60	19.41
楽しげ	19.34	19.75	18.48	18.67
怒り	23.30	23.41	22.41	22.37
平均	19.72	19.96	19.12	19.05

また認識結果だけでなく，パラ言語情報である入力音声のスタイルを得ることができる点で優位であると考えられる．

4. ま と め

本論文では，重回帰 HMM に基づくスタイル推定を用いた音声認識手法を，容易に任意の話者へ適用することを目的として，重回帰 HMM の学習に話者非依存モデルとモデル適応手法を利用する手法を提案し，プロのナレータと一般の発話者の模擬感情音声に対して，音素認識実験とスタイル推定の結果から，性能を評価した．その結果，提案法は話者非依存モデルより認識性能が大幅に改善され，また 3 次元空間の重回帰 HMM では，各スタイル 5 文章しか学習に使用していないにも関わらず，約 87 %のスタイル識別率が得られた．

今後は，自然発話などのより現実的な音声データに対して，提案法の有効性を検討したいと考えている．また，文章中のポーズ毎や音節毎といった現在の文章毎より短い区間でスタイル推定を行うことも今度の課題である．

謝　辞

本研究の一部は，日本学術振興会特別研究員奨励費 (1910295) の助成を受けたものである．

文　献

[1] 井島 勇祐, 橘 誠, 能勢 隆, 小林 隆夫, "スタイル推定に基づく音響モデルのオンライン適応手法," 信学技報, vol.108, no.42, SP2008-48, July 2008, pp.31–36.

[2] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, "Multiple-regression hidden Markov model," in *Proc. ICASSP 2001*, May 2001, pp. 513–516.

[3] T. Nose, Y. Kato, and T. Kobayashi, "Style estimation of speech based on multiple regression hidden semi-Markov model," in *Proc. INTERSPEECH 2007*, Oct. 2007, pp. 2285–2288.

[4] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 695–707, Nov. 2000.

[5] 宮永 圭介, 益子 貴史, 小林 隆夫, "HMM 音声合成における多様なスタイル実現のための制御法," 信学技報, vol.104, no.30, SP2004-7, April 2004, pp.35–40.

[6] T. Nose, Y. Kato, M. Tachibana, and T. Kobayashi, "An estimation technique of style expressiveness for emotional speech using model adaptation based on multiple-regression HMM," in *Proc. INTERSPEECH 2008*, Sept. 2008, pp. 2759–2762.

[7] 井澤 信介, 橘 誠, 能勢 隆, 小林 隆夫, "重回帰 HSMM に基づく合成音声のスタイル制御のための平均声からの話者適応手法," 信学技報, vol.107, no.282, SP2007-85, Oct. 2007, pp.81–86.

[8] V. Digalakis and L. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 294–300, July 1996.