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

2007－ＳＬＰ－69 (18)
2007／12／20

声質変換のためのスペクトル・F_0の同時モデリング

宇藤 陽介† 南角 吉彦† 李 晃伸† 徳田 恵一†

† 名古屋工業大学 大学院工学研究科 情報工学専攻
〒466–8555 名古屋市 昭和区 御器所町

あらまし 声質変換とは，ある話者が発した音声を別の話者が発したかのような音声に変換する技術であり，任意の音声を合成する音声合成システムよりも少量の学習データで実現可能である．従来の声質変換ではスペクトルをガウス混合モデル (Gaussian Mixture Model; GMM) でモデル化し，非線形に変換する手法が広く用いられる．しかし，F_0の変換に関してはスペクトルとは独立に線形変換が用いられることが多かった．これは，F_0が有声区間のみで定義されており，無声区間では値を持たず，系列全体を通常の連続分布や離散分布でモデル化することが容易ではないためである．本報告では，多空間上の確率分布 (Multi-Space Probability Distribution; MSD) に基づく GMM (MSD-GMM) を用いたスペクトルとF_0の同時変換手法を提案する．提案法では，F_0の非線形変換が可能になるだけでなく，有声から無声や無声から有声への変換も可能となる．さらに本研究では，F_0の時間方向の変動をモデル化するために MSD-HMM への拡張を検討する．

キーワード 声質変換，F_0変換，MSD-GMM，MSD-HMM

Simultaneous Modeling of Spectrum and F_0 for Voice Conversion

Yosuke UTO†, Yoshihiko NANKAKU†, Akinobu LEE†, and Keiichi TOKUDA†

† Depertment of Computer Science and Engineering, Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya, 466–8555 Japan

Abstract This paper proposes a simultaneous modeling spectrum and F_0 for voice conversion based on MSD (Multi-Space Probability Distribution) models. In conventional voice conversion, the spectral conversion technique based on GMM (Gaussian Mixture Model) has been proposed. Although spectral feature sequences are nonlinearly converted based on GMM, F_0 sequences are converted by a simple linear function. This is because F_0 is undefined in unvoiced seguments; therefore F_0 sequences cannot be modeled by neither continuous nor discrete distributions. To overcome this problem, we apply MSD-GMM to the voice conversion. The MSD-GMM allows to model continuous F_0 values in voice frames and discrete symbol representing unvoiced frames in a unified framework. Futhermore, the MSD-HMM is adopted to model a long time dependency in F_0 sequences.

Key words Voice conversion, F_0 conversion, MSD-GMM, MSD-HMM

1. はじめに

近年，合成音の需要の高まりに伴い多様な合成音の必要性が高まっている．隠れマフコフモデル (Hidden Marcov Model; HMM) を用いた音声合成手法 [1] では比較的少量の学習データで高音質の音声の合成が可能であるが，多様な合成音の作成には話者ごとにデータを用意する必要があり，コスト面などから実現は容易ではない．この問題を解決する手法の１つとして，声質変換の研究が進められている．声質変換とは，ある話者の音声を別の話者の音声に変換する技術であり，音声合成システムよりもさらに少量の学習データで実現可能である．

近年，最も研究が進められている声質変換は，GMM に基づく声質変換手法 [2] である．この手法では元話者と目標話者の関係を GMM により学習し，入力データが与えられた時の事後確率に基づいてスペクトルを非線形に変換することができる．しかし，F_0の変換に関してはスペクトルとは独立に線形に変換されることが多かった．これは，F_0が有声区間のみで定義されており，無声区間では値を持たず，系列全体を通常の連続分布や離散分布ではモデル化ができないためである．

声質変換では，元話者と目標話者の特徴量を結合するため，F_0の学習データとしては元話者，目標話者の順に「有声–有声」「有声–無声」「無声–有声」「無声–無声」の 4 種類が存在する．

このうち「有声–有声」の特徴量のみを使用して GMM を学習し，変換する手法 [3] も提案されているが，他のフレームを捨てていることになり統計モデルとしては不十分である．

そこで，本報告では，F_0のモデル化に MSD モデル [4] を利用したスペクトルとF_0の同時モデリング手法を提案する．提案法では，「有声–有声」「有声–無声」「無声–有声」「無声–無声」の特徴量をそれぞれ別の空間でモデル化する．また提案法では，変換時に「有声–無声」「無声–有声」の空間の分布を使用する場合があり，有声から無声への変換やその逆の変換が可能となる．本研究では，MSD モデルとして MSD-GMM と MSD-HMM を使用する．MSD-GMM では，従来のスペクトル変換と同様にF_0の非線形変換を実現できる．さらに MSD-HMM では，音素コンテキストを考慮したトポロジーを用いることにより，音素より長い単位での時間的な相関をモデル化可能と期待される．

以下，2. では GMM に基づく声質変換について，3. では MSD-GMM に基づく声質変換について，4. では MSD-HMM に基づく声質変換について，5. では声質変換実験について述べる．

2. GMM に基づく声質変換

2.1 スペクトル変換

GMM に基づく声質変換におけるスペクトル変換 [5], [6] について説明する．まず，GMM の学習データとして，DP マッチングなどによりフレームごとに対応付けられた変換前後のスペクトル特徴量を用意する．すなわち，時刻 t における変換前後の特徴量をそれぞれ $\boldsymbol{X}_t$, $\boldsymbol{Y}_t$ とし，これらを結合したベクトル列 $\boldsymbol{Z} = \left[\boldsymbol{Z}_1^\top, \boldsymbol{Z}_2^\top, \ldots, \boldsymbol{Z}_T^\top\right]^\top$, $\boldsymbol{Z}_t = \left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top\right]^\top$ を学習データとする．このデータを用いて次式の尤度関数を最大にする GMM のモデルパラメータ $\boldsymbol{\lambda}$ を推定する．

$$p(\boldsymbol{Z} \mid \boldsymbol{\lambda}) = \prod_t^T \sum_i^M w_i \mathcal{N}(\boldsymbol{X}_t; \boldsymbol{\mu}_i^{(Z)}, \boldsymbol{\Sigma}_i^{(Z)}) \tag{1}$$

$$\boldsymbol{\mu}_i^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_i^{(X)} \\ \boldsymbol{\mu}_i^{(Y)} \end{bmatrix}, \boldsymbol{\Sigma}_i^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{(XX)} & \boldsymbol{\Sigma}_i^{(XY)} \\ \boldsymbol{\Sigma}_i^{(XX)} & \boldsymbol{\Sigma}_i^{(YY)} \end{bmatrix} \tag{2}$$

ここで w_i は混合重み，M は混合数である．変換前後の話者の特徴量系列をそれぞれ $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \ldots, \boldsymbol{X}_T^\top\right]^\top$, $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, \ldots, \boldsymbol{Y}_T^\top\right]^\top$ とすると，$\boldsymbol{X}$ が与えられた際の $\boldsymbol{Y}$ の確率分布は次式で与えられる．

$$\begin{aligned} p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\lambda}) &= \sum_{\text{all } \boldsymbol{m}} p(\boldsymbol{m} \mid \boldsymbol{X}, \boldsymbol{\lambda}) p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}) \\ &= \sum_{\text{all } \boldsymbol{m}} \prod_{t=1}^T \Big[p(m_t \mid \boldsymbol{X}_t, \boldsymbol{\lambda}) p(\boldsymbol{Y}_t \mid \boldsymbol{X}_t, m_t, \boldsymbol{\lambda}) \Big] \end{aligned} \tag{3}$$

ここで，$\boldsymbol{m} = (m_1, m_2, \ldots, m_T)$ は混合要素番号の系列を表す．また，t 番目のフレームにおける $p(m_t \mid \boldsymbol{X}_t, \boldsymbol{\lambda})$, $p(\boldsymbol{Y}_t \mid \boldsymbol{X}_t, m_t, \boldsymbol{\lambda})$ は以下で与えられる．

$$\begin{aligned} &p(m_t = i \mid \boldsymbol{X}_t, \boldsymbol{\lambda}) \\ &= \frac{w_i \mathcal{N}(\boldsymbol{X}_t; \boldsymbol{\mu}_i^{(X)}, \boldsymbol{\Sigma}_i^{(XX)})}{\sum_{j=1}^M w_j \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_j^{(X)}, \boldsymbol{\Sigma}_j^{(XX)})} \end{aligned} \tag{4}$$

$$p(\boldsymbol{Y}_t \mid \boldsymbol{X}_t, m_t = i, \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{Y}_t; \boldsymbol{E}_t(i), \boldsymbol{D}(i)) \tag{5}$$

ただし，

$$\boldsymbol{E}_t(i) = \boldsymbol{\mu}_i^{(Y)} + \boldsymbol{\Sigma}_i^{(YX)} \boldsymbol{\Sigma}_i^{(XX)^{-1}} (\boldsymbol{X}_t - \boldsymbol{\mu}_i^{(X)}) \tag{6}$$

$$\boldsymbol{D}(i) = \boldsymbol{\Sigma}_i^{(YY)} - \boldsymbol{\Sigma}_i^{(YX)} \boldsymbol{\Sigma}_i^{(XX)^{-1}} \boldsymbol{\Sigma}_i^{(XY)} \tag{7}$$

である．尤度最大化基準として最適な変換特徴量は式 (3) を最大化することで得られる．ただし，本研究では簡単のため，式 (3) の $\boldsymbol{m}$ に関する和を，事後確率 $p(\boldsymbol{m} \mid \boldsymbol{X}, \boldsymbol{\lambda})$ を最大にする 1 本の $\boldsymbol{m}$ で近似する．このとき，式 (3) の対数は次式で表される．

$$\begin{aligned} &\log p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}) \\ &= -\frac{1}{2} \boldsymbol{Y}^\top \boldsymbol{D}^{-1} \boldsymbol{Y} + \boldsymbol{Y}^\top \boldsymbol{D}^{-1} \boldsymbol{E} + K \end{aligned} \tag{8}$$

ただし，

$$\boldsymbol{E} = [\boldsymbol{E}_1(m_1), \boldsymbol{E}_2(m_2), \ldots, \boldsymbol{E}_T(m_T)] \tag{9}$$

$$\boldsymbol{D}^{-1} = \text{diag}\left[\boldsymbol{D}^{-1}(m_1), \boldsymbol{D}^{-1}(m_2), \ldots, \boldsymbol{D}^{-1}(m_T)\right] \tag{10}$$

であり，K は $\boldsymbol{Y}$ と独立な定数である．よって式 (8) を最大とする系列は $\boldsymbol{Y} = \boldsymbol{E}$ となる．さらに本研究では，滑らかな特徴系列を得るために動的特徴量を考慮した変換を行う [7]．スペクトルの静的特徴量 $\boldsymbol{x}_t$, $\boldsymbol{y}_t$ にそれぞれの動的特徴量 $\Delta\boldsymbol{x}_t$, $\Delta\boldsymbol{y}_t$ を加えたベクトル $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$, $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$ を新たな特徴ベクトルとする．このとき，静的特徴系列 $\boldsymbol{y}$ と静的・動的特徴系列 $\boldsymbol{Y}$ の間には $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}$ の関係が成り立つ．よって，静的特徴量 $\boldsymbol{y}$ について式 (8) を最大化することにより，動的特徴量を考慮した最適なスペクトル特徴量は次式で与えられる．

$$\boldsymbol{y} = (\boldsymbol{W}^\top \boldsymbol{D}^{-1} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{D}^{-1} \boldsymbol{E} \tag{11}$$

2.2 F_0変換

F_0の変換には次式で表される線形方程式が用いられる．

$$p_t^{(y)} = \frac{p_t^{(x)} - \mu^{(x)}}{\sigma^{(x)}} \times \sigma^{(y)} + \mu^{(y)} \tag{12}$$

ここで，$p_t^{(x)}$, $p_t^{(y)}$ はそれぞれ変換前，変換後のF_0の値，$\mu^{(x)}$, $\sigma^{(x)}$ は，元話者のF_0の平均と分散，$\mu^{(y)}$, $\sigma^{(y)}$ は，目標話者のF_0の平均と分散を表す．

3. MSD-GMM に基づく声質変換

3.1 MSD による特徴量のモデル化

MSD モデルでは，図 1 のように空間インデックス $g = 1, 2, \ldots, G$ により参照される G 個の空間 $(R^1, R^2, \ldots, R^G)$ を考える．これらの空間はそれぞれ異なった次元 n_g をもつことができる．また，各空間上には，確率密度分布 $(\mathcal{N}_1^{n1}, \mathcal{N}_2^{n2}, \ldots, \mathcal{N}_G^{nG})$ と重み $(w_1, w_2, \ldots, w_G)$ が定義されている．ただし，$\sum_{g=1}^G w_g =$

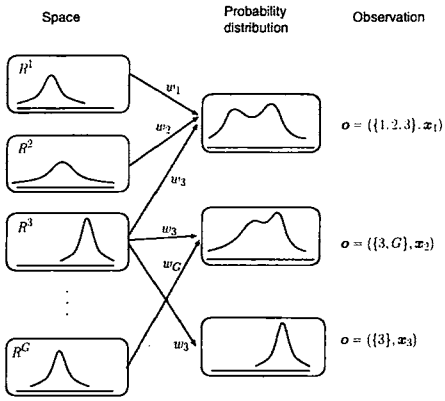

図 1 MSD と観測事象

Fig. 1 MSD and observation vectors

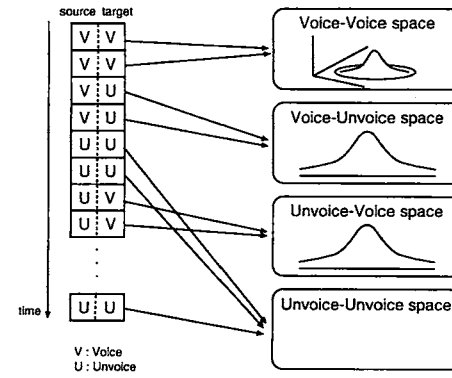

図 2 MSD によるF_0のモデル化

Fig. 2 F_0 modeling based on MSD

1とする．観測事象 o は次式のように，n 次元のベクトル $\boldsymbol{x}$ と，$\boldsymbol{x}$ がどの空間から出力されたかを表す空間インデックスの集合 $\boldsymbol{X}$ からなるものとする．

$$o = (\boldsymbol{X}, \boldsymbol{x}) \tag{13}$$

ただし，$\boldsymbol{X}$ に含まれる空間インデックスが表す空間は全て $\boldsymbol{x}$ と同じ n 次元とする．このとき，o の観測確率は次式で定義することができる．

$$p(o) = \sum_{g \in S(o)} w_g \mathcal{N}_g^{n_g}(V(o)) \tag{14}$$

ただし，

$$V(o) = \boldsymbol{x}, \quad S(o) = \boldsymbol{X} \tag{15}$$

を満たす．零次元空間からの観測系列 o は，空間インデックスの集合 $\boldsymbol{X}$ だけからなり，$\boldsymbol{x}$ の値は存在しないが，既述の便宜上，$\mathcal{N}_g^0(V(o)) = 1$ と定義する．このとき，全事象に関する $p(o)$ の積分値は

$$\int p(o)do = \sum_{g=1}^{G} w_g \int \mathcal{N}_g^{n_g}(\boldsymbol{x})d\boldsymbol{x} = 1 \tag{16}$$

となる．

3.2 スペクトル・F_0のモデル化

声質変換における MSD-GMM によるスペクトル・F_0 の同時モデル化について説明する．学習部ではスペクトルとF_0を 1 つの特徴量と考え，元話者・目標話者の特徴量 $\boldsymbol{X}_t = [c_t^{(x)\top}, \Delta c_t^{(x)\top}, p_t^{(x)\top}, \Delta p_t^{(x)\top}]^\top$, $\boldsymbol{Y}_t = [c_t^{(y)\top}, \Delta c_t^{(y)\top}, p_t^{(y)\top}, \Delta p_t^{(y)\top}]^\top$ を結合した系列，$\boldsymbol{Z} = [\boldsymbol{Z}_1^\top, \boldsymbol{Z}_2^\top, \ldots, \boldsymbol{Z}_T^\top]^\top$, $\boldsymbol{Z}_t = [\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top$ を学習データとし，式 (1) を最大にするモデルパラメータ $\boldsymbol{\lambda}$ を推定する．ただし，c, p はそれぞれスペクトル，F_0特徴量であり，$\Delta(\cdot)$ は動的特徴量を表す．

F_0の動的特徴量を考慮した場合，有声と無声の境目では静的特徴量が有声であっても，動的特徴量が計算できないフレームが存在する．しかし，そのようなフレームは非常に少量であり，直接的に MSD モデルで学習した場合，過学習の可能性がある．そこで，本研究ではこれらの特徴量は，静的・動的特徴量ともに無声シンボルへと変更した．その結果，結合されたF_0は，元話者，目標話者の順に「有声–有声」「有声–無声」「無声–有声」「無声–無声」の 4 種類が存在することになる．MSD モデルではこれらの 4 種類の特徴量を図 2 のように 4 つの空間でモデル化する．動的特徴量を考慮すると「有声–有声空間」は 4 次元，「有声–無声空間」と「無声–有声空間」は 2 次元，「無声–無声空間」は 0 次元の空間となる．また，各空間データは単一ガウス分布でモデル化することとした．

3.3 F_0の変換方法

MSD によるF_0の変換方法について説明する．まず，変換前の系列を用いて変換後の系列の有声/無声区間を決定する．変換前のF_0として有声データが与えられた場合は，「有声–有声空間」「有声–無声空間」，無声シンボルが与えられた場合は「無声–有声空間」「無声–無声空間」のうち事後確率が最大となる空間を用いて変換される．ただし，入力が無声シンボルの場合は，分布を持たないため事後確率は空間重みのみによって決定される．次に，有声区間の値を推定する．変換は各有声区間に対してスペクトルの変換と同様に，式 (11) を適用することで与えられる．ただし，無声区間との隣接フレームでは式 (7) における動的特徴量の分散が計算できないため，分散は無限大とした．

従来法との比較のため，混合数が 1 の場合を考えると，変換されたF_0は次式で表される (ただし，動的特徴量は考慮していない)．

$$p_t^{(y)} = \frac{p_t^{(x)} - \mu^{(x)}}{\sigma^{(xx)}} \times \sigma^{(yx)} + \mu^{(y)} \tag{17}$$

ここで，$p_t^{(x)}$, $p_t^{(y)}$ はそれぞれ変換前後のF_0の値，$\mu^{(x)}$, $\sigma^{(xx)}$ はそれぞれ元話者のF_0の平均と分散，$\mu^{(y)}$ は目標話者のF_0の平均を表す．また，$\sigma^{(yx)}$ は元話者と目標話者の相互共分散である．式 (17) は，式 (12) と同様に線形変換であるが元話者，目標話者の特徴量の相関を考慮した変換となっていることがわかる．

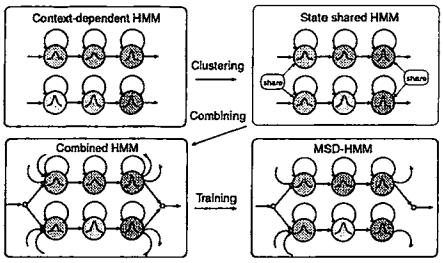

図 3　MSD-HMM の作成手順
Fig. 3　The training process of MSD-HMM

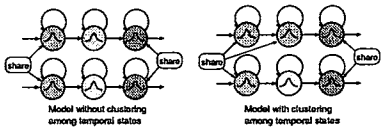

図 4　時間方向のクラスタリング
Fig. 4　The clustering among temporal states

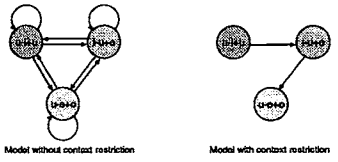

図 5　コンテキストの制限
Fig. 5　The context restriction

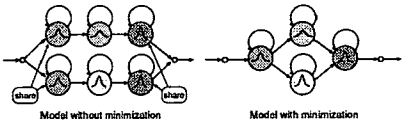

図 6　モデル構造の最小化
Fig. 6　The model structure minimization

4. MSD-HMM に基づく声質変換

4.1 MSD-HMM の構築

本研究では，F_0の音素より長い区間での時間的な相関をモデル化するため，音素コンテキストを考慮した MSD-HMM を構築する．しかし，変換時にコンテキストラベルを用いないため，コンテキストを隠れ要素とみなす巨大な 1 つの HMM に変換する必要がある．以下に HMM の作成手順を示す．また，図 3 に概要図を示す．

（1）　コンテキスト依存 HMM の作成

本研究ではフラットスタートによるモノフォンの学習後，トライフォンへコンテキストを展開し，再学習を行った．

（2）　コンテキストクラスタリング

表 1　実験条件
Table 1　Experimental condiction

学習データ	ATR 日本語データベース B-set 450 文
変換データ	ATR 日本語データベース B-set 53 文
変換話者	mtk → mht, mho → myi
特徴量	メルケプストラム (24 次元)+1 次動的特徴量 +F_0(1 次元)+1 次動的特徴量
GMM の混合数	32, 64, 128, 256, 512
HMM の分布数	メルケプストラム：32, 64, 128, 256, 512 F_0：メルケプストラムと同じ
共分散行列	全共分散行列

過学習を防止するために，コンテキストクラスタリング [8] により状態共有を行う．本研究では，スペクトルとF_0は別々のクラスタイリングを行った．さらに，時間方向の変動をより柔軟にモデル化するために，図 4 の右図のように時間方向の状態共有も行った．

（3）　コンテキスト依存 HMM の結合

コンテキスト情報を隠れ変数と見なすために，全てのコンテキスト依存 HMM を結合して 1 つの巨大な HMM を作成する．その際，モデルの最終状態から開始状態に向けて図 5 の右図のようにコンテキスト情報を考慮し結合する．

（4）　モデルの再学習

結合したモデルを再学習することによりコンテキスト情報を隠れ要素とする HMM を作成する．しかし，結合した HMM は非常に巨大な状態遷移のネットワークを保持しており連結学習が困難である．そこで，図 6 の右図のように同一のクラスタに属する状態を共有することにより，モデル学習が容易となる．ただしこの場合，状態を共有しモデルを最小化することにより実際には存在しなかった状態パスを含むモデルとなることに注意する．

5. 声質変換実験

5.1 実験条件

提案法の有効性を示すために，声質変換実験を行った．表 1 に実験条件を示す．また，本実験で使用したモデルの説明を以下にまとめる．

- GMM：スペクトルは GMM を用いた非線形変換，F_0は線形変換
- MSD-GMM：スペクトル・F_0は MSD-GMM を用いた非線形変換
- MSD-HMM1：スペクトル・F_0の MSD-HMM を用いた非線形変換，モデルの最小化なし．ただし，モデル結合後の連結学習が困難であったため，前節の（3）までのモデルである
- MSD-HMM2：スペクトル・F_0の MSD-HMM を用いた非線形変換，モデルの最小化あり

5.2 客観評価実験

メルケプストラムの変換精度の客観評価として，目標話者のメルケプストラムと変換されたメルケプストラムの間でメルケプストラム歪みを求めた．この値が小さいほど話者性を再現し

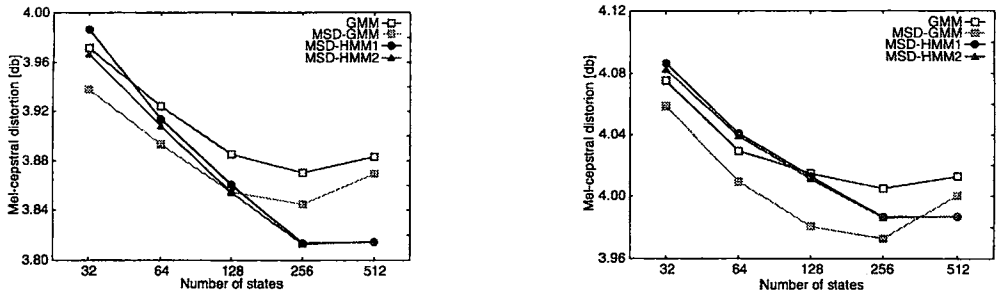

図 7　メルケプストラム歪み (左図：mtk → mht，右図：mho → myi)

Fig. 7　Mel-cepstral distorion (left : mtk → mht, right : mho → myi)

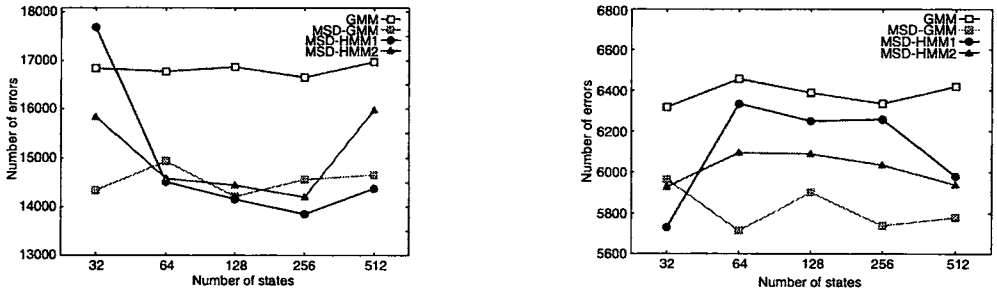

図 8　有声/無声誤り (左図：mtk → mht，右図：mho → myi)

Fig. 8　Number of errors (left : mtk → mht, right : mho → myi)

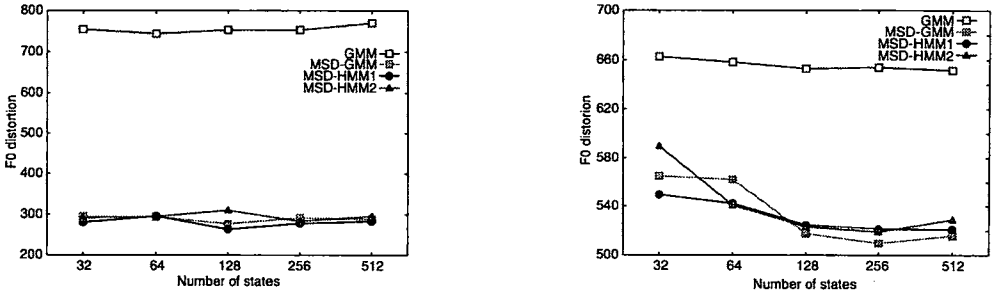

図 9　F_0歪み (左図：mtk → mht，右図：mho → myi)

Fig. 9　F_0 distortion (left : mtk → mht, right : mho → myi)

た変換といえる．図 7 にメルケプストラム歪みの結果を示す．

さらに，F_0の変換精度を調べるために，客観評価として有声/無声誤りとF_0歪みを求めた．有声/無声誤りは，目標話者のF_0と変換後のF_0でアライメントをとり有声/無声が一致しないフレームの数を数えた．ただし，アライメントはスペクトルの歪みに基づいて DP マッチングにより求めた．F_0歪みは，アライメント後のF_0において共に有声となたフレームを抽出し，2 乗誤差を求めた．図 8, 9 にそれぞれ有声/無声の誤り，F_0歪みの結果を示す．

図 7 より，メルケプストラム歪みでは，どちらの変換話者においても，モデルの大きさによっては提案法が従来法よりも小さな値となっている．しかし，これらの差は非常に小さく誤差の範囲と考えられ，必ずしも有効性があるとはいえない．ただし，提案法ではスペクトル・F_0 を同時にモデル化しているにも関わらず，従来法と同等の歪みが得られていることから，少なくとも本実験ではF_0の悪影響を受けることなくスペクトルのモデル化ができていることがわかる．また，提案法の MSD-GMM と MSD-HMM では大きな歪みの差が得られなかった．さらに，混合数，状態数を 512 と増やすと 256 のモデルと比較して歪みが悪化していることから，過学習が発生しているものと考えられる．

図 8 より有声/無声誤りにおいては，提案法では従来法と比較して改善がみられる．これは MSD モデルでは有声から無声や無声から有声の変換が可能となるが，それが有効に働いたためと見られる．ただし，MSD-HMM1 の 32 状態のモデルでは誤りの数が非常に多くなっている．この理由として，たまたま

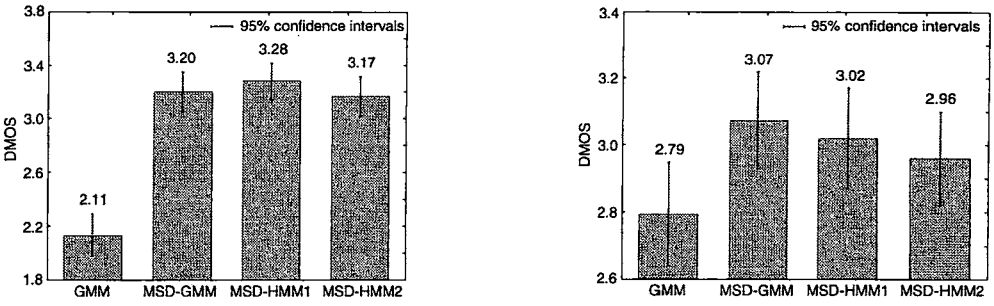

図 10 DMOS (左図 : mtk → mht, 右図 : mho → myi)

Fig. 10 Subjective evaluation (left : mtk → mht, right : mho → myi)

発生した学習精度が低い状態が，HMM に拡張したことによるパスの制約と，モデルサイズが小さいことから頻繁に変換に用いられたためと考えられる．状態数の増加にともない，誤りの大幅な改善が見られるのは，学習精度が低い状態が用いられる頻度が低下したためと考えられる．

図 9 よりF_0歪みでは，mtk から mht への変換において提案法では従来法と比較して非常に大きな改善が見られる．これは，提案法でのF_0の非線形変換が大きな要因と考えられる．また，mho から myi への変換においても mtk から mht への変換ほどではないが歪みの改善が見られた．ただし，どちらの変換話者においても MSD-GMM と MSD-HMM では大きな違いが出なかった．さらに，状態数を増加させた場合にも歪みの改善が見られなかった．これは，音素コンテキストを考慮した HMM のトポロジーがF_0の時間変動を十分にモデル化できていないことを示している．この原因として，本研究で使用した音素コンテキスト (トライフォン) がF_0の変換には不十分であった可能性が考えられる．

5.3 主観評価実験

主観評価実験として，話者性を判断する DMOS テストを行った．被験者は 10 人，各人 15 文章を評価させた．実験に用いたモデルは混合数が 256 の GMM, MSD-GMM，分布数が 256 の MSD-HMM1, MSD-HMM2 である．実験結果を図 10 に示す．

図 10 より，提案法では従来法と比較して大きな改善が見られた．これは，提案法で非線形変換によるF_0 の変換と MSD による有声・無声間の変換が効果的に働いたためと考えられる．提案法である MSD-GMM と MSD-HMM を比較すると客観評価と同様に HMM への拡張による効果は見られなかった．また，MSD-HMM1 と MSD-HMM2 を比較すると MSD-HMM2 の方が劣る結果となった．これは，モデルの最小化によってコンテキスト依存モデルとしては存在しない状態パスを許したことが原因と考えられる．

6. むすび

本報告では，声質変換において MSD モデルを使用し，スペクトル・F_0の同時モデリングを提案した．同時モデリングによりF_0の非線形変換が可能となり，また，有声・無声間の変換が可能となった．客観評価においては，提案法のF_0の有声/無声誤り，F_0歪みの減少などで有効性が示された．また，主観評価においても，提案法は従来法と比較して高い話者性の再現が可能であることがわかった．しかし，MSD-HMM を作成の際に使用した音素コンテキスト (トライフォン) ではF_0を十分にモデル化できず，MSD-GMM と比較して評価値の改善は見られなかった．より効率的なF_0のモデル化のためには，音声合成で用いられている韻律に関するコンテキストを追加する必要があると考えられる．今後の課題として，F_0をモデル化する際の最適なコンテキストの調査や少量の学習データでの有効性の検証などが挙げられる．

謝辞 この研究の一部は，文部科学省リーディングプロジェクト「e-Society 基盤ソフトウェアの総合開発」によって行われたものである．

文　献

[1] 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.

[2] Yannis Stylianou, Olivier Cappe, Eric Moulines, "Continuous Probabilistic Transform for Voice Conversion," *Proc. of IEEE Trans. Speech Audio Process., vol.6*, pp.131–142, Mar. 1998.

[3] Taoufik En-Najjary, Olivier Rosec, Thierry Chonavel, "A voice conversion method based on joint pitch and spectral envelope transformation," *Proc.of Interspeech*, pp.1225–1228, Oct. 2004.

[4] 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.J83-D-II, no.7, pp.1579–1589, Jul. 2000.

[5] Tomoki Toda, Alan W Black, Keiichi Tokuda, "Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis," *Proc.of ISCA Speech Synthesis Workshop*, pp.31–36, Jun. 2004.

[6] Tomoki Toda, Alan W Black, Keiichi Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," *IEEE Trans. Audio Speech Language Process., vol.15*, pp.2222–2235, Nov. 2007.

[7] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, Tadasi Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *Proc.of ICASSP, vol.3*, pp.1315–1318, Jun. 2000.

[8] J.J. Odell, "The Use of Context in Large Vocabulary Speech Recognition," PhD dissertation, Cambridge University, 1995.