Vol.2016-SLP-112 No.4
2016/7/28

時系列構造を考慮した 行列変量混合正規分布モデルによる声質変換

内田 秀継[1,a)]　楊 奕[1,b)]　齋藤 大輔[1,c)]　峯松 信明[1,d)]

概要：本稿では、行列変量混合正規分布モデル (MV-GMM) を用いた声質変換における、特徴量の時系列特性のモデル化法について報告する。声質変換では、混合正規分布モデル (GMM) を用いた手法が、その扱い易さと拡張性の高さから広く用いられている。GMM 声質変換では、入力話者と出力話者の音声特徴量を結合ベクトルで表現し、その確率分布をモデル化する。その際に、当該フレームの前後フレームから導出される動的特徴量を各話者の特徴量にさらに結合することで、変換時に特徴量の時間方向の関係性を考慮することが可能となり、変換性能が向上する。MV-GMM を用いた声質変換では、入力話者と出力話者の特徴量を結合行列として表現することで、特徴量空間と話者空間を明示的に分離でき、それぞれの空間を適切にモデル化することができる。そこで、本稿では、MV-GMM 声質変換において特徴量の時間方向の関係性を考慮するために、時間的に連続した複数のフレームの特徴量を話者空間に沿って連結した結合行列を用いたモデル構築法と変換法を提案した。実験の結果、客観評価と主観評価のどちらにおいても、その有効性が示された。

キーワード：声質変換, 混合正規分布モデル, 行列変量混合正規分布モデル, 時系列構造

Modeling temporal structure in speech for MV-GMM based voice conversion

Hidetsugu Uchida[1,a)]　Yi Yang[1,b)]　Daisuke Saito[1,c)]　Nobuaki Minematsu[1,d)]

1. はじめに

声質変換 (Voice Conversion: VC) は、音声の言語的情報を維持したまま、音声の特性、例えば声の印象や明瞭度などを変換する技術の総称である。その中でも、音声の話者性に着目したものを話者変換と呼ぶ。話者変換では、変換の対象となる話者（入力話者）と変換の目標となる話者（出力話者）の同一文読み上げ音声（パラレル音声コーパス）を用意し、それを用いて各話者空間を対応づける，統計的変換モデルを構築する。統計的変換モデルとしては、混合正規分布モデル (Gaussian Mixture Model: GMM)[1] やニューラルネットワーク (Neural Network: NN)[2]、非負値行列因子分解 (Nonnegative Matrix Factorization: NMF)[3] などが検討されている。

GMM 声質変換は、その柔軟性の高さから様々な拡張が行われてきた。その一例が動的特徴量を用いた系列変換である [4]。この手法では、当該フレームの前後フレームの特徴量から計算される動的特徴量を、静的特徴量と合わせてモデル化する。変換時には、入力特徴量に対して、動的特徴量を考慮した上での最尤な出力特徴の系列を出力する。音声は時間方向の相関が高く、その時系列特性を考慮するこの手法は、変換音声の品質を大幅に改善する。

一般に GMM 声質変換では、パラレル音声コーパスから抽出した入力話者と出力話者のそれぞれの音声特徴量を連結し、一つの長いベクトル（結合ベクトル）として表現し、その確率分布を GMM によってモデル化する。動的特徴量

[1] 東京大学　東京都文京区本郷 7-3-1
[a)] uchida@gavo.t.u-tokyo.ac.jp
[b)] yang@gavo.t.u-tokyo.ac.jp
[c)] dsk_saito@gavo.t.u-tokyo.ac.jp
[d)] mine@gavo.t.u-tokyo.ac.jp

を用いる場合は、各話者の静的特徴量に動的特徴量を連結した上で、さらに各話者の特徴量を連結した結合ベクトルをモデル化する。このモデル化では、静的特徴量に関する情報、動的特徴量に関する情報、各話者の関係性に関する情報が、結合ベクトルによって表現される一つの空間の中で混在した状態で扱われる。より正確に変換関係をモデル化するためには、それらの情報を適切に扱う必要がある。

行列変量混合正規分布 (Matrix Variate Gaussian Mixture Model : MV-GMM) を用いた声質変換では、各話者の特徴量をベクトルではなく行列として結合し（結合行列)、その確率分布をモデル化する [5]。結合行列は、行方向に特徴量の次元が並び、列方向に話者が並ぶような行列である。この結合行列の確率分布を MV-GMM によってモデル化する場合、その分布は、二つの分散行列を持つことになる。二つの分散行列は、結合行列の行方向と列方向のそれぞれの分散構造をモデル化したものである。つまり、特徴量空間の分散構造と話者空間の分散構造をそれぞれ独立に表現している。

GMM と MV-GMM は、どちらも特徴量を正規分布でモデル化するものであるため、GMM 声質変換で検討されてきた拡張は、MV-GMM 声質変換でも有効であると考えられる。そこで、本稿では、GMM 声質変換における動的特徴量を用いた系列変換を、MV-GMM 声質変換に導入した変換法を提案する。提案手法では、各話者の特徴量について、時間的に連続した複数のフレームを結合行列として連結し、さらにその結合行列を入力話者と出力話者について連結した結合行列を MV-GMM によってモデル化する。このモデル化によって、特徴量の時間方向の特性は、特徴量空間とは分離された分散共分散行列によって表現されるため、GMM 声質変換における結合ベクトルを用いた手法よりも適切に時間方向の特性がモデル化できると考えられる。

2. GMM 声質変換

本節では、結合ベクトルを用いた GMM 声質変換について述べる [1]。入力話者と出力話者の音声から抽出される音声特徴量をそれぞれ $\boldsymbol{x}_t = [x_1, x_2, \ldots, x_D]^\top$, $\boldsymbol{y}_t = [y_1, y_2, \ldots, y_D]^\top$ とする。 ここで、t は特徴量の時間インデックス、D は特徴量の次元数、 $(\cdot)^\top$ は転置を表す。このとき、同じ時間インデックスにおける音声特徴量は、同じ言語的特徴を持つ。結合ベクトル $\boldsymbol{z}_t = [\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top]^\top$ の確率分布は、GMM によって以下のようにモデル化される。

$$P\left(\boldsymbol{z}_t|\boldsymbol{\lambda}^{(z)}\right) = \sum_{m=1}^{M} w_m \mathcal{N}\left(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}\right) \tag{1}$$

ここで、 w_m, $\boldsymbol{\mu}_m^{(z)}$,$\boldsymbol{\Sigma}_m^{(z)}$ はそれぞれ m 番目の混合成分の正規分布における混合重み、平均ベクトル、分散共分散行列である。$\boldsymbol{\mu}_m^{(z)}$, $\boldsymbol{\Sigma}_m^{(z)}$ の各要素は、各話者の特徴量の平均ベクトルと分散行列を用いて以下のように表される。

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix} \tag{2}$$

$\boldsymbol{\Sigma}_m^{(\cdot)}$ は、入力話者と出力話者の分散行列、および共分散行列を表す。一般にこれらの分散行列・共分散行列は、過学習を避けるために対角行列でモデル化されることが多い。

入力特徴量 $\boldsymbol{x}_t$ を出力特徴量 $\boldsymbol{y}_t$ に変換するマッピング関数は、$\boldsymbol{x}_t$ に対する $\boldsymbol{y}_t$ の条件付き確率に基づき導出される。$\boldsymbol{y}_t$ の条件付き確率は、結合確率分布のパラメータを用いて以下のように表される。

$$P\left(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) = \sum_{m=1}^{M} P\left(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) P\left(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}\right) \tag{3}$$

ここで、

$$P\left(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) = \frac{w_m \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)}\right)}{\sum_{m=1}^{M} w_m \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)}\right)} \tag{4}$$

$$P\left(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}\right) = \mathcal{N}\left(\boldsymbol{y}_t; \boldsymbol{E}_{m,t}^{(y)}, \boldsymbol{D}_{m,t}^{(y)}\right) \tag{5}$$

$$\boldsymbol{E}_{m,t}^{(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} (\boldsymbol{x_t} - \boldsymbol{\mu}_m^{(x)}) \tag{6}$$

$$\boldsymbol{D}_m^{(y)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} \boldsymbol{\Sigma}_m^{(xy)} \tag{7}$$

$\boldsymbol{x}_t$ が与えられたときの、最尤基準に基づく出力特徴量 $\hat{\boldsymbol{y}}_t$ は、以下の通りである。

$$\hat{\boldsymbol{y}}_t = \left(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(y)-1}\right)^{-1} \left(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(y)-1} \boldsymbol{E}_{m,t}^{(y)}\right)$$
$$\gamma_{m,t} = P\left(m|\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{\lambda}^{(z)}\right) \tag{8}$$

GMM 声質変換において、特徴量の時間方向の特性をする手法が検討されている [4]。特徴量の当該フレームの前後のフレームから計算される動的特徴量を $\Delta\boldsymbol{x}_t$, $\Delta\boldsymbol{y}_t$ とし、静的特徴量と結合したベクトル $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$, $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ を考える。この結合ベクトルは、特徴量の次元数の 2 倍の次元数 ($2D$ 次元) を持ち、特徴量の各フレームにおける静的な特性と動的な特性を記述した特徴量ベクトルとなる。 入力特徴量と出力特徴量の時系列データを、それぞれ $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \ldots, \boldsymbol{x}_T^\top]^\top$, $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, \ldots, \boldsymbol{y}_T^\top]^\top$ としたとき、動的特徴量を考慮した場合の時系列データは、$\boldsymbol{X} = [\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \ldots, \boldsymbol{X}_T^\top]^\top$, $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top \boldsymbol{Y}_2^\top, \ldots, \boldsymbol{Y}_T^\top]^\top$ となる。各フレーム毎の $\boldsymbol{X}_t$ と $\boldsymbol{Y}_t$ の結合ベクトル $\boldsymbol{Z}_t = [\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top$ は、静的特徴量の結合ベクトル $\boldsymbol{z}_t$ と同様に GMM によってモデル化することができる。このとき、入力特徴量の時系列データ $\boldsymbol{X}$ が与えられたときの出力特徴量の時系列データ $\boldsymbol{Y}$ の条件付き確率は、

$$P\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}\right) = \sum_{m=1}^{M} P\left(m|\boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}\right) P\left(\boldsymbol{Y}|\boldsymbol{X}, m, \boldsymbol{\lambda}^{(Z)}\right)$$
$$= \prod_{t=1}^{T} \sum_{m=1}^{M} P\left(m|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(Z)}\right) P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(Z)}\right) \tag{9}$$

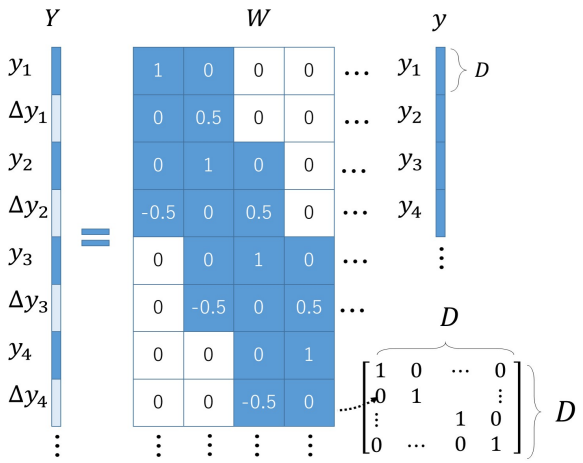

図 1 $\Delta \boldsymbol{y}_t = 0.5(\boldsymbol{y}_{t+1} - \boldsymbol{y}_{t-1})$ における $\boldsymbol{y}_t$ と $\boldsymbol{Y}_t$ の関係
Fig. 1 Relation between $\boldsymbol{y}_t$ and $\boldsymbol{Y}_t$. $\Delta \boldsymbol{y}_t$ is defined as $0.5(\boldsymbol{y}_{t+1} - \boldsymbol{y}_{t-1})$.

となる。ここで、

$$P\left(m|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(Z)}\right) = \frac{w_m \mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_m^{(X)}, \boldsymbol{\Sigma}_m^{(XX)}\right)}{\sum_{m=1}^{M} w_m \mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_m^{(X)}, \boldsymbol{\Sigma}_m^{(XX)}\right)} \quad (10)$$

$$P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(Z)}\right) = \mathcal{N}\left(\boldsymbol{Y}_t; \boldsymbol{E}_{m,t}^{(Y)}, \boldsymbol{D}_m^{(Y)}\right) \quad (11)$$

$$\boldsymbol{E}_{m,t}^{(Y)} = \boldsymbol{\mu}_m^{(Y)} + \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)-1} (\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(X)}) \quad (12)$$

$$\boldsymbol{D}_m^{(Y)} = \boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)-1} \boldsymbol{\Sigma}_m^{(XY)} \quad (13)$$

である。変換後の特徴量系列 $\hat{\boldsymbol{y}}$ は、以下の最適化問題を解くことによって求まる。

$$\hat{\boldsymbol{y}} = \arg\max P\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}\right) \text{ s.t. } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \quad (14)$$

図 1 に、$\boldsymbol{W}$ の例を示す。 この例では、動的特徴量 $\Delta \boldsymbol{y}_t$ は、$0.5(\boldsymbol{y}_{t+1} - \boldsymbol{y}_{t-1})$ によって定義され、それを満たすように窓 $\boldsymbol{W}$ が選ばれる。マッピング関数は、最尤推定に基づいた以下の式によって表される。

$$\hat{\boldsymbol{y}} = \left(\boldsymbol{W}^\top \overline{\boldsymbol{D}^{(Y)-1}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \overline{\boldsymbol{D}^{(Y)-1} \boldsymbol{E}^{(Y)}} \quad (15)$$

ここで、$\overline{\boldsymbol{D}^{(Y)-1}}, \overline{\boldsymbol{D}^{(Y)-1}\boldsymbol{E}^{(Y)}}$ は、時系列に対するパラメータであり、以下で表されるものである。

$$\overline{\boldsymbol{D}^{(Y)-1}} = \mathrm{diag}[\overline{\boldsymbol{D}_1^{(Y)-1}}, \overline{\boldsymbol{D}_2^{(Y)-1}}, \ldots, \overline{\boldsymbol{D}_t^{(Y)-1}}, \ldots, \overline{\boldsymbol{D}_T^{(Y)-1}}] \quad (16)$$

$$\overline{\boldsymbol{D}^{(Y)-1}\boldsymbol{E}^{(Y)}} = [\overline{\boldsymbol{D}_1^{(Y)-1}\boldsymbol{E}^{(Y)}}^\top, \overline{\boldsymbol{D}_2^{(Y)-1}\boldsymbol{E}^{(Y)}}^\top, \ldots, \overline{\boldsymbol{D}_t^{(Y)-1}\boldsymbol{E}^{(Y)}}^\top, \ldots, \overline{\boldsymbol{D}_T^{(Y)-1}\boldsymbol{E}^{(Y)}}^\top]^\top \quad (17)$$

$$\overline{\boldsymbol{D}_t^{(Y)-1}} = \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(Y)-1} \quad (18)$$

$$\overline{\boldsymbol{D}_t^{(Y)-1}\boldsymbol{E}^{(Y)}} = \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(Y)-1} \boldsymbol{E}_{m,t}^{(Y)} \quad (19)$$

$$\gamma_{m,t} = P\left(m|\boldsymbol{X}_t, \boldsymbol{Y}_t, \boldsymbol{\lambda}^{(z)}\right)$$

式 (15) の最尤推定は、特徴量の隣り合ったフレーム間の関係を考慮したものとなっており静的特徴量のみを考慮した場合に比べ、その変換精度は大きく向上する。動的特徴量は静的特徴量とは異なった特性をもっている。したがって、静的特徴量と動的特徴量の結合ベクトルを用いた特徴量のモデル化は、本来は異なる特性を持った二つの空間を一つの空間で表現していることになる。この点に関しては、さらなる改善の余地があると言える。

3. 行列変量混合正規分布モデルを用いた声質変換

本節では、行列変量混合正規分布モデル (MV-GMM) を用いた声質変換について述べる [5]. 今、$\boldsymbol{X}$ をサイズ $n \times p$ の確率変数としての行列とする。$\boldsymbol{X}$ が正規分布に従うとき、

$$\boldsymbol{X} \sim \mathcal{N}_{mv}(\boldsymbol{X}; \boldsymbol{M}, \boldsymbol{U}, \boldsymbol{V}) \quad (20)$$

と表される。ここで、$\boldsymbol{M}$ は、 正規分布の平均を表すサイズ $n \times p$ の行列、$\boldsymbol{U}$ と $\boldsymbol{V}$ は、それぞれ行方向と列方向の分散構造を表す行列であり、そのサイズは $n \times n$ と $p \times p$ である。 行列変量正規分布を従来のベクトルの確率変数を用いて表した場合、以下のようになる [6]。

$$P\left(\mathrm{vec}(\boldsymbol{X})|\boldsymbol{\lambda}\right) = \mathcal{N}\left(\mathrm{vec}(\boldsymbol{X}); \mathrm{vec}(\boldsymbol{M}), \boldsymbol{V} \otimes \boldsymbol{U}\right) \quad (21)$$

ここで、 vec() は、行列をベクトルに展開する演算子である。また、$\otimes$ は、行列のクロネッカー積を表す。式 (21) の分散項に注目すると、行列変量の正規分布は、分散行列の構造がクロネッカー積によって制約されているという点で、ベクトル変量の正規分布と異なることがわかる。 分散行列を、$\boldsymbol{U}$ と $\boldsymbol{V}$ という二つの分散行列に分離することで、$\boldsymbol{X}$ の行方向の空間と列方向の空間の分散構造をそれぞれ個別にモデル化することができる。これは、ベクトルを確率変数とした場合の分散共分散行列に対してある仮定を置き，二つの行列に分解していることになるが，この操作の妥当性は，タスクに依存する。

行列の正規分布は、ベクトルの正規分布と同様に声質変換に用いることができる。$\boldsymbol{x}_t$ と $\boldsymbol{y}_t$ をそれぞれ入力話者と出力話者の音声特徴量とし、それらのベクトルを並べることで結合行列 $\boldsymbol{Z}_t = [\boldsymbol{x}_t, \boldsymbol{y}_t] \in \mathcal{R}^{D \times S}$ を定義する。ここで、D は特徴量空間の次元数、 S は話者空間の次元数を表す。入力話者 1 名、出力話者 1 名の場合は $S = 2$ である。$\boldsymbol{Z}_t$ の確率分布は、混合モデルを用いて以下のように表される。

$$P\left(\boldsymbol{Z}_t|\boldsymbol{\lambda}^{(Z)}\right) = \sum_{m=1}^{M} w_m \mathcal{N}_{mv}(\boldsymbol{Z}_t; \boldsymbol{M}_m, \boldsymbol{U}_m, \boldsymbol{V}_m) \quad (22)$$

上式の確率分布は、行列変量正規分布の重み付け和である。各混合成分の行列変量正規分布は、 $\boldsymbol{M}_m, \boldsymbol{U}_m$, および $\boldsymbol{V}_m$ の三つの行列をパラメータとしている。$\boldsymbol{M}_m$ は平均行列、

$\boldsymbol{U}_m$ は特徴量空間の分散行列、 $\boldsymbol{V}_m$ は話者空間の分散行列である。これらのパラメータは、以下の EM アルゴリズムによって推定できる。

$$\gamma_{m,t} = \frac{w_m \mathcal{N}_{mv}(\boldsymbol{Z}_t; \boldsymbol{M}_m, \boldsymbol{U}_m, \boldsymbol{V}_m)}{\sum_{m=1}^{M} w_m \mathcal{N}_{mv}(\boldsymbol{Z}_t; \boldsymbol{M}_m, \boldsymbol{U}_m, \boldsymbol{V}_m)} \quad (23)$$

$$\hat{\boldsymbol{M}}_m = \frac{1}{T_m} \sum_{t=1}^{T} \gamma_{m,t} \boldsymbol{Z}_t \quad (24)$$

$$\hat{\boldsymbol{U}}_m = \frac{1}{ST_m} \sum_{t=1}^{T} \gamma_{m,t} (\boldsymbol{Z}_t - \hat{\boldsymbol{M}}_m) \hat{\boldsymbol{V}}_m^{-1} (\boldsymbol{Z}_t - \hat{\boldsymbol{M}}_m)^\top \quad (25)$$

$$\hat{\boldsymbol{V}}_m = \frac{1}{DT_m} \sum_{t=1}^{T} \gamma_{m,t} (\boldsymbol{Z}_t - \hat{\boldsymbol{M}}_m)^\top \hat{\boldsymbol{U}}_m^{-1} (\boldsymbol{Z}_t - \hat{\boldsymbol{M}}_m) \quad (26)$$

$$T_m = \sum_{t=1}^{T} \gamma_{m,t} \quad (27)$$

マッピング関数は、式 (3) と同様に条件付き確率 $P(\boldsymbol{y}_t|\boldsymbol{x}_t)$ に基づき導出される。m 番目の混合成分における条件付き確率は以下の通りである。

$$P\left(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(Z)}\right) = \mathcal{N}\left(\boldsymbol{y}_t; \boldsymbol{E}_{m,t}, \boldsymbol{D}_m\right) \quad (28)$$

$$\boldsymbol{E}_{m.t} = \boldsymbol{\mu}_m^{(y)} + \frac{v_m^{(yx)}}{v_m^{(xx)}} (\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}) \quad (29)$$

$$\boldsymbol{D}_m = \left(v_m^{(yy)} - \frac{v_m^{(yx)} v_m^{(xy)}}{v_m^{(xx)}} \right) \boldsymbol{U}_m \quad (30)$$

ここで、$v_m^{(\cdot)}$ は、$\boldsymbol{V}_m$ の要素を表す。MV-GMM 声質変換では、分散行列を行方向と列方向に分離したことで、式 (25) および式 (26) による効率的な推定が可能になり、簡潔で適切なモデル化が可能になる。

4. 複数フレームの特徴量を用いた MV-GMM 声質変換

本節では、複数フレームの特徴量を用いた MV-GMM 声質変換を提案する。$\boldsymbol{x}_t$ と $\boldsymbol{y}_t$ を、それぞれ時刻 t における入力話者の音声特徴量と出力話者の音声特徴量とする。特徴量の時系列に沿った関係性を考慮するため、時刻 $t - N_x$ から $t + N_x$ における入力話者の音響特徴量 $\boldsymbol{x}_{t-N_x}, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_{t+N_x}$ および時刻 $t - N_y$ から $t + N_y$ における出力話者の音響特徴量 $\boldsymbol{y}_{t-N_y}, \ldots, \boldsymbol{y}_t, \ldots, \boldsymbol{y}_{t+N_y}$ に注目する。これらの特徴量を並べた結合行列 $\boldsymbol{Z}_t$ は、$\boldsymbol{Z}_t = [\boldsymbol{x}_{t-N_x}, \ldots, \boldsymbol{x}_t \ldots \boldsymbol{x}_{t+N_x}, \boldsymbol{y}_{t-N_y} \cdots \boldsymbol{y}_t \cdots \boldsymbol{y}_{t+N_y}] \in \mathcal{R}^{D \times S}$ となる。ここで、D は特徴量の次元数、S は各話者において考慮するフレーム数の合計となる。$\boldsymbol{Z}_t$ の確率分布は、式 (22) の MV-GMM と同様の形式で表される。 このとき、特徴量空間の分散構造を表す $\boldsymbol{U}_m$ は、複数フレームを考慮しない場合と同じ次元数を持ち、以下の式で表される $\boldsymbol{V}_m$ の次元数のみが増加する。

$$\boldsymbol{V}_m = \begin{bmatrix} \boldsymbol{V}_m^{(xx)} & \boldsymbol{V}_m^{(xy)} \\ \boldsymbol{V}_m^{(yx)} & \boldsymbol{V}_m^{(yy)} \end{bmatrix} \quad (31)$$

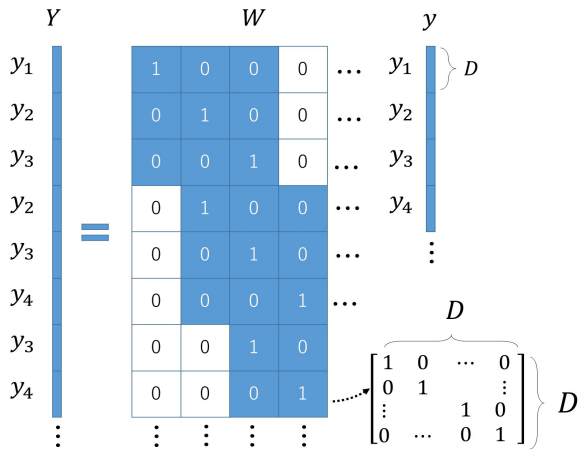

図 2　vec($\boldsymbol{Y}$) と $\boldsymbol{y}$ の関係
Fig. 2　Relation between vec($\boldsymbol{Y}$) and $\boldsymbol{y}$

出力に関して複数フレームを考慮しない場合、つまり、入力話者の特徴量は複数のフレームを用いるが、出力話者の特徴量は単一のフレームのみを用いる場合 ($N_x \neq 0, N_y = 0$) のマッピング関数は、$\boldsymbol{y}_t$ の条件付き確率に基づきフレーム毎の出力を求める式になる。複数フレームの入力特徴量 $\boldsymbol{X}_t = [\boldsymbol{x}_{t-N_x}, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_{t+N_x}]$ が与えられたときの $\boldsymbol{y}_t$ の条件付き確率は以下の通りである。

$$P\left(\boldsymbol{y}_t|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(Z)}\right) = \sum_{m=1}^{M} P\left(m|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(Z)}\right) \times P\left(\boldsymbol{y}_t|\boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(Z)}\right) \quad (32)$$

ここで、

$$P\left(m|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(Z)}\right) = \frac{w_m \mathcal{N}\left(\mathrm{vec}(\boldsymbol{X}_t); \mathrm{vec}(\boldsymbol{M}_m^{(x)}), \boldsymbol{V}_m^{(xx)} \otimes \boldsymbol{U}_m\right)}{\sum_{m=1}^{M} w_m \mathcal{N}\left(\mathrm{vec}(\boldsymbol{X}_t); \mathrm{vec}(\boldsymbol{M}_m^{(x)}), \boldsymbol{V}_m^{(xx)} \otimes \boldsymbol{U}_m\right)} \quad (33)$$

$$P\left(\boldsymbol{y}_t|\boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(Z)}\right) = \mathcal{N}\left(\boldsymbol{y}_t; \boldsymbol{E}_{m,t}^{(y)}, \boldsymbol{D}_m^{(y)}\right) \quad (34)$$

$$\boldsymbol{E}_{m,t}^{(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{V}_m^{(yx)} \boldsymbol{V}_m^{(xx)-1} \left(\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(X)}\right) \quad (35)$$

$$\boldsymbol{D}_m^{(y)} = \left(V_m^{(yy)} - \boldsymbol{V}_m^{(yx)} \boldsymbol{V}_m^{(xx)-1} \boldsymbol{V}_m^{(xy)}\right) \boldsymbol{U}_m \quad (36)$$

である。変換後の特徴量 $\hat{\boldsymbol{y}}_t$ は、式 (8) と同様の形式で求めることができる。

一方、出力に関しても複数フレームを考慮する場合 ($N_x \neq 0, N_y \neq 0$) は、式 (15) で表される最尤系列推定を採用することができる。このとき、マッピング関数は、$\boldsymbol{X} = [\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \ldots, \boldsymbol{X}_T^\top]^\top$ および $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, \ldots, \boldsymbol{Y}_T^\top]^\top$ を用いて、以下の最適化問題として表される。

$$\hat{\boldsymbol{y}} = \arg\max P\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}\right) \text{ s.t. } \mathrm{vec}(\boldsymbol{Y}) = \boldsymbol{W}\boldsymbol{y} \quad (37)$$

ここで、vec($\boldsymbol{Y}$) と $\boldsymbol{y}$ の関係は、図 2 にようになる [7]。式 (32) に基づく変換は、複数フレームを考慮した入力特徴量 $\boldsymbol{X}_t$ から、その時刻 t における出力特徴量 $\hat{\boldsymbol{y}}_t$ が出力される、フレーム単位のマッピングであるに対して、式 (37) に

基づく変換は、入力発話の系列 $\boldsymbol{X}$ に対して、最尤な出力系列 $\hat{\boldsymbol{y}}$ が出力される、系列単位のマッピングである。

動的特徴量を用いた GMM と本稿の提案手法は、どちらも隣接するフレーム間の特徴量の関係を考慮したモデルという点で一致する。提案手法では、静的特徴量と性質が異なる動的特徴量の代わりに、隣接した複数フレームの静的特徴量を用いることで、時系列の特性を捉えることを目指した。この複数フレームを用いたモデル化では、複数のフレームから得られる特徴量が静的特徴量のみで構成される一つの特徴量空間を共有することになる。その点において、異なる性質を持つ静的特徴量と動的特徴量を単一の空間でモデル化する従来手法よりも合理的であると考えられる。

5. 評価実験

5.1 実験条件

提案法の性能を確認するために客観評価実験と主観評実験をおこなった。実験では、パラレル音声コーパスとして ATR 日本語音声データベースを用いた [8]。データベース内の話者 MMY を入力話者、話者 MHT を出力話者とした。250 文のパラレル音声データを学習データ、学習データに用いた文を含まない 50 文を評価データとした。提案手法として、入力特徴量に関してのみ複数フレームを用いた MV-GMM 声質変換 (MVGMM-VC(3 to 1)) と、入力特徴量と出力特徴量の両方に関して複数フレームを用いた MV-GMM 声質変換 (MVGMM-VC(3 to 3)) を評価した。MVGMM-VC(3 to 1) のマッピング関数は、式 (32) に基づいたフレーム単位のマッピング関数、MVGMM-VC(3 to 3) のマッピング関数は、式 (37) に基づいた発話単位のマッピング関数をそれぞれ用いた。また、従来手法として、動的特徴量を用いない GMM 声質変換 (GMM-VC w/o delta) と動的特徴量を用いた GMM 声質変換 (GMM-VC w/ delta)、入力・出力特徴量ともに単一フレームのみを用いた MV-GMM 声質変換 (MVGMM-VC(1 to 1)) を評価した。GMM 声質変換における共分散行列は対角行列とし、MV-GMM 声質変換における共分散行列は全共分散とした。音声特徴量は、STRAIGHT 分析 [9] を用いて抽出した 24 次元のメルケプストラムとした。

5.2 客観評価

メルケプストラム歪み (Mel-cepstral distortion: MCD) を用いて客観評価を行った結果を図 3 に示す。図 3 の縦軸は MCD、横軸は GMM と MV-GMM の混合数である。図 3 より、混合数が 64 を超えると提案手法である MVGMM-VC(3 to 1) と MVGMM-VC(3 to 3) がいずれの従来手法よりも高い変換精度となることがわかる。また、二つの提案手法を比較すると、混合数 64 以上では、MVGMM-VC(3 to 3) が MVGMM-VC(3 to 1) を上回っている。このとき、MVGMM-VC(1 to 1) の変換精度は、MVGMM-VC(3 to 1) を下回っていることから、MV-GMM 声質変換において、入力特徴量の時系列特性を考慮することは変換精度の向上に有効であり、それに加えて出力特徴量の時系列特性も考慮することで変換精度はさらに向上することが明らかになった。静的特徴と動的特徴の結合ベクトルによって出力特徴量の時系列特性を考慮する GMM-VC w/ delta よりも、静的特徴量を並べて出力特徴量の時系列特性を考慮する MVGMM-VC(3 to 3) のほうが高い変換精度を示すことから、分散構造において特徴量の空間と話者・時系列の空間を分離することの有効性、および時系列に沿った複数フレームで静的特徴量の空間を共有してモデル化することの有効性が示唆された。

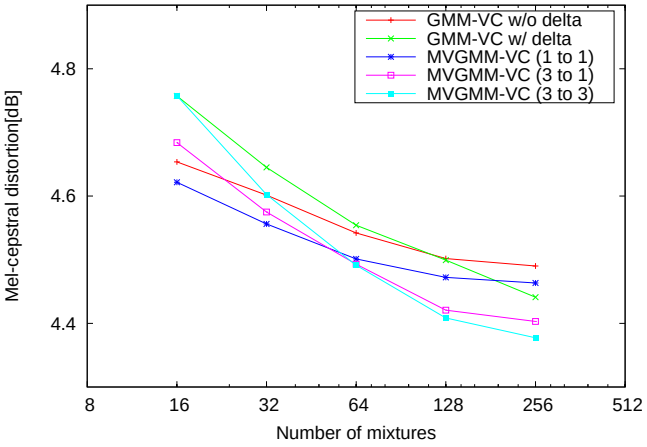

図 3 客観評価結果
Fig. 3 Results of objective evaluations

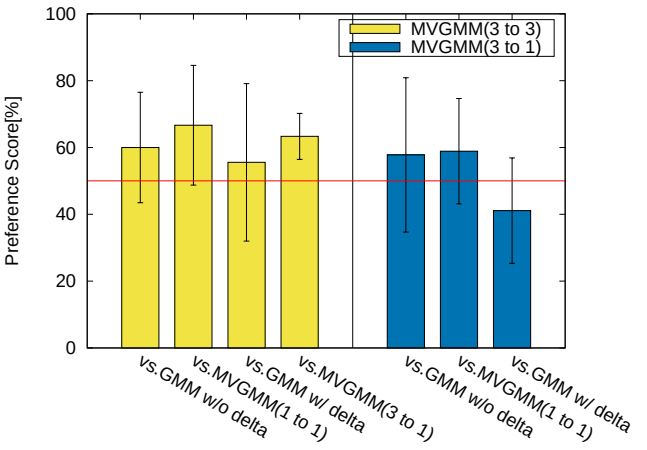

図 4 主観評価結果
Fig. 4 Results of subjective evaluations

5.3 主観評価

XAB 法を用いた主観評価実験を行った。XAB 法では、

二種類の手法によって変換された音声を被験者に提示し（A:手法 1、B:手法 2)、出力話者の音声を参照として聞かせた上で（X:参照音声)、どちらの変換音声のほうが出力話者の音声に近いかを選択させる。被験者は 20 歳から 35 歳の日本語母語話者 10 名である。各被験者は、提案手法–従来手法のペア (6 ペア) と提案手法–提案手法のペア (1 ペア) の計 7 ペアを 1 セットした 10 文、つまり計 70 ペアの変換音声が提示した。音声波形の生成には、STRAIGHT のボコーダを用いた。その際、基本周波数は、$\log F_0$ に対して平均値と分散に基づいた線形変換を行ったものを用いた。また各変換手法の混合数は、いずれも 256 とした。図 4 に主観評価の結果を示す。図中のエラーバーは、95% 信頼区間を表している。 図 4 から、入力・出力特徴量ともに複数フレームを用いた MVGMM-VC(3 to 3) は、他の全ての手法に比べてより出力話者に近い音声であると評価されている。一方で、入力特徴量のみ複数フレームを用いた MVGMM-VC(3 to 1) は、出力特徴量の時系列特性を考慮していない他の手法（GMM-VC w/o delta・MVGMM-VC(1 to 1)）よりも高評価であるが、出力特徴量の時系列特性を考慮した手法（GMM-VC w/ delta・MVGMM-VC(3 to 3)) に比べると低評価であった。MVGMM-VC(3 to 1) は、フレーム単位の変換でり、GMM-VC w/ delta および MVGMM-VC(3 to 3) は、発話単位のマッピングであることから、特徴量の時系列特性を考慮した発話単位のマッピングを行うことが話者性の知覚において重要であると考えられる。

6. おわりに

本稿では、複数フレームから得られる特徴量を用いた MV-GMM 声質変換について検討した。MV-GMM 声質変換は、入力特徴量と出力特徴量の結合行列の確率分布をモデル化することで、特徴量空間と話者空間で独立した分散構造を得ることできるという特徴がある。提案手法では、MV-GMM 声質変換の枠組みで特徴量の時系列特性を考慮することを目指し、隣接したフレームの特徴量を各話者に関して並べた行列の結合行列の確率分布をモデル化する。この提案手法では、各話者および各時刻において共有された特徴量空間と、各話者と各時刻の関係を表現した空間の二つの独立した分散構造を得ることができる。

客観評価と主観評価のいずれにおいても、提案手法である入力特徴量と出力特徴量の両方の複数フレームを用いた MV-GMM 声質変換は、動的特徴量を用いた GMM 声質変換よりも高い評価を示すことが明らかになった。したがって、提案手法が声質変換における特徴量の時系列特性のモデル化に有効であることが示された。

謝辞

本研究は MEXT 科研費 JP26118002 および JSPS 科研費 JP25730105 の助成を受けたものである。

参考文献

[1] A. Kain and M. Macon, “Spectral voice conversion for text-to-speech synthesis,” IEEE International Conference on, pp. 285-288, 1998.

[2] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black and K. Prahallad, “Voice conversion using artificial neural networks,” ICASSP, pp. 3893-3896, 2009.

[3] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, “Exemplar-based voice conversion in noisy environment,” in SLT, pp. 313-317, 2012.

[4] T. Toda, A. W. Black, and K. Tokuda, “Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajctory, ” IEEE TRANSACTIONS 15 (8) pp. 2222-2235, 2007.

[5] D. Saito, H. Doi, N. Minematsu, and K. Hirose, “Application of Matrix Variate Gaussian Mixture Model to Statistical Voice Conversion, ” INTERSPEECH 2014

[6] A. K. Gupta, D. K. Nagar, 『Matrix Variate Distributions』, 2000

[7] L. Chen, Z. Ling, and L. Dai, “Voice Conversion Using Generative Trained Deep Neural Networks with Multiple Frame Spectral Envelopes, ” Interspeech, pp. 2313-2317, 2014.

[8] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol.9, pp.357-363, 1990.

[9] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigne, “Re-structuring speech representations using a pitch-adaptive time- frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, vol.27, pp.187-207, 1999.