論　文

HMMに基づく音声合成におけるピッチ・スペクトルの話者適応

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

Speaker Adaptation of Pitch and Spectrum for HMM-Based Speech Synthesis

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

あらまし　本論文では，不特定話者の音声合成単位である“平均声”モデルから，任意話者の特徴をもつ音声を合成する手法を提案する．提案手法は，HMMに基づくテキスト音声合成システムに基づいている．HMMに基づく音声合成システムでは，多空間上の確率分布（MSD）に基づくHMMを用いてスペクトル及びピッチパラメータを同時にモデル化しており，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や韻律特徴を変換できる．本論文では，MLLRアルゴリズムをMSD-HMMに拡張し，ピッチ及びスペクトルモデルの話者適応を行うことにより，目標話者の少量の文章を用いて，声質のみでなく韻律情報も適応できることを示す．主観評価試験により，ピッチ及びスペクトルを同時に話者適応することにより，平均声モデルを数文章で適応したモデルから，特定話者モデルからの合成音声に近い音声を合成できることを示した．

キーワード　音声合成，話者適応，声質変換，MLLR，平均声

1. まえがき

テキスト音声合成は自然なヒューマンコンピュータインタラクションの実現のための重要な要素技術であり，音声合成システムには，合成音声の自然性・明りょう性とともに，話者性や感情表現など，多様なスタイルで音声を合成できることが求められる．近年広く用いられている波形素片接続型の音声合成手法（例えば[1], [2]）では，自然性の高い合成音声が生成できる反面，ある話者の声質・韻律特徴をもつ音声を合成するためには，その話者が発声した大量の音声データを必要とする．このことは，任意の話者の多様な声質・韻律特徴をもつ音声を合成するためには，対応する任意の話者の大量のデータをあらかじめ用意しておかなければならないことを意味しており，現実的には困難であるといえる．更に，実際には存在しない話者の声質をもつ音声を合成することも非常に困難である．

一方我々の提案するHMMに基づく音声合成システム[3], [4]は，Vocoder型であるため波形素片接続型と比較して多少音質は劣るものの，滑らかで明りょうな合成音声が得られている．また，合成の基本単位として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で，合成音声の声質を変換することができるという特徴をもっている．実際，MLLR [5]や，MAP/VFS [6], [7]などの話者適応技術を適用することにより，任意の話者の声質に近い音声を合成できること[8], [9]，また話者補間[10]や固有声[11]などの手法により，多様な声質の音声を合成できることを示している．

しかし，これまでのHMMに基づく音声合成システムにおける合成音声の多様化の検討においては，スペクトルパラメータの変換のみを扱っており，声質（スペクトル）とともに話者性に関して重要な要素である韻律については考慮していなかった．そこで本論文では，声質の話者適応のみでなく，韻律特徴の一つであるピッチパラメータを話者適応し，任意話者の声質・韻律特徴に近い音声を合成する手法を提案する．HMMに基づく音声合成システム[4]では，ピッチパターンを多空間上の確率分布（MSD-HMM）によりモデル化している[12], [13]．そこで，MLLRアルゴリズムをMSD-HMMに拡張し，ピッチパラメータの話者適応に用いる．

[†] 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

話者適応の際，初期モデルを学習した話者と目標となる話者の組合せによって適応の精度が変化すると考えられる．本論文ではこの影響を低減するため，複数の話者により発声された音声データを用いて学習した不特定話者モデルを初期モデルとして用いることにする．この不特定話者モデルは，学習に用いた複数の話者の平均的なスペクトル，ピッチ，継続長をモデル化しており，不特定話者モデルから合成された音声は複数の話者の平均的な特徴をもつと考えられることから，ここではこの不特定話者モデルを「平均声モデル」，平均声モデルから合成された音声を「平均声」と呼ぶことにする．以下では，初期モデルとして数名の話者のデータから得られた平均声モデルを用い，これを拡張した MLLR により目標とする話者に適応し，得られたモデルから音声を合成する手法を提案する．評価実験を通して，声質及び韻律特徴を変換することで目標とする話者に近い音声を合成することができることを示す．

2. HMM に基づく音声合成システム

2.1 音声合成システム

提案する HMM に基づく音声合成システムの構成を図 1 に示す．システムは大きく分けて学習部，適応部，合成部からなる．

学習部では，複数話者音声データを用いて，話者適応の初期モデルとなる平均声音素 HMM を作成する．まず，複数の話者の音声データベースから，スペクトルパラメータとしてメルケプストラム，ピッチパラメータとして対数基本周波数を求め，静的特徴量とする．得られたパラメータから動的特徴量を求め，静的特徴量と結合して，特徴パラメータとする．ピッチパラメータは，有声音では 1 次元の連続値となるが無声音では値をもたないため，HMM の出力分布として通常用いられる連続分布若しくは離散分布でモデル化することはできない．そこで，ピッチの値を表す 1 次元の連続値と無声を表す離散シンボルを統一的にモデル化するために，多空間上の確率分布に基づいた HMM（MSD-HMM）[12] を用いる．スペクトル部には通常の連続分布を，ピッチ部には多空間上の確率分布を用いたマルチストリーム MSD-HMM によりスペクトルとピッチを同時にモデル化する [4].

学習の際，スペクトル，ピッチ，継続長は音韻のみでなく，呼気段落，形態素，アクセントなどの言語的な情報も考慮してモデル化し，MDL 基準による決定木に基づくコンテクストクラスタリング [14] をスペクトル部，ピッチ部それぞれに適用することで，出力分布を共有化する．音韻継続長は，HMM の各モデルの状態継続長を多次元のガウス分布でそれぞれモデル化しスペクトル・ピッチと同様に決定木に基づくコンテクストクラスタリングを適用する [15].

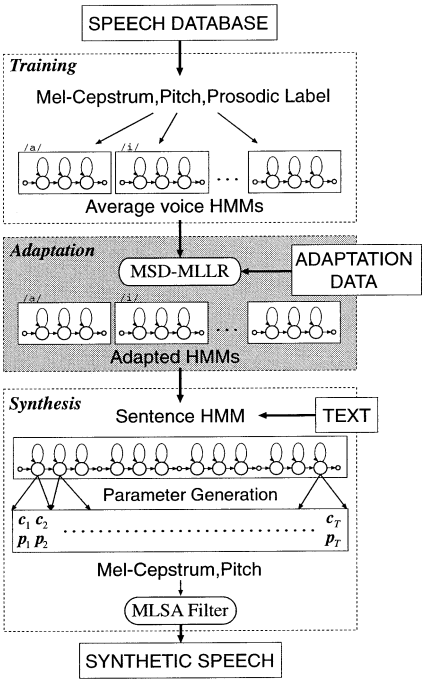


図 1　音声合成システム
Fig. 1　HMM-based speech synthesis system.

適応部では，学習部で得られた平均声 HMM のパラメータを目標とする話者の少量の音声データを用いて話者適応する．本論文では，ピッチパラメータとスペクトルパラメータに対し **3.** で述べる MSD-HMM に基づく MLLR を適用する．これにより，スペクトル及びピッチパラメータを同時に適応することができる．

合成部では，まず，合成する文章を韻律情報を含むコンテクスト依存音素ラベル列に変換する．得られたラベル列に従って音素 HMM を接続し，文章 HMM とする．音素モデルの継続長分布の平均値から HMM のそれぞれの状態の継続長を決定し，メルケプストラム及びピッチパターンをゆう度最大化基準に基づく HMM からのパラメータ生成アルゴリズム [13], [16] を用いて生成し，MLSA フィルタ [17] を用いて音声を合成する．

2.2 MSD-HMM によるピッチのモデル化

ピッチは無声区間では値をもたず，有声区間では連続な値をもつ．そこで，ピッチパターンを有声区間では 1 次元空間 Ω_1，無声区間では 0 次元空間 Ω_2 から出力される系列であるとする．各空間 Ω_g は重み w_g $(\sum_{g=1}^{2} w_g = 1)$ をもち，Ω_1 においては 1 次元確率密度関数 $\mathcal{N}_1(\boldsymbol{x})$ $(\int \mathcal{N}_1(\boldsymbol{x})d\boldsymbol{x} = 1)$ が定義され，Ω_2 は 1 点のみを示す．

観測ベクトル $\boldsymbol{o}$ は空間インデックス集合 X と変数 $\boldsymbol{x}$ からなる．

$$\boldsymbol{o} = (X, \boldsymbol{x}) \tag{1}$$

ここで，有声区間では $X = \{1\}$，無声区間では $X = \{2\}$ となる．このとき，$\boldsymbol{o}$ の観測確率は次式で定義される．

$$b(\boldsymbol{o}) = \sum_{g \in S(\boldsymbol{o})} w_g \mathcal{N}_g(V(\boldsymbol{o})) \tag{2}$$

ただし，

$$V(\boldsymbol{o}) = \boldsymbol{x}, \quad S(\boldsymbol{o}) = X \tag{3}$$

である．ここで，空間 Ω_2 上では $\mathcal{N}_2(\boldsymbol{x})$ は存在しないが，簡単のため $\mathcal{N}_2(\boldsymbol{x}) = 1$ と定義する．

各状態の出力分布が式 (2) で表される HMM を MSD-HMM と呼び，MSD-HMM を用いることにより，ピッチパターンを統一的にモデル化することができる [12]．

図 2 に本論文で用いる観測ベクトルの構成を示す．図中 $\boldsymbol{c}_t$，X_t^p，$\boldsymbol{x}_t^p$ はそれぞれ時刻 t におけるスペクトルパラメータ，ピッチの有声，無声を表す空間，ピッチパラメータを示し，Δ，Δ^2 はデルタ及びデルタデルタパラメータを示している．学習時には，スペクトル，ピッチ，Δ ピッチ，Δ^2 ピッチを別々のストリームに分割し，スペクトル部は連続出力分布，ピッチ部は MSD により同時にモデル化する．

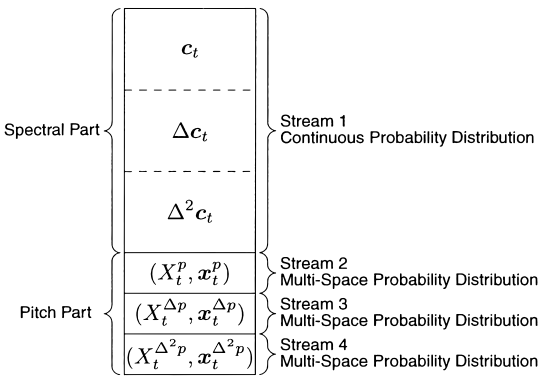


図 2　観測ベクトル
Fig. 2　Observation vector.

3. MLLR を用いた話者適応

3.1 多空間上の確率分布における MLLR

任意話者の声質・韻律特徴をもった音声を合成するため，目標とする話者の少量の音声データにより平均声 HMM の話者適応を行う．話者適応手法としては様々な手法が提案されているが，ここでは MLLR [5] を用いる．

MLLR では，状態 i の出力分布 $\mathcal{N}_i(\boldsymbol{x})$ の平均を $\boldsymbol{\mu}_i$ としたとき，平均ベクトルの推定値を

$$\hat{\boldsymbol{\mu}}_i = \boldsymbol{W}_i \boldsymbol{\xi}_i, \qquad \boldsymbol{\xi}_i = [1 \;\; \boldsymbol{\mu}_i^T]^T \tag{4}$$

により求める．ここで，$\boldsymbol{W}_i$ は適応のための回帰行列であり，適応データから最ゆう推定により求める．

スペクトルパラメータの話者適応の場合，通常の連続出力分布に対する MLLR を直接適用することができる [9]．一方，ピッチモデルの適応にはそのままでは適用できない．そこで，以下では MLLR を MSD-HMM に拡張する．

MSD-HMM の状態 i，空間 g（ただし次元 $n_g > 0$）の出力分布 $\mathcal{N}_{ig}(\boldsymbol{x})$ の平均を $\boldsymbol{\mu}_{ig}$，共分散行列を $\boldsymbol{U}_{ig}$ とする．

適応データ $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T)$ が与えられたとき，平均 $\boldsymbol{\mu}_{ig}$ の推定値 $\hat{\boldsymbol{\mu}}_{ig}$ を

$$\hat{\boldsymbol{\mu}}_{ig} = \boldsymbol{W}_{ig} \boldsymbol{\xi}_{ig}, \qquad \boldsymbol{\xi}_{ig} = [1 \;\; \boldsymbol{\mu}_{ig}^T]^T \tag{5}$$

と表す．行列 $\boldsymbol{W}_{ig}$ は平均ベクトルの適応のための回帰行列であり，適応データに対し，適応文章から生成される文章 HMM のゆう度を最大とするように推定される（教師付き適応）．しかし，最ゆう推定値を直接求めることは一般に困難なことから，現在のモデルを λ'，再推定したモデルを λ としたときの Q 関数を

$$Q(\lambda', \lambda) = \sum_{\text{all } \boldsymbol{q}, \boldsymbol{l}} P(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{l} | \lambda') \log P(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{l} | \lambda) \tag{6}$$

と定義し，これに関して最大化を行う．ただし，$\boldsymbol{q} = (q_1, q_2, \ldots, q_T)$ を許される状態系列，$\boldsymbol{l} = (l_1, l_2, \ldots, l_T) \in \{S(\boldsymbol{o}_1) \times S(\boldsymbol{o}_2) \times \ldots \times S(\boldsymbol{o}_T)\}$ を

観測系列 $\boldsymbol{O}$ に対して許される空間インデックスの系列としている．

ある条件のもとで，Q 関数は臨界点において唯一の大局的最大値をとることが証明されている [12]．

ここで時刻 t の観測ベクトル $V(\boldsymbol{o}_t)$ が，状態 i，空間 g において出力される確率 $\gamma_{ig}(t)$ を定義する．

$$\begin{aligned}\gamma_{ig}(t) &= P(q_t = i, l_t = g|\boldsymbol{O}, \lambda) \\ &= \frac{1}{P(\boldsymbol{O}|\lambda)} P(\boldsymbol{O}, q_t = i, l_t = g|\lambda) \end{aligned} \tag{7}$$

観測事象 $\boldsymbol{o}_t$ の空間インデックス集合が空間インデックス g を含むような時刻 t の集合を

$$T(\boldsymbol{O}, g) = \{t | g \in S(\boldsymbol{o}_t)\} \tag{8}$$

と定義し

$$\begin{aligned}&h(\boldsymbol{o}_t, i, g) \\ &\quad = (V(\boldsymbol{o}_t) - \boldsymbol{W}_{ig}\boldsymbol{\xi}_{ig})^T \boldsymbol{U}_{ig}^{-1} (V(\boldsymbol{o}_t) - \boldsymbol{W}_{ig}\boldsymbol{\xi}_{ig})\end{aligned} \tag{9}$$

とする．これらを用いて式 (5) は

$$\begin{aligned}&Q(\lambda', \lambda) \\ &\quad = K + \sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{t\in T(\boldsymbol{O},g)} P(\boldsymbol{O}, q_t = i, l_t = g|\lambda') \\ &\qquad \cdot \Big(\log w_{ig} - \frac{n_g}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{U}_{ig}^{-1}| \\ &\qquad - \frac{1}{2} h(\boldsymbol{o}_t, i, g) \Big) \\ &\quad = K + P(\boldsymbol{O}|\lambda') \sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{t\in T(\boldsymbol{O},g)} \gamma_{ig}(t) \Big(\log w_{ig} \\ &\qquad - \frac{n_g}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{U}_{ig}^{-1}| - \frac{1}{2} h(\boldsymbol{o}_t, i, g) \Big)\end{aligned} \tag{10}$$

と書くことができる．ここで，n_g は空間 Ω_g の次元を，w_{ig} は状態 i，空間 Ω_g の空間重みを，また K は定数項を表す．

ゆう度を最大化する $\boldsymbol{W}_{ig}$ を求めるため，式 (10) で表される Q 関数を $\boldsymbol{W}_{ig}$ で微分することにより

$$\begin{aligned}&\frac{\partial}{\partial \boldsymbol{W}_{ig}} Q(\lambda', \lambda) \\ &= -\frac{1}{2} P(\boldsymbol{O}|\lambda') \frac{\partial}{\partial \boldsymbol{W}_{ig}} \sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{t\in T(\boldsymbol{O},g)} \gamma_{ig}(t) h(\boldsymbol{o}_t, i, g) \\ &= P(\boldsymbol{O}|\lambda') \sum_{t\in T(\boldsymbol{O},g)} \gamma_{ig}(t) \boldsymbol{U}_{ig}^{-1} (V(\boldsymbol{o}_t) - \boldsymbol{W}_{ig}\boldsymbol{\xi}_{ig}) \boldsymbol{\xi}_{ig}^T\end{aligned} \tag{11}$$

が得られ，$\partial Q(\lambda', \lambda)/\partial \boldsymbol{W}_{ig} = 0$ とおくことにより，次式の解として $\boldsymbol{W}_{ig}$ が求まる．

$$\begin{aligned}&\sum_{t\in T(\boldsymbol{O},g)} \gamma_{ig}(t) \boldsymbol{U}_{ig}^{-1} V(\boldsymbol{o}_t) \boldsymbol{\xi}_{ig}^T \\ &\quad = \sum_{t\in T(\boldsymbol{O},g)} \gamma_{ig}(t) \boldsymbol{U}_{ig}^{-1} \boldsymbol{W}_{ig} \boldsymbol{\xi}_{ig} \boldsymbol{\xi}_{ig}^T\end{aligned} \tag{12}$$

3.2 スペクトル，ピッチの話者適応

一般に，適応データは少量なため，すべての状態に対する変換行列を求めることはできない．そこで，いくつかの状態で回帰行列 $\boldsymbol{W}_{ig}$ を共有することで適応データの存在しない分布の適応を行う．本論文では，LBG アルゴリズムにより，リーフノードが分布となる 2 分木を作成し適応データ量があるしきい値より大きくなる最下位ノードにおいて分布の適応を行う．

まず，空間 g のすべての分布の集合を $C = \{\mathcal{N}_1, \mathcal{N}_2, \cdots, \mathcal{N}_I\}$ とする．ただし I は総分布数である．このとき，C の適応データ量の期待値は，

$$d = \sum_{i=1}^{I}\sum_{t\in T(\boldsymbol{O},g)} \gamma_{ig}(t) \tag{13}$$

となる．C を各分布の平均ベクトルのユークリッド距離に基づいて LBG アルゴリズムにより二つのノードに分割し，C_l，C_r を求める．C_l，C_r それぞれを新たに C とし，C の適応データ量の期待値 d があるしきい値より小さくなるまで再帰的に繰り返し，2 分木を作成する．この 2 分木のリーフノードの親ノードにおいて回帰行列を求め，当該リーフノードに含まれるすべての分布の話者適応を行う．

スペクトル，ピッチの同時適応の際には，ストリームごとに 2 分木を作成し，話者適応を行う．スペクトル部は連続出力分布に対する MLLR，ピッチ，Δ ピッチ，Δ^2 ピッチ部は MSD-HMM に対する MLLR により話者適応を行う．実際には，ピッチ，Δ ピッチ，Δ^2 ピッチ部は，分布 $\mathcal{N}_1(\boldsymbol{x})$ の平均ベクトルを適応する．各ストリームに式 (12) を適用する際，$\gamma_{ig}(t)$ は

全ストリームで共通の状態遷移を用いて求める．また，$\boldsymbol{\xi}_{ig}$，$\boldsymbol{U}_{ig}$ は各ストリームのモデルパラメータを用い，$T(\boldsymbol{O}, g)$，$\boldsymbol{W}_{ig}$ はストリームごとに求める．

4. 実　　験

4.1 実験条件

HMM の学習データとして，ATR 音韻バランス文を用いた．無音を含む 42 種類の音素を単位とし，コンテクスト情報の含まれるラベルを作成して用いた．用いたコンテクストの詳細は，文献 [4] に示されている．サンプリングレート 16 kHz の音声信号を，フレーム長 25 ms，フレーム周期 5 ms のブラックマン窓を用いてメルケプストラム分析し，0 次から 24 次のメルケプストラムを求めた．ピッチは対数基本周波数を特徴パラメータとした．これらのパラメータに，デルタ及びデルタデルタパラメータを加えた 78 次のベクトルを特徴ベクトルとし，5 状態の left-to-right HMM によりモデル化した．

4.2 話者 MHT への適応実験

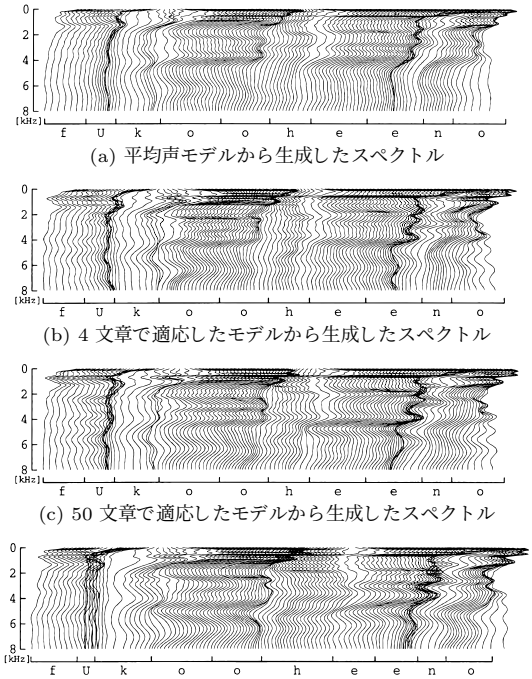
提案手法の有効性を確認するため，目標話者を男性話者 MHT とし，話者適応を行った．

平均声 HMM は，ATR データベース B セットに含まれる話者 6 名中目標話者 MHT を除いた男性 5 名，各 400 文章を学習データとして作成した．MDL 基準に基づく決定木によるコンテクストクラスタリングにより状態の共有を行った結果，平均声 HMM の状態数は，スペクトル部が 3,765，ピッチ部が 12,761，継続長の分布数が 6,318 となった．適応データとしては学習データに含まれない 4，20，または 50 文章を用いた．話者適応の際，適応データ量のしきい値は，実験的にスペクトル部 1,500，ピッチ部 100 とした．この結果回帰行列の数は，4 文章で適応した場合では，スペクトル部 2，ピッチ部 25，Δ，Δ^2 ピッチ部はそれぞれ 21 となり，50 文章では，スペクトル部 13，ピッチ部 175，Δ ピッチ部 163，Δ^2 ピッチ部 165 となった．また評価のために話者 MHT の 450 文章を用いて特定話者モデルを作成した．特定話者モデルの状態数は，スペクトル部 906，ピッチ部 1,894 である．

なお，本論文では継続長分布の話者適応は行っていないため，適応モデルの継続長分布は平均声モデルの継続長分布をそのまま用いている．

4.2.1 スペクトル生成例

図 3 に，学習データに含まれない「不公平の存在は否認しなかった」という文章に対して生成したスペクトルの例を示す．図 3 (a) は平均声モデル，(b) は 4 文章で話者適応したモデル，(c) は 50 文章で話者適応したモデル，(d) は目標話者モデルからそれぞれ生成したスペクトルである．継続長分布の話者適応を行っていないため，適応モデル，平均声モデルの継続長は常に同じで，目標話者モデルの継続長と異なっていることに注意する．



(a) 平均声モデルから生成したスペクトル

(b) 4 文章で適応したモデルから生成したスペクトル

(c) 50 文章で適応したモデルから生成したスペクトル

(d) 目標話者モデルから生成したスペクトル

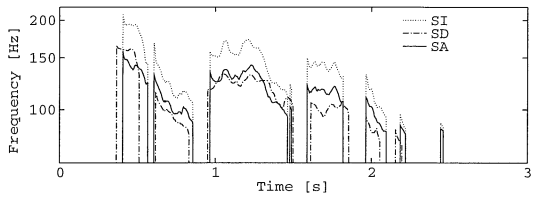
図 3　スペクトルの生成例
Fig. 3　Examples of generated spectra.

この図より，話者適応を行うことで，少ない文章数でも目標とする話者のスペクトルに近づくことがわかる．

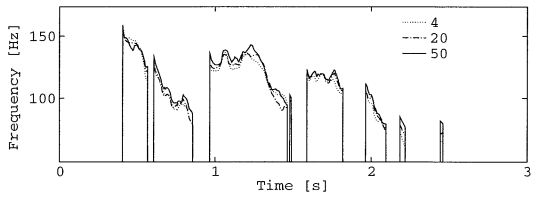
4.2.2 ピッチパターン生成例

図 4 に，スペクトルと同様「不公平の存在は否認しなかった」という文章に対して生成したピッチパターンの例を示す．図 4 (a) は平均声モデル（SI），目標話者モデル（SD），50 文章で適応したモデル（SA）からそれぞれ生成したピッチパターンである．

この図より，スペクトルと同様，話者適応を行うことで，目標とする話者のピッチパターンに近づくことがわかる．図 4 (b) は適応文章数を 4，20，50 文章としたときの適応モデルから生成したピッチパターンを示す．図より，4 文章程度の少ない文章数でも，50 文

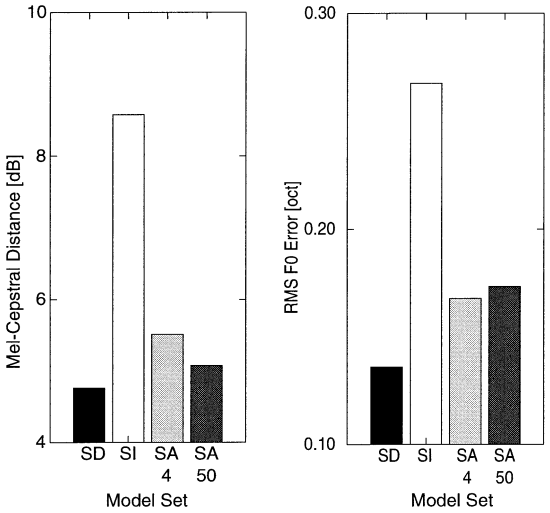


(a) 話者適応の効果



(b) 適応文章数の影響

図 4　ピッチパターンの生成例
Fig. 4　Examles of generated pitch contour.



(a) メルケプストラム距離　(b) 対数基本周波数の RMS 誤差

図 5　話者 MHT の客観評価結果
Fig. 5　Results of objective evaluations for speaker MHT.

章での適応と同程度なピッチパターンが生成できていることがわかる．また，4，20，50 文章を用いて適応したモデルから生成したピッチパターンを比較すると，概形は近いものの，細部において異なるパターンが生成されているため，単純な平行移動やスケーリングによるピッチ変換とは異なり適応データに応じて変化していることがわかる．

4. 2. 3　客観評価試験

客観評価として，平均メルケプストラム距離及び対数基本周波数の RMS 誤差を求めた．評価用データは，学習データ，適応データに含まれない 53 文章とし，目標話者の実際の発声から求めたパラメータと比較した．距離計算のため，合成時の状態継続長は Viterbi アルゴリズムにより目標話者の実際の発声にアラインメントした結果を用いた．

図 5 (a) に生成したメルケプストラムと目標話者の実際の発声から求めたメルケプストラムとの全フレームでの平均メルケプストラム距離を示す．左から，目標話者モデル，平均声モデル，4 文章で適応したモデル，50 文章で適応したモデルの場合を示している．図より，話者適応することで目標話者の実際の発声に近づいていることがわかる．4 文章程度の少ない文章数でも，目標話者の SD モデル，50 文章で適応したモデルに近い距離となっている．

図 5 (b) に生成した対数基本周波数と目標話者の実際の発声から求めた対数基本周波数との RMS 誤差を示す．平均メルケプストラム距離と同様，左から，目標話者モデル，平均声モデル，4 文章で適応したモデル，50 文章で適応したモデルの場合を示している．対数基本周波数は，無声音部では値をもたないため，モデルから生成したパラメータ及び目標話者の発声から求めたパラメータ両方が値をもつ部分で計算した．図より，対数基本周波数も話者適応することにより目標話者の実際の発声に近づき，4 文章程度の少ない文章数でも，平均声モデルと比べ，目標話者の SD モデルに近い値となっている．

4. 2. 4　主観評価試験

提案法の有効性を確認するため，ABX 法による主観評価試験を行った．ABX 法による受聴試験においては，A を SI モデルからの合成音声，B を SD モデルからの合成音声，X を適応モデルからの合成音声とし，被験者に A，B，X または B，A，X の順に提示し，X が 1 番目または 2 番目のどちらに近いかを判定させた．このとき，判定できない場合はどちらともいえない（“Undecided”）としている．被験者は 6 名である．主観評価試験の適応データは，学習データに含まれない 4 または 50 文章を用いた．評価に用いた音声サンプルは，学習データ及び適応データに含まれない 4 文章の合成音声とした．被験者は，各音声サンプルに対し 2 回ずつ（A，B，X とした場合及び B，A，X とした場合）判断を行うことになる．

図 6 に ABX 法による受聴試験の結果を示す．図中“Spectrum”，“Pitch” はそれぞれスペクトルまたは

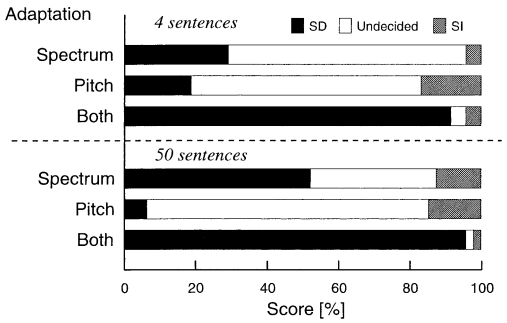


図 6　話者 MHT の主観評価結果
Fig. 6　Results of subjective evaluations for speaker MHT.

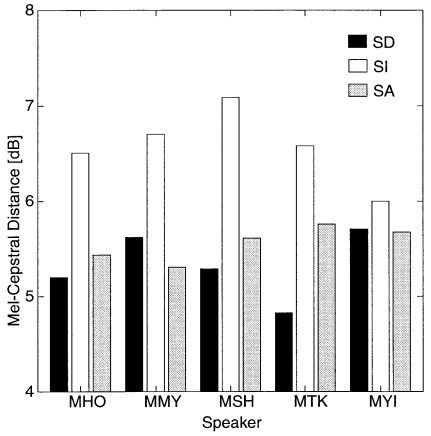


図 7　話者別のメルケプストラム距離
Fig. 7　Mel-cepstral distance for each speaker.

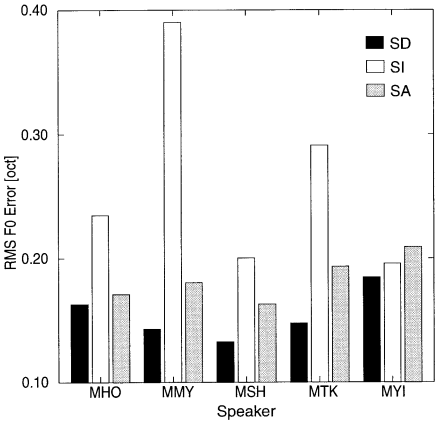


図 8　話者別の対数基本周波数の RMS 誤差
Fig. 8　RMS value of log fundamental frequency for each speaker.

ピッチのみを適応した場合を表し，“Both” はピッチ及びスペクトルを同時に適応した場合を表す．主観評価の結果，適応モデルからの合成音声が目標話者の SD モデルに近いと判定された割合は，4 文章では，スペクトルのみを適応した場合が 29.2%，ピッチのみ適応した場合が 18.8%，同時適応した場合が 91.7%となり，50 文章では，スペクトルのみ適応した場合が 52.1%，ピッチのみ適応した場合が 6.2%，同時適応した場合が 95.8%となった．

評価試験の結果より，スペクトルのみ，ピッチのみの話者適応ではどちらともいえないという判定が多いのに対し，スペクトルとピッチを同時に適応することにより，適応した合成音声が目標とする話者の音声に大幅に近づくことがわかる．また，50 文章で適応した結果を見ると，ピッチのみ適応した場合より，スペクトルのみ適応した場合の方が，目標とする話者の個人性を再現しているという結果になっている．更に，4 文章程度の少ない文章でもスペクトル，ピッチを同時に適応することにより，目標話者の SD モデルからの合成音にかなり近づくことがわかる．

4.3　様々な話者への適応

平均声と目標話者の組合せにより，話者適応の効果が変化すると考えられることから，複数の話者を目標話者として実験を行った．ATR データベース B セットに含まれる話者 6 名中，1 人を目標話者，残り 5 名を平均声モデルの学習話者とし，話者 MHT 以外の 5 名について実験を行った．

4.2.3 と同様の方法で目標話者の実際の発声とのメルケプストラム距離及び対数基本周波数の RMS 誤差を求めた結果を図 7，図 8 に示す．評価用データは学習データ，適応データに含まれない 53 文章である．話者 MHT における結果より，話者適応に用いた文章数は 4 文章とした．図 7 より，話者により，効果に違いがあるものの，目標話者の実際の発声とのメルケプストラム距離は，話者適応することにより平均声より小さくなることがわかる．また，図 8 より，対数基本周波数も，話者 MYI 以外の 4 名では，適応モデルの誤差は，平均声より小さくなり，目標話者モデルとの誤差に近づいていることがわかる．話者 MYI は，平均声モデルより適応モデルの誤差が大きくなっているが，これはもともと平均声モデルの誤差と目標話者モデルの誤差が近いためであり，適応モデルの誤差も平均声モデル，目標話者モデルの誤差とほぼ同等である

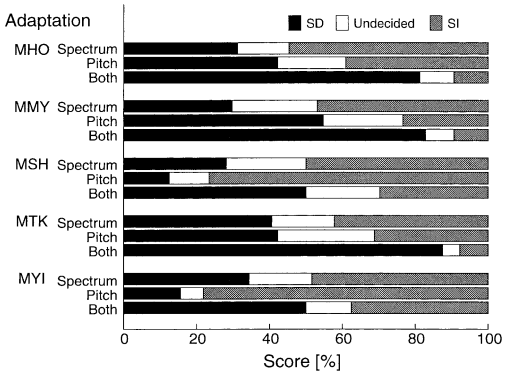


図 9　話者別の主観評価結果
Fig. 9　Results of subjective evaluations for each speaker.

と考えられる．

図 9 に，主観評価試験の結果を示す．**4. 2. 4** と同様，ABX 法により主観評価試験を行った．被験者は 8 名である．評価に用いた音声サンプルは，適応データ，学習データに含まれない 53 文章から被験者ごと及び目標話者ごとにランダムに選んだ 4 文章とした．図 9 より，いずれの話者においても，スペクトルのみ，ピッチのみの適応と比べ，ピッチとスペクトルを同時に適応することにより，目標話者に近いと判定された割合が増えることがわかる．同時適応した場合のスコアは，MHO が 81.2%，MMY が 82.8%，MSH が 50.0%，MTK が 87.5%，そして，MYI が 50.0%となった．話者 MHO，MMY，MTK においては，80%以上が目標話者モデルからの合成音声に近いと判定されており，話者適応が有効であることがわかる．話者 MSH は 50.0%とスコアが低いが，対数基本周波数の RMS 誤差を見ると，平均声モデルとの RMS 誤差が他の話者に比べて小さく，また平均声モデルと適応モデルの RMS 誤差の差及び適応モデルと目標話者モデルの RMS 誤差の差がほぼ同じである．このため，平均声モデル，目標話者モデルでどちらともいえないという判定が増えたものと考えられる．話者 MYI も 50.0%と低いスコアになっているが，平均声モデルのメルケプストラム距離及び対数基本周波数の RMS 誤差が目標話者モデルとほぼ同等であるため，評価の際に判定が困難であったものと考えられる．

5. むすび

本論文では，HMM に基づく音声合成システムにおいて，話者適応により，スペクトルだけでなく，韻律特徴の一つであるピッチも任意の話者の音声に似せる手法を提案した．ピッチパラメータの話者適応のため，MSD-HMM に対する MLLR を導出し，主観評価実験により評価した．その結果，目標話者の少量の音声データを用いてスペクトルとピッチを同時に適応することにより，目標話者の声に近い声で音声が合成できることを示した．今後の課題は音素継続長の話者適応が挙げられる．

謝辞　本研究の一部は，科学研究費補助金（課題番号 13878070）によった．また，音声データベースの言語ラベルを提供して頂いた名古屋工業大学岩田鋼司氏，沢辺敦氏に感謝します．

文　献

[1] A.W. Black and N. Cambpbell, “Optimising selection of units from speech database for concatenative synthesis,” Proc. EUROSPEECH-95, pp.581–584, Sept. 1995.

[2] A.K. Syrdal, C.W. Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Storm, K. Lee, and M.J. Makashay, “Corpus-based techniques in the AT&T NEXTGEN synthesis system,” Proc. ICSLP-2000, pp.411–416, Oct. 2000.

[3] 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.

[4] 吉村貴克，徳田恵一，益子貴史，小林隆夫，北村　正，“HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化，” 信学論（D-II），vol.J83-D-II, no.11, pp.2099–2107, Nov. 2000.

[5] 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, no.2, pp.171–185, 1995.

[6] M. Tonomura, T. Kosaska, and S. Matsunaga, “Speaker adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation,” Computer Speech and Language, vol.10, no.2, pp.117–132, 1996.

[7] J. Takahashi and S. Sagayama, “Vector-field-smoothed Baysian learning for fast and incremental speaker/telephone-channel adaptation,” Computer Speech and Language, vol.11, no.2, pp.127–146, 1997.

[8] 益子貴史，田村正統，徳田恵一，小林隆夫，今井　聖，“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，” 信学論（D-II），vol.J83-D-II, no.12, pp.2509–2516, Dec. 2000.

[9] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “Speaker adaptation for HMM-based speech synthesis system using MLLR,” The Third ESCA/COCOSDA Workshop on Speech Synthesis, pp.273–276, Nov. 1998.

[10] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Speaker interpolation for HMM-based speech synthesis system,” J. Acoust. Soc. Jpn. (E), vol.21, no.4, pp.199–206, April 2000.
[11] 沢部　敦，七里健吾，吉村貴克，徳田恵一，益子貴史，小林隆夫，北村　正，“固有声に基づいた triphone による音声合成，” 音講論，pp.299–301, March 2001.
[12] 徳田恵一，益子貴史，宮崎　昇，小林隆夫，“多空間上の確率分布に基づいた HMM，” 信学論（D-II），vol.J79-D-II, no.7, pp.1579–1589, July 2000.
[13] 益子貴史，徳田恵一，宮崎　昇，小林隆夫，“多空間確率分布 HMM によるピッチパターン生成，” 信学論（D-II），vol.J83-D-II, no.7, pp.1600–1609, July 2000.
[14] K. Shinoda and T. Watanabe, “Acoustic modeling based on the MDL criterion for speech recognition,” Proc. EUROSPEECH-97, pp.99–102, Sept. 1997.
[15] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Duration modeling for HMM-based speech synthesis,” Proc. ICSLP-98, pp.29–32, Dec. 1998.
[16] 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.
[17] 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.

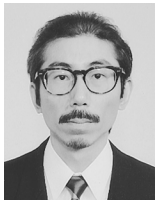
（平成 13 年 6 月 7 日受付，10 月 9 日再受付）

田村　正統　（学生員）

平 9 東工大・工・情工卒．現在同大大学院総合理工学研究科物理情報システム創造専攻博士後期課程在学中．音声合成，マルチモーダルインタフェースの研究に従事．日本音響学会会員．

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システム創造専攻助手．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．平 13 本会論文賞，猪瀬賞各受賞．日本音響学会，IEEE，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．平 13 電気通信普及財団賞，平 13 本会論文賞，猪瀬賞各受賞．日本音響学会，情報処理学会，人工知能学会，IEEE，ISCA 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システム創造専攻教授．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．平 13 電気通信普及財団賞，平 13 本会論文賞，猪瀬賞各受賞．日本音響学会，情報処理学会，IEEE，ISCA 各会員．