論　文

HMMに基づく音声合成システムにおけるMAP-VFSを用いた声質変換

益子　貴史[†]　　田村　正統[†]　　徳田　恵一[††]　　小林　隆夫[†]

Voice Characteristics Conversion for HMM-Based Speech Synthesis System Using MAP-VFS

Takashi MASUKO[†], Masatsune TAMURA[†], Keiichi TOKUDA[††], and Takao KOBAYASHI[†]

あらまし　隠れマルコフモデル（HMM）に基づく音声合成システムにおける合成音声の声質変換手法を提案する．提案システムでは合成の基本単位として音素HMMを用いているため，HMMのパラメータを適切に変換すれば合成音声の声質を変換することができると考えられる．そこで本論文では，音声認識の分野で研究されているHMMの話者適応手法を適用することにより，合成音声の声質を少量の学習データを用いて目標話者の声質に変換する手法について検討する．話者適応手法として最大事後確率（MAP）推定法と移動ベクトル場平滑化（VFS）法を組み合わせたMAP-VFS法を適用し，不特定話者モデルから目標話者への適応を行った適応モデルを用いて音声を合成した．主観評価実験により，合成音声の声質について，目標話者による十分なデータで学習した特定話者モデルとの比較を行った．その結果，数文章程度で適応した適応モデルから合成した音声の声質は特定話者モデルを用いた場合と同等となり，合成音声の声質を容易に目標話者の声質に変換できることが示された．

キーワード　HMM，音声合成，声質変換，MAP-VFS

1. まえがき

近年提案されている音声合成手法の多くは，音素や音節などの音声単位を波形あるいはスペクトルパラメータの時系列の形でそのまま蓄積し，それらを接続して音声を合成する手法がとられており，自然性やめいりょう性の高い音声を合成できるようになってきている．しかし，一般的に音声合成システムには，合成音声の自然性，めいりょう性だけではなく，声質や話し方，感情表現など，多様なスタイルで音声を合成できることが求められる．例えば，多数話者間での自動音声翻訳システムの実現を考えた場合，翻訳された発言がどの話者によるものかを区別するために，発話者の声質で音声を合成できることが望ましい．また，仮想空間内でのコンピュータによるエージェントとの自然な対話を実現するためには，様々な声質や話し方で感情表現を伴った音声を合成できることが必要となる．

[†] 東京工業大学大学院総合理工学研究科，横浜市
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama-shi, 226–8502 Japan
[††] 名古屋工業大学知能情報システム学科，名古屋市
Faculty of Engineering, Nagoya Institute of Technology, Nagoya-shi, 466–8555 Japan

このような観点から，これまで，音声の個人性をある話者から別の話者に変換する話者変換の研究がなされている．音声の個人性はスペクトル，基本周波数，音韻継続長などに現れるため[1], [2]，個人性を変換するためにはこれらすべてを変換する必要がある．しかし，これらの中でも，特にスペクトルの変換は聴覚的に大きな影響があるため[1]，スペクトル写像による声質変換に関する研究が多くなされている[3]〜[5].

一方，音声認識の分野では，認識精度を向上させるために音響モデルを入力話者に適応させる話者適応の手法が研究されている[6]〜[12]．話者変換と話者適応は，ある話者の（または多数の話者の平均的な）スペクトルパラメータの分布を別の話者へと変換するという点で類似しており，話者適応の手法を声質変換に応用する試みもなされている[4].

ところで，我々はこれまでに隠れマルコフモデル（HMM）に基づいた音声合成システムの枠組みを提案してきた[13]．提案手法では，HMMから静的，動的特徴の分布を考慮してゆう度最大化基準に従ってスペクトルパラメータを生成しており[14]，各時刻でのスペクトルの形状とともにスペクトルが時間的にどのように変化するかも考慮することによって，滑らか

で自然性の高い合成音声が得られている．この HMM に基づく音声合成システムでは，合成単位として音素 HMM を用いているため，HMM のパラメータを適切に変換することにより，合成音声の声質を変換できると考えられる．

そこで本論文では，話者適応手法の一つである，最大事後確率（MAP）推定法 [6]〜[8] と移動ベクトル場平滑化（VFS）法 [9] を組み合わせた MAP-VFS 法 [10], [11] を適用し，少量の適応データを用いて合成音声の声質を目標となる話者へ変換する手法について述べる．我々は既に文献 [15] で，MAP-VFS 法により特定話者モデルを目標話者へ適応することにより，合成音声の声質を目標話者へ変換できることを示している．しかし特定話者モデルからの変換においては，初期話者の選び方によって適応精度が異なると考えられる．そこで本論文では，不特定話者モデルから変換する場合について検討する．

2. HMM に基づく音声合成システム

HMM に基づく音声合成システムのブロック図を図 1 に示す．システムは学習部，適応部，合成部の三つの部分に分けられる．

学習部では，まず音声データベースからメルケプストラム分析 [16] によりメルケプストラムを求める．更に動的特徴量である Δ，Δ^2 パラメータを計算し，静的特徴量と合わせて特徴ベクトルとし，音素 HMM を学習する．音素 HMM は前後の音韻環境を考慮した triphone モデルとし，決定木に基づくコンテキストクラスタリング [17] により，HMM の各状態を中心音素別，モデル内での状態位置別にクラスタリングして共有化する．HMM の学習後，学習データに対する Viterbi アラインメントにより HMM の各状態の継続長のヒストグラムを求め，これをガウス分布で近似して各状態の状態継続長分布とする．

適応部では，学習した HMM を話者適応手法を用いて目標となる話者のモデルに変換する．話者適応には，音声認識の分野で研究されている様々な手法を適用することができる．

合成部では，まず合成したい任意のテキストを音素列に変換する．この音素列に従って前述の音素 HMM を接続し，与えられたテキストに対応する一つの文 HMM をつくる．この文 HMM から，ゆう度最大化基準に基づくパラメータ生成アルゴリズム [14] によりメルケプストラム系列を生成する．これに適当なピッチを与え，MLSA（メル対数スペクトル近似）フィルタ [18], [19] を用いて合成音声を得る．

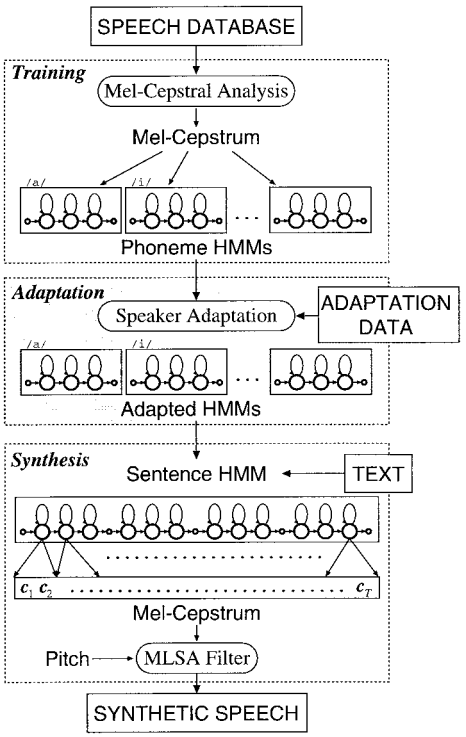


図 1　HMM に基づく音声合成システム
Fig. 1　A block diagram of an HMM-based speech synthesis system.

2.1 ゆう度最大基準に基づく HMM からのパラメータ生成

連続出力分布 HMM のパラメータセットを λ で表し，長さ T の状態遷移系列を $\boldsymbol{Q} = (q_1, q_2, \ldots, q_T)$，各時刻 t $(1 \leqq t \leqq T)$ の出力ベクトル $\boldsymbol{o}_t$ を縦に並べることにより構成されるベクトルを $\boldsymbol{O} = [\boldsymbol{o}_1', \boldsymbol{o}_2', \ldots, \boldsymbol{o}_T']'$ とする．ここでは簡単のため，HMM は単一ガウス出力分布型 left-to-right モデルであると仮定する．

与えられた HMM のパラメータセット λ 及び状態遷移系列 $\boldsymbol{Q}$ に対し，出力確率 $P(\boldsymbol{O}\,|\,\boldsymbol{Q}, \lambda)$ を最大化する出力ベクトル $\boldsymbol{O}$ を求めることを考える [14]．出力確率 $P(\boldsymbol{O}\,|\,\boldsymbol{Q}, \lambda)$ の対数は，

$$\begin{aligned}&\log P(\boldsymbol{O}\,|\,\boldsymbol{Q}, \lambda)\\ &\quad = -\frac{1}{2}(\boldsymbol{O}-\boldsymbol{\mu})'\boldsymbol{U}^{-1}(\boldsymbol{O}-\boldsymbol{\mu}) - \frac{1}{2}\log|\boldsymbol{U}|\\ &\qquad - Const. \end{aligned} \tag{1}$$

となる．ここで，

$$\boldsymbol{\mu} = [\boldsymbol{\mu}'_{q_1}, \boldsymbol{\mu}'_{q_2}, \ldots, \boldsymbol{\mu}'_{q_T}]' \tag{2}$$

$$\boldsymbol{U} = \mathrm{diag}[\boldsymbol{U}_{q_1}, \boldsymbol{U}_{q_2}, \ldots, \boldsymbol{U}_{q_T}] \tag{3}$$

であり，$\boldsymbol{\mu}_{q_t}$ 及び $\boldsymbol{U}_{q_t}$ はそれぞれ状態 q_t の平均及び共分散である．また，*Const.* は出力分布の正規化係数の対数であり，$\boldsymbol{O}$，$\boldsymbol{\mu}$，$\boldsymbol{U}$ と独立な定数項である．

$\boldsymbol{Q}$ が与えられたとき，$P(\boldsymbol{O}\,|\,\boldsymbol{Q}, \lambda)$ を最大化する $\boldsymbol{O}$ は，静的特徴量 $\boldsymbol{c}_t = [c_t(1), c_t(2), \ldots, c_t(M)]'$（例えばメルケプストラム係数，ただし M は次数）のみを考慮する場合 ($\boldsymbol{o}_t = \boldsymbol{c}_t$)，$\boldsymbol{O} = \boldsymbol{\mu}$ となる．すなわち，個々のフレームにおける出力は前後のフレームでの出力とは独立に決定され，そのフレームに対応する状態の出力分布の平均となるため，ある状態から次の状態へ遷移する部分で階段状に変化し，不連続となる．

そこで，出力パラメータをフレームごとに独立な静的特徴量 $\boldsymbol{c}_t$ と，前後の複数フレームの静的特徴量から定まる動的特徴量 $\Delta\boldsymbol{c}_t$，$\Delta^2\boldsymbol{c}_t$ を導入して $\boldsymbol{o}_t = [\boldsymbol{c}'_t, \Delta\boldsymbol{c}'_t, \Delta^2\boldsymbol{c}'_t]'$ とすることを考える．このとき，$P(\boldsymbol{O}\,|\,\boldsymbol{Q}, \lambda)$ を最大化する静的特徴量 $\boldsymbol{c}_t$ $(1 \leqq t \leqq T)$ からなるベクトル $\boldsymbol{C} = [\boldsymbol{c}'_1, \boldsymbol{c}'_2, \ldots, \boldsymbol{c}'_T]'$ は，$\boldsymbol{0}_{TM}$ を $T \times M$ 次の零ベクトルとして，

$$\frac{\partial}{\partial \boldsymbol{C}} \log P(\boldsymbol{O}\,|\,\boldsymbol{Q}, \lambda, T) = \boldsymbol{0}_{TM} \tag{4}$$

として与えられる線形連立方程式を解くことにより一意に求められる [14]．これにより，スペクトルの形状とともに時間的な変化も考慮した滑らかで自然性の高いパラメータ系列が得られる．

3. MAP-VFS を用いた声質変換

2. で説明した HMM に基づく音声合成システムでは，合成単位として音素 HMM を用いている．この音素 HMM は音韻性だけではなく話者性も同時にモデル化しており，学習により得られたモデルパラメータのうち話者性を表す部分を適切に変換することができれば，合成音声の声質を変換することができると考えられる．ここでモデルパラメータの変換手法としては，話者適応による手法 [15], [20] や複数の話者のモデルから補間により新たなモデルを作成する手法 [21] などが考えられるが，本論文では最大事後確率（MAP）推定法 [6]〜[8] と移動ベクトル場平滑化（VFS）法 [9] を組み合わせた MAP-VFS 法 [10], [11] による話者適応を用いる．

MAP-VFS では，まず目標話者による少量の適応データを用いて最大事後確率（MAP）推定を行い，その後移動ベクトル場平滑化（VFS）法により局所的連続性の仮定のもとで，適応データの存在しない分布に対しては補間を，また適応データの存在する分布に対しても平滑化を行うことにより，すべてのモデルの適応を行う．

3.1 最大事後確率（MAP）推定法

与えられた学習データ $\boldsymbol{x}$ に対し，推定すべきパラメータを λ とすると，λ の最大事後確率（MAP）推定値 λ^{MAP} は，

$$\begin{aligned}\lambda^{MAP} &= \operatorname*{argmax}_{\lambda} g(\lambda\,|\,\boldsymbol{x}) \\ &= \operatorname*{argmax}_{\lambda} f(\boldsymbol{x}\,|\,\lambda) g(\lambda)\end{aligned} \tag{5}$$

ただし，$f(\cdot)$，$g(\cdot)$ はそれぞれ確率変数 $\boldsymbol{x}$，λ の確率密度関数である．事前分布 $g(\lambda)$ が λ とは関係なく定数である場合には，MAP 推定は最ゆう（ML）推定となる．

$\boldsymbol{x} = \{x_1, \ldots, x_n\}$ を平均 μ，分散 σ^2 のガウス分布に従う独立な観測サンプル系列とする．分散 σ^2 が既知であると仮定すると，平均 μ の共役事前分布は平均 μ_0，共分散 σ_0^2 のガウス分布で表される [22]．μ_0 は μ の事前知識のみによる最も確からしい推定値であり，σ_0 は μ_0 の不確かさを表している．このとき，μ の MAP 推定値 μ^{MAP} は次式で表される．

$$\mu^{MAP} = \frac{n}{n+\tau}\bar{x} + \frac{\tau}{n+\tau}\mu_0 \tag{6}$$

ただし，n はサンプル数であり，$\bar{x}$ はサンプル平均 $\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$ である．また，τ は $\tau = \sigma^2/\sigma_0^2$ であり，事前分布の確からしさを表す係数であるが，本論文では文献 [7] と同様に一定値とした．

多変量正規分布の場合にも式 (6) と同様にして MAP 推定値が得られる．また連続分布 HMM については出力分布の平均ベクトルだけではなく，共分散行列や状態遷移確率など MAP 推定が可能である [8] が，本論文では平均ベクトルのみの推定を行った．

3.2 移動ベクトル場平滑化法

一般的に，話者適応では適応データが非常に少ないため，学習サンプルが得られずに MAP 推定をできない状態が存在する．そこで，学習サンプルの得られないパラメータに対する適応後の値を補間により推定する．

ある分布 q の近傍の MAP 推定値の得られた K 個の分布の集合を $G_K(q)$ で表す．HMM の出力分布の集合のうち，MAP 推定値の得られた分布 i に対する移動ベクトル $\boldsymbol{v}_i$ は，次式で表される．

$$\boldsymbol{v}_i = \boldsymbol{\mu}_i^{MAP} - \boldsymbol{\mu}_i \tag{7}$$

ただし，$\boldsymbol{\mu}_i^{MAP}$，$\boldsymbol{\mu}_i$ はそれぞれ分布 i の平均ベクトルの MAP 推定値及び初期値である．このとき，MAP 推定値の得られなかった分布 j の VFS により補間された移動ベクトル $\boldsymbol{v}_j^I$ は，以下の式により求められる．

$$\boldsymbol{v}_j^I = \frac{\sum_{k \in G_K(j)} w_{jk} \boldsymbol{v}_k}{\sum_{k \in G_K(j)} w_{jk}} \tag{8}$$

ただし，w_{jk} は $\boldsymbol{\mu}_j$ と $\boldsymbol{\mu}_k$ の距離によって決まる重み係数である．この補間された移動ベクトルを用いて，$\boldsymbol{\mu}_j$ の推定値 $\boldsymbol{\mu}_j^I$ は

$$\boldsymbol{\mu}_j^I = \boldsymbol{\mu}_j + \boldsymbol{v}_j^I \tag{9}$$

と求められる．

また，移動ベクトルの連続性の仮定に基づき，すべての学習された移動ベクトル $\boldsymbol{v}_i$ について次式により平滑化を行う．

$$\boldsymbol{v}_i^S = \frac{\boldsymbol{v}_i + \sum_{k \in G_K(i)} w_{ik} \boldsymbol{v}_k}{1 + \sum_{k \in G_K(i)} w_{ik}} \tag{10}$$

この平滑化された移動ベクトルを用い，ガウス分布の平均値を以下のように修正する．

$$\boldsymbol{\mu}_i^S = \boldsymbol{\mu}_i + \boldsymbol{v}_i^S \tag{11}$$

本論文では重み係数は次式により計算した．

$$w_{jk} = \exp(-d_{jk}/s) \tag{12}$$

ただし，d_{jk} は $\boldsymbol{\mu}_j$ と $\boldsymbol{\mu}_k$ の間のマハラノビス距離であり，s は平滑化の度合を制御する係数（s が大きくなるに従って平滑化が強まる）である．

4. 評価実験

4.1 実験条件

ABX 法による受聴試験により，不特定話者モデルから目標話者に適応した適応モデルからの合成音声の適応の度合を評価した．

音声データは ATR 日本語音声データベースを用いた．データベースに含まれるラベルデータに基づいて，35 種類の音素及び無音でラベル付けした．男女各 10 名による 3,000 文章（各話者 150 文章）を用いて性別に依存しない不特定話者モデルを学習し，学習データの話者 20 名に含まれない女性話者 2 名（FKN, FYM）及び男性話者 2 名（MYI, MHT）の計 4 名を目標話者として適応を行った．また，目標話者 4 名による音韻バランス文（各話者 450 文章）を用い，各目標話者それぞれに対して特定話者モデルを学習した．

サンプリング周波数は 10 kHz とし，フレーム長 25.6 ms，フレーム周期 5 ms のブラックマン窓を用い，メルケプストラム分析により 0 次から 15 次までのメルケプストラムを求め，更に式 (13)，(14) により Δ，Δ^2 メルケプストラムを計算し，これらを特徴ベクトルとした．

$$\Delta \boldsymbol{c}_t = \frac{\boldsymbol{c}_{t+1} - \boldsymbol{c}_{t-1}}{2} \tag{13}$$

$$\Delta^2 \boldsymbol{c}_t = \frac{\Delta \boldsymbol{c}_{t+1} - \Delta \boldsymbol{c}_{t-1}}{2} \tag{14}$$

HMM は 5 状態 left-to-right モデルで，それぞれの状態の出力分布は単一の対角共分散ガウス分布とした．決定木に基づくコンテクストクラスタリングにより，triphone HMM の各状態を中心音素別，状態別にクラスタリングして共有化し，tied triphone HMM のセットを生成した．不特定話者モデル，特定話者モデルにかかわらず，クラスタリングの終了条件は各モデルセットで同じとした．

適応データは学習データとは異なる 12 文章とした．適応データに出現する triphone の種類及び学習サンプルの得られた分布数は，目標話者によって多少異なるが，目標話者を FKN とし，適応文章数を 1，3，5，8，10，12 文章とした場合，triphone の種類はそれぞれ 103，182，244，372，450，507，また学習サンプルの得られた分布数はそれぞれ 413，722，956，1,407，1,668，1,837 となった．ただし，不特定話者モデルの総分布数は 4,620 分布である．MAP 推定を行う際，文献 [15] では Segmental MAP 推定法 [8] を用いているが，本実験では予備実験においてより良い結果の得られた Forward-Backward MAP 推定法 [8] を用いた．

評価用データは学習データ及び適応データとは異なる 53 文章とし，4 文章を主観評価実験に，残りの 49

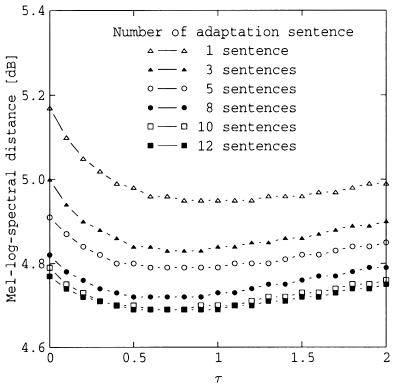


図 2　メル対数スペクトル距離に対する τ の影響
Fig. 2　Mel-log-spectral distance depending on different value of τ.

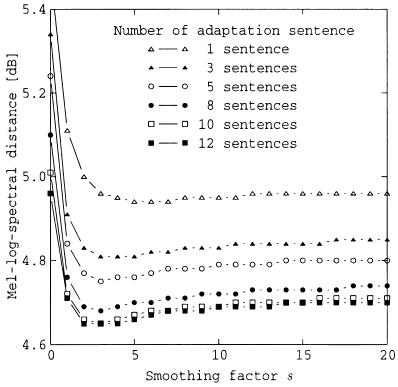


図 3　メル対数スペクトル距離に対する s の影響
Fig. 3　Mel-log-spectral distance depending on different value of s.

文章を MAP-VFS のパラメータを決定するための予備実験に用いた．

なお，本実験では声質のみの比較を行うため，合成時の各状態の継続長は目標話者による自然発声から分析して得られたパラメータ列に対して文 HMM を Viterbi アラインメントして得られた値を用いた．またピッチパターンについても，データベースに付属するピッチデータをそのまま用いた．

4.2 予備実験

MAP 推定におけるパラメータ τ 及び VFS における平滑化係数 s を決定するため，評価用データのうち主観評価実験で用いる 4 文章を除く 49 文章を用いて，HMM から生成されたスペクトルと自然発声から分析して得られたスペクトルとの 0 次を除いた平均メル対数スペクトル距離を求めた．ただし，VFS における近傍数は $K = 10$ とした．

まず，図 2 に $s = 10$ として τ を 0 から 2 まで 0.1 ごとに変化させた場合の目標話者 4 名の平均スペクトル距離を適応文章数別に示す．縦軸は平均メル対数スペクトル距離，横軸は τ の値を示しており，グラフは上から順に適応文章数が 1，3，5，8，10，12 の場合を示している．ここで，$\tau = 0$ の場合は，MAP 推定は最ゆう（ML）推定と等しくなり，文献 [9] で示されている VFS 法に相当する．

図 2 より，ML 推定に基づく VFS 法は MAP-VFS 法と比べてスペクトル距離が大きく，$\tau = 0.8$ 付近でスペクトル距離が最小となっていることがわかる．

次に，図 3 に $\tau = 0.8$ として s を 0 から 20 まで 1 ごとに変化させた場合のスペクトル距離を示す．ここで，$s = 0$ の場合は，VFS を行わず MAP 推定のみの場合に相当する．

図 3 より，$s = 0$ の場合は他と比べてスペクトル距離が非常に大きくなっており，VFS を用いた場合の方が距離が小さくなることがわかる．また，$s = 3$ 付近でスペクトル距離が最小となっており，s が 5 より大きい部分ではほぼ一定となっていることがわかる．

以上の結果より，主観評価実験における MAP-VFS 法のパラメータは，$\tau = 0.8$，$s = 3$ とした．話者によりパラメータの最適値には多少のばらつきがあるものの，$s = 3$，$\tau = 0.8$ 付近でパラメータを変化させても合成音声の主観的な品質，話者性はほとんど変化せず，話者ごとに最適な値を用いた場合と $s = 3$，$\tau = 0.8$ とした場合で合成音声の話者性はほとんど変化しないことから，以下の主観評価実験では全話者共通のパラメータを用いることとした．

なお，ここには示さないが，$s = 3$ として τ を変化させた場合にも，$\tau = 0.8$ 付近でスペクトル距離が最小となることを確認している．また，s や τ を大きくしすぎると不特定話者モデルからの声質の変化が小さくなることを確認している．

4.3 主観評価

主観評価実験は ABX 法により行った．A を不特定話者モデルから合成された音声，B を特定話者モデルから合成された音声，X を適応モデルから合成された音声とし，被験者に (A，B，X) または (B，A，X) の順に提示し，X の声質が A と B のどちらに近いかを強制判定させた．被験者は男性 8 名である．

目標話者 4 名について平均した結果を図 4 に示す．

縦軸は適応モデルから合成された音声が不特定話者モデルから合成された音声よりも特定話者モデルから合成された音声に近いと判断された割合，横軸の数字は適応データとして用いた文章数を表す．

図 4 より，適応データが 1 文章のみの場合でも 88.7% のスコアが得られており，適応モデルからの合成音声が目標話者に十分近いことがわかる．また，8 文章程度で適応がほぼ収束していることもわかる．

次に，図 5 に適応文章数が 1 文章（白い棒グラフ）及び 8 文章（グレーの棒グラフ）の場合の話者別の評価結果を示す．図 5 より，話者によって多少のばらつきはあるものの，どの話者の場合にも 1 文章で 80% 以上，8 文章で 90% 以上のスコアが得られ，ほぼ安定して適応できていることがわかる．なお，話者 FYM では適応文章数が 1 文章の場合よりも 8 文章の場合の

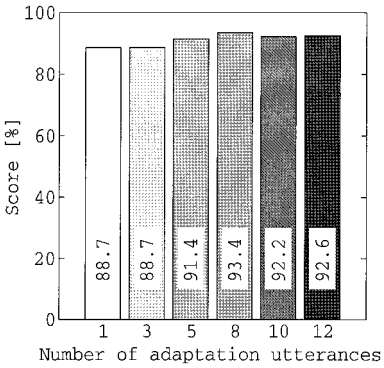


図 4　主観評価結果
Fig. 4　Results of ABX listening tests.

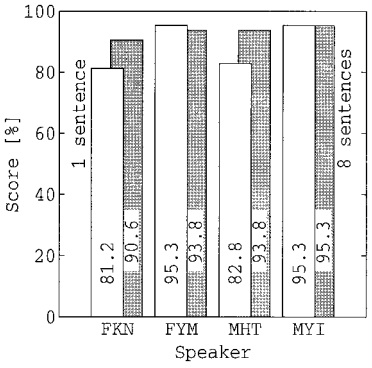
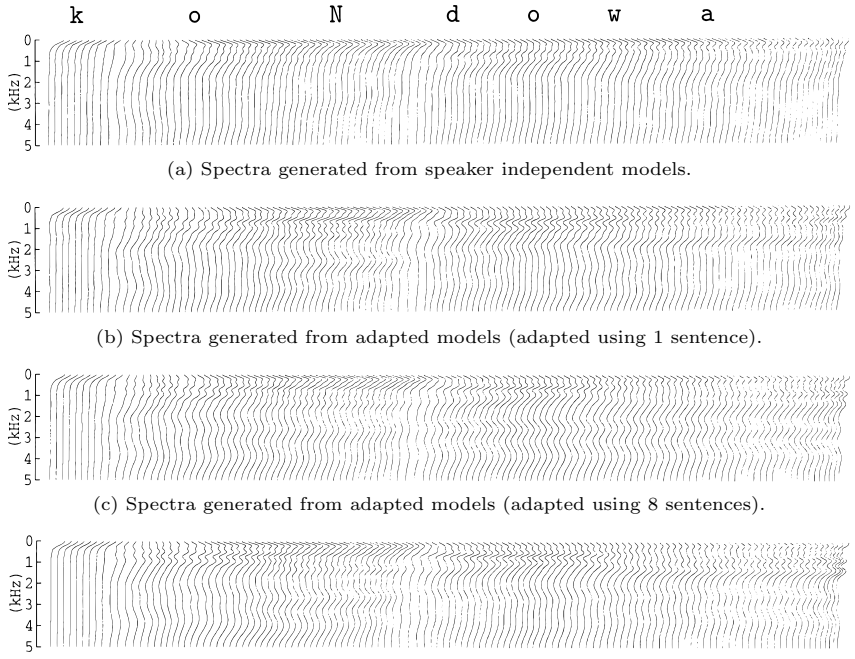


図 5　話者別の主観評価結果
Fig. 5　Results of ABX listening tests for each speaker.



(a) Spectra generated from speaker independent models.

(b) Spectra generated from adapted models (adapted using 1 sentence).

(c) Spectra generated from adapted models (adapted using 8 sentences).

(d) Spectra generated from speaker dependent models.

図 6　HMM から生成されたスペクトルの例（「こんどは」）
Fig. 6　Spectra generated from HMMs (/k-o-N-d-o-w-a/).

方がスコアがわずかに低くなっているが，両者の声質にはほとんど差がなく，誤差の範囲内であると考えられる．

図 6 に，不特定話者モデルを話者 MHT へ適応させた場合の HMM から生成されたスペクトルの例を示す．図 6 (a) は初期モデルから生成されたスペクトル，(b)，(c) はそれぞれ 1，8 文章を用いて適応した適応モデルから生成されたスペクトル，(d) は特定話者モデルから生成されたスペクトルである．図より，不特定話者モデルではスペクトルの山や谷がはっきりせず平たんになっているのに対し，適応を行うことによって目標話者のスペクトルに近づき，8 文章程度で特定話者モデルとほぼ同等のスペクトルが得られていることがわかる．

本実験で用いた MAP-VFS のパラメータの値は，実音声とのスペクトル距離に基づいて求めているが，スペクトル距離と聴感上の距離とは必ずしも一致しないため，最適な値ではない可能性もある．しかし，わずか 1 文章で適応できること，また異なる話者に対しても安定して適応できることから，妥当な値であると考えられる．

5. むすび

話者適応手法の一つである MAP-VFS 法を用いた多様な声質による音声合成システムを提案し，数文章程度で合成音声の声質を変換できることを示した．

今後の課題としては，逐次的に MAP-VFS を行う手法 [11] との比較や最ゆう線形回帰（MLLR）[12] などの他の話者適応手法を用いた場合 [20] との比較，更にピッチモデル，状態継続長モデル [23]〜[25] に対する適応手法の検討などが挙げられる．また，話者性だけではなく，自然性やめいりょう性などの観点から合成音声の品質を評価することも今後の課題となる．

謝辞 本研究を進めるにあたり，有益な御助言を頂いた千葉工業大学 今井聖教授に感謝致します．本研究の一部は文部省科学研究費補助金（課題番号 08750506, 10555125）によった．

文　献

[1] 伊藤憲三，斉藤収三，“音声の音響的特徴パラメータが個人性の知覚に及ぼす影響，” 信学論（A），vol.J65-A, no.1, pp.101–108, Jan. 1982.

[2] N. Higuchi and M. Hashimoto, “Analysis of acoustic features affecting speaker identification,” Proc. EUROSPEECH95, pp.435–438, Sept. 1995.

[3] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Voice conversion through vector quantization,” Proc. ICASSP88, pp.655–658, April 1988.

[4] 橋本　誠，樋口宜男，“話者選択と移動ベクトル場平滑化による声質変換のためのスペクトル写像，” 信学論（D-II），vol.J80-D-II, no.1, pp.1–9, Jan. 1997.

[5] Y. Stylianou and O. Cappé, “A system for voice conversion based on probabilistic classification and a harmonic plus noise model,” Proc. ICASSP98, pp.281–284, May 1998.

[6] C.H. Lee, C.H. Lin, and B.H. Juang, “A study on speaker adaptation of the parameters of continuous density hidden Markov models,” IEEE Trans. Signal Processing, vol.39, no.4, pp.806–814, 1991.

[7] C.H. Lee and J.L. Gauvain, “Speaker adaptation based on MAP estimation of HMM parameters,” Proc. ICASSP93, pp.558–561, April 1993.

[8] J.L. Gauvain and C.H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” IEEE Trans. Speech & Audio Processing, vol.2, no.2, pp.291–298, 1994.

[9] 大倉計美，杉山雅英，嵯峨山茂樹，“混合連続分布 HMM を用いた移動ベクトル場平滑化話者適応方式，” 信学技報，SP96-12, June 1992.

[10] M. Tonomura, T. Kosaka, and S. Matsunaga, “Speaker adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation,” Computer Speech and Language, vol.10, pp.117–132, 1996.

[11] J. Takahashi and S. Sagayama, “Vector-field-smoothed Bayesian learning for fast and incremental speaker/telephone-channel adaptation,” Computer Speech and Language, vol.11, pp.127–146, 1997.

[12] C.J. Leggetter and P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” Computer Speech and Language, vol.9, pp.171–185, 1995.

[13] 益子貴史，徳田恵一，小林隆夫，今井　聖，“動的特徴を用いた HMM に基づく音声合成，” 信学論（D-II），vol.J79-D-II, no.12, pp.2184–2190, Dec. 1996.

[14] 徳田恵一，益子貴史，小林隆夫，今井　聖，“動的特徴を用いた HMM からの音声パラメータ生成アルゴリズム，” 音響誌，vol.53, no.3, pp.192–200, March 1997.

[15] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, “Voice characteristics conversion for HMM-based speech synthesis system,” Proc. ICASSP97, pp.1611–1614, 1997.

[16] 徳田恵一，小林隆夫，深田俊明，斎藤博徳，今井　聖，“メルケプストラムをパラメータとする音声のスペクトル推定，” 信学論（A），vol.J74-A, no.8, pp.1240–1248, Aug. 1991.

[17] S.J. Young, J.J. Odell, and P.C. Woodland, “Tree-based state tying for high accuracy acoustic modelling,” Proc. ARPA Workshop on Human Language Technology, pp.307–312, 1994.

[18] 今井　聖，住田一男，古市千枝子，“音声合成のためのメ

ル対数スペクトル近似（MLSA）フィルタ," 信学論（A）, vol.J66-A, no.2, pp.122–129, Feb. 1983.
[19] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," Proc. ICASSP92, pp.137–140, March 1992.
[20] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," Proc. The Third ESCA/COSCOSDA International Workshop on Speech Synthesis, pp.273–276, Nov. 1998.
[21] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," Proc. EUROSPEECH97, pp.2523–2526, Sept. 1997.
[22] M.H. DeGroot, Optimal Statistical Decisions, McGraw-Hill, 1970.
[23] 宮崎　昇，徳田恵一，益子貴史，小林隆夫，"多空間上の確率分布に基づいた HMM とピッチパタンモデリングへの応用," 信学技報，SP98-11, April 1998.
[24] 宮崎　昇，徳田恵一，益子貴史，小林隆夫，"多空間上の確率分布に基づいた HMM によるピッチパタン生成の検討," 信学技報，SP98-12, April 1998.
[25] 吉村貴克，徳田恵一，益子貴史，小林隆夫，北村　正，"HMM に基づく音声合成のための状態継続長モデルの構築," 信学技報，DSP98-85, SP98-64, Sept. 1998.

（平成 11 年 11 月 9 日受付，12 年 3 月 9 日再受付）

益子　貴史　（正員）

平 5 東工大・工・情工卒．平 7 同大大学院博士前期課程了（知能科学専攻）．同年東工大精密工学研究所助手．平 10 東工大大学院総合理工学研究科物理情報工学専攻助手．音声分析・合成・認識，マルチモーダルインタフェースの研究に従事．日本音響学会，IEEE，ISCA 各会員．

田村　正統

平 9 東工大・工・情工卒．現在同大大学院総合理工学研究科知能システム科学専攻博士前期課程在学中．音声合成，マルチモーダルインタフェースの研究に従事．日本音響学会，ISCA 各会員．

徳田　恵一　（正員）

昭 59 名工大・工・電子卒．平 1 東工大大学院博士課程了．同年東工大電気電子工学科助手．平 8 名工大知能情報システム学科助教授．工博．音声分析・合成・符号化・認識，ディジタル信号処理，マルチモーダルインタフェースの研究に従事．日本音響学会，情報処理学会，人工知能学会，IEEE 各会員．

小林　隆夫　（正員）

昭 52 東工大・工・電気卒．昭 57 同大大学院博士課程了．同年東工大精密工学研究所助手．同助教授を経て平 10 東工大大学院総合理工学研究科物理情報工学専攻教授．工博．ディジタルフィルタ，音声分析・合成・符号化・認識，マルチモーダルインタフェースの研究に従事．日本音響学会，IEEE，ISCA 各会員．