社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

統計的音声–調音マッピングにおける声質変換を利用した話者正規化の検討

内田　秀継†　齋藤　大輔†　峯松　信明†　広瀬　啓吉†

† 東京大学大学院 〒 113 − 8656 東京都文京区本郷 7 − 3 − 1

あらまし　本稿では、音声–調音マッピングにおける声質変換を利用した話者正規化法について報告する。調音観測技術の発展による高品質な音声–調音パラレルデータベースの登場と、統計的メディア変換手法の発展を背景に、混合正規分布モデル（Gaussian mixture model : GMM) などを用いた、統計的な音声–調音マッピング法が検討されている。しかし、特定話者の音声–調音パラレルデータから構築された GMM による音声–調音変換モデルは、話者依存モデルとなり、入力音声の発話者と音声–調音パラレルデータに収録されている話者（モデル話者）が異なる場合、変換精度が低下するという問題がある。そこで、本研究では、声質変換を用いて、入力音声をモデル話者音声に近づける話者正規化を行い、音声–調音変換モデルの話者依存性を緩和することによって、変換精度の向上を試みた。入力音声に対して話者変換と音声–調音変換を多段に適用する連結モデルと、話者変換と音声–調音変換の二つの変換モデルを、一つの変換モデルによって実現する分布共有モデルを提案する。その二つの変換モデルに対して入力音声の話者とモデル話者が異なるという条件で音声–調音変換実験を行い、その性能を確かめた。

キーワード　音声–調音マッピング　混合正規分布モデル　話者変換　話者正規化　発音トレーニング

A study of speaker normalization based on voice conversion for statistical acoustic-to-articulatory mapping

Hidetsugu UCHIDA†, Daisuke SAITO†, Nobuaki MINEMATSU†, and Keikichi HIROSE†

† The University of Tokyo, 7 − 3 − 1, Hongo, Bunkyo-ku, Tokyo, 113 − 8656 Japan

Abstract　In this paper, we apply speaker normalization based on voice conversion to acoustic-to-articulatory mapping and report its effectiveness. These days, high-quality acoustic-to-articulatory parallel databases have become available and good efforts have been made for statistical voice conversion study. By combining these two, statistical acoustic-to-articularoty mapping based on Gaussian mixture model (GMM) has drawn researchers' attention. However, GMM-based mapping developed with a single speaker's acoustic-to-articulatory data can be applied only to that specific speaker. If voices of a different speaker is taken as input to the mapping model, the estimated articulatory movements will include large errors. In this study, to solve this problem, we introduce a speaker normalization step before doing acoustic-to-articularoty mapping. Here, the speaker identity of any input speaker will be converted to that of the speaker of the parallel database, i.e., reference speaker. we propose two methods for speaker-normalized acoustic-to-articulatory mapping, one is by concatenating speaker conversion model and acoustic-to-articulatory model and the other integrates the two models into one model, where acoustic observations of a speaker can be converted into articulatory movements of the reference speaker. Experimental evaluation is done for the two methods and their accuracy is discussed.

Key words　acoustic-to-articulatory mapping, GMM, voice conversion, speaker normalization, pronunciation training

1. はじめに

構音障害者や語学学習者のための発音トレーニングにおいて、舌や口唇といった調音器官の運動情報を利用したトレーニングが検討されている [1]。調音情報を利用した発音トレーニングでは、ユーザーの調音状態と正しい発音における調音状態がリアルタイムで提示され、ユーザーは自身の口の中を擬似的に眺めながら、調音運動を矯正することができる。発音の音情報に加

え調音情報という新たな手がかりをユーザーに与えることで、従来の発音トレーニングよりも効果的に発音を改善できる可能性があることが報告されている [2]。また、聴覚障害者など通常の発音トレーニングを行うことが難しいユーザーへの応用も期待できる。

調音情報を利用した発音トレーニングにおいて、ユーザーの調音状態をどのように推定するかが重要な課題となる。調音状態を推定する方法として、磁気センサシステム [3] や電気的口蓋計 [4] などの調音観測システムを用いて、ユーザーの発話中の調音運動を直接測定する方法がある [5]。この方法には、高い精度で計測された調音運動を利用できる利点があるが、測定コストが高く、数多くのユーザーに提供することが難しい。そこで、調音観測システムを用いずに音声から調音運動を推定する手法が検討されている [1]。音声から調音運動を推定する方法として、物理モデルによる推定法とデータベースに基づく変換法が存在する。物理モデルによる推定法では、声道を音響管で近似し、その音響管の共鳴周波数（フォルマント周波数）が，音響的に観測されたそれと等しくなるように声道形状を推定する [6]。この手法は、フォルマント情報から推定を行うので、子音のような明確なフォルマントが存在しない音声の調音運動を推定することが難しく、また、推定されるのが声道形状であるため、舌や口唇などの個別の調音器官の動きを知ることが出来ないという問題がある。

データベースに基づく変換法では、音声–調音パラレルデータと統計的な写像推定手法を用いて、音声と調音運動の関係をモデル化し音声を調音運動に変換する。隠れマルコフモデル (Hidden Markov Model： HMM) や混合正規分布モデル (Gaussian Mixture Model:GMM) などの統計的手法を用いたものや、ディープニューラルネットワーク（Deep Neural Netwark:DNN) を用いたものが検討されている [7] [8] [9]。データベースに基づく変換法は、物理モデルに比べ、子音に対しても比較的良好な変換精度を得ることができ、調音器官の動きを直接推定することが可能である。しかし、これらの手法において、特定の発話者のパラレルデータによって構築された変換モデルは話者依存モデルとなり、モデル話者以外の音声を入力とした場合、変換精度が低下するという問題がある。この問題を避けるため、ユーザー毎に音声–調音パラレルデータを用意し、各ユーザーの話者依存モデルを構築するのは、調音観測の測定コストを考えるとやはり現実的ではない。そこで、特定話者の音声–調音パラレルデータのみを用いて、モデル話者以外の発話者の音声の調音運動を推定可能な変換モデルを構築する手法が必要となる。

統計的手法に基づく写像推定は、声質変換の分野で広く用いられている手法である [10]。特に、任意の音声の言語的情報を保存しながら話者性を発話者から目標話者に変換する技術を話者変換と呼ぶ。そこで、特定話者によって構築された音声ー調音変換モデルに任意話者の音声を入力する場合、特定話者の音声に近づける話者正規化を事前に行うことによって、音声–調音変換モデルの話者依存性の影響を軽減し、任意話者の音声に対する変換精度の向上が期待できる。本稿では、話者変換と音声–調音変換を GMM という同じ枠組みの中で統合することで、任意の話者の入力音声に対して、音声–調音変換を行うことが可能な手法を検討する。

2. 話者変換を利用した音声-調音変換モデルの話者正規化法

2.1 GMM による特徴量変換

まず、本研究で用いる統計的変換手法である、GMM による特徴量変換について説明する。GMM による変換モデルは、変換元となる特徴量と変換先となる特徴量のパラレルデータから構築される。変換元の特徴量を $\boldsymbol{x} = [x(1), x(2), \cdots, x(d_x)]^\top$、変換先の特徴量を $\boldsymbol{y} = [y(1), y(2), \cdots, y(d_y)]^\top$ とし、二つの特徴量を合わせた結合ベクトル $\boldsymbol{z} = [\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$ を定義する。ここで、d_x, d_y は、それぞれの特徴量の次元数である。そして、結合ベクトル $\boldsymbol{z}$ を以下の GMM によってモデル化する。

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

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

ここで、$\boldsymbol{\lambda}$ はモデルパラメータ、$\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は、平均ベクトル $\boldsymbol{\mu}$、分散共分散行列 $\boldsymbol{\Sigma}$ で表される正規分布である。また、M は GMM の混合数、α は各混合成分の重みパラメータである。この GMM は、変換元の特徴量と変換先の特徴量の関係を、多数の正規分布の足し合わせによって表現したものである。この変換モデルを用いて、特徴量変換を行う場合は、以下ような基準に基づきパラメータを得る。

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\lambda}^{(z)}) \tag{3}$$

この式は、変換元の特徴量が入力ベクトルとして与えられた場合、変換モデルの尤度を最大とする出力ベクトルを変換結果とすることを表している。

GMM による音声–調音変換モデルは、調音観測システムによって音声と調音運動が同期観測された音声–調音変換パラレルデータから構築される。特定話者（以下モデル話者）のパラレルデータから変換モデルを構築した場合、その変換モデルは、あくまでモデル話者の音声をモデル話者の調音運動に変換する話者依存モデルである。つまり、モデル話者の音声–調音変換モデルに、モデル話者以外の音声を入力した場合、その音声とモデル構築に用いたモデル話者の音声との間に、音響的特徴の不一致が生じるため変換精度が低下する。そこで、本研究では、音声の話者間の差異を取り除くために、GMM による話者正規化法を導入する。

話者変換は、入力音声の音響的特徴を所望の話者の音声の特徴に変換する技術である。つまり、任意話者の音声を入力とする音声–調音変換の前処理として、話者変換を用いて入力音声をモデル話者音声に変換することができれば（話者正規化)、入力音声とモデル話者音声の差異が取り除かれ、話者依存性は緩和されると考えられる。このとき得られる変換音声は、入力音声の時間構造を持ったモデル話者の音声、言わばモデル話者が入力音声の話速で発話した音声である。そして、その変換音声を音声–調音変換によって調音運動に変換する。つまり、この手法で求めることが出来る調音運動は、モデル話者が入力音声の話速で発話した場合の、モデル話者自身の調音運動となる。例えば語学学習のコンテキストで言えば，教師が学習者の不適切な発音を模擬し，学習者に呈示することに相当する。学習者に示されるのは学習者自らの調音運動ではなく，教師による模擬運動である。このような形式の調音運動推定でも実用的には十

分利用価値はあると考える。

2.2 話者変換と音声ー調音変換の統合

本研究では、a) モデル話者の音声–調音パラレルデータ及び、b) 任意の入力話者とモデル話者との音声のパラレルデータが存在する条件のもと、話者変換と音声–調音変換の統合を行う。本稿では、1) モデル話者への話者変換と音声–調音変換を多段で適用し連結する変換手法（連結モデル）と、2) モデル話者の特徴量空間の確率分布を共有する事で、任意の入力話者の音声を調音運動に直接変換する手法（分布共有モデル）の 2 つについて検討する。

2.3 連結モデル

話者変換と音声-調音変換を多段で適用し連結する変換手法では、音声-調音変換と話者変換を別々の変換モデルとしてそれぞれ構築する。今、入力話者の音声特徴量を $\boldsymbol{x}^{(s)}$、モデル話者の音声特徴量及び調音運動特徴量を $\boldsymbol{y}^{(s)}$、$\boldsymbol{y}^{(a)}$ とし、パラレルデータから構成される結合ベクトル $\boldsymbol{z}^{(xy)} = [\boldsymbol{x}^{(s)\top}, \boldsymbol{y}^{(s)\top}]^\top$ および $\boldsymbol{z}^{(sa)} = [\boldsymbol{y}^{(s)\top}, \boldsymbol{y}^{(a)\top}]^\top$ の確率密度を以下の二つの混合ガウス分布（GMM）で表す;

$$P(\boldsymbol{z}^{(xy)}; \boldsymbol{\lambda}^{(xy)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{z}^{(xy)}; \boldsymbol{\mu}_m^{(xy)}, \boldsymbol{\Sigma}_m^{(xy)}) \tag{4}$$

$$P(\boldsymbol{z}^{(sa)}; \boldsymbol{\lambda}^{(sa)}) = \sum_{n=1}^{N} \beta_n \mathcal{N}(\boldsymbol{z}^{(sa)}; \boldsymbol{\mu}_n^{(sa)}, \boldsymbol{\Sigma}_n^{(sa)}) \tag{5}$$

$$\boldsymbol{\mu}_m^{(xy)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix} \boldsymbol{\Sigma}_m^{(xy)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(x,x)} & \boldsymbol{\Sigma}_m^{(x,y)} \\ \boldsymbol{\Sigma}_m^{(y,x)} & \boldsymbol{\Sigma}_m^{(y,y)} \end{bmatrix} \tag{6}$$

$$\boldsymbol{\mu}_n^{(sa)} = \begin{bmatrix} \boldsymbol{\mu}_n^{(s)} \\ \boldsymbol{\mu}_n^{(a)} \end{bmatrix} \boldsymbol{\Sigma}_n^{(sa)} = \begin{bmatrix} \boldsymbol{\Sigma}_n^{(s,s)} & \boldsymbol{\Sigma}_n^{(s,a)} \\ \boldsymbol{\Sigma}_n^{(a,s)} & \boldsymbol{\Sigma}_n^{(a,a)} \end{bmatrix} \tag{7}$$

入力話者の音声を調音運動に変換する場合は、まず、入力話者の音声と話者変換モデルを用いて、以下の式に従って話者変換音声 $\hat{\boldsymbol{y}}^{(s)}$ を求める。

$$\hat{\boldsymbol{y}}^{(s)} = \arg\max_{\boldsymbol{y}^{(s)}} P(\boldsymbol{y}^{(s)} | \boldsymbol{x}^{(s)}; \boldsymbol{\lambda}^{(xy)}) \tag{8}$$

この $\hat{\boldsymbol{y}}^{(s)}$ は、入力話者の音声をモデル話者の音声に変換したものである。次に、この変換音声を音声–調音変換モデルの入力として、以下の式に従って調音運動に変換する。

$$\hat{\boldsymbol{y}}^{(a)} = \arg\max_{\boldsymbol{y}^{(a)}} P(\boldsymbol{y}^{(a)} | \hat{\boldsymbol{y}}^{(s)}; \boldsymbol{\lambda}^{(sa)}) \tag{9}$$

この手法では、入力話者の音声に対して、音声変換と音声-調音運動変換という 2 つの変換が多段に作用することになる。

2.4 分布共有モデル

話者変換と音声-調音変換を統合し、一つの変換モデルとして構築することによって、話者変換音声を経由せずに、入力話者の音声を調音運動へと変換する手法を検討する。

連結モデルにおいて、式 (4),(5) で表される 2 つのモデルの確率分布はともに、部分空間としてモデル話者の音声特徴量 $\boldsymbol{y}^{(s)}$ の空間を有する。このとき、これらの部分空間が共通のモデルによって表されていると考える事で（分布共有）、推定すべきパラメータ数を削減することができる [11]。そして、モデル話者の音声特徴量を周辺化することで、入力話者の音声を調音運動に直接変換するモデルを構築することが可能になる。

入力話者の音声特徴量が与えられたときの、調音運動は以下のように表される。

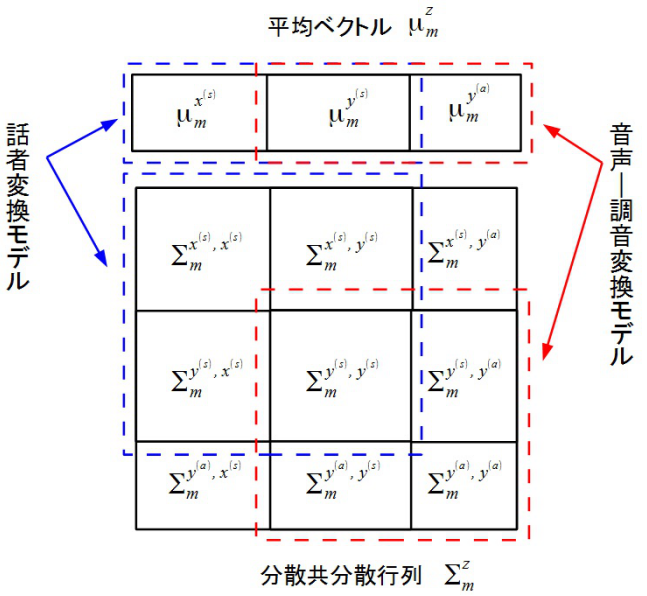

図 1 結合ベクトル $\hat{\boldsymbol{z}}$ から得られるGMMにおける、一つの混合成分の平均ベクトルと分散共分散行列

Fig. 1 Mean vector and covariance matrix of a component of GMM for the joint vector $\hat{\boldsymbol{z}}$

$$\hat{\boldsymbol{y}}^{(a)} = \arg\max_{\boldsymbol{y}^{(a)}} P(\boldsymbol{y}^{(a)} | \boldsymbol{x}^{(s)}; \boldsymbol{\lambda}) \tag{10}$$

上式を、結合ベクトルの混合成分 m 及び、モデル話者の音声特徴量 $\boldsymbol{y}^{(s)}$ について展開すると、

$$\hat{\boldsymbol{y}}^{(a)} = \arg\max_{\boldsymbol{y}^{(a)}} P(\boldsymbol{y}^{(a)} | \boldsymbol{x}^{(s)}) \tag{11}$$

$$\approx \arg\max_{\boldsymbol{y}^{(a)}} \sum_{m=1}^{M} P(m | \boldsymbol{x}^{(s)}) \times \int_{\mathcal{Y}^{(s)}} P(\boldsymbol{y}^{(a)} | \boldsymbol{y}^{(s)}, m) P(\boldsymbol{y}^{(s)} | \boldsymbol{x}^{(s)}, m) d\boldsymbol{y}^{(s)} \tag{12}$$

$$= \arg\max_{\boldsymbol{y}^{(a)}} \sum_{m=1}^{M} P(m | \boldsymbol{x}^{(s)}) \mathcal{N}(\boldsymbol{y}^{(a)} | \boldsymbol{E}_m, \boldsymbol{D}_m) \tag{13}$$

を得る。ここで、

$$\boldsymbol{E}_m = \boldsymbol{\mu}_m^{(a)} + \boldsymbol{\Sigma}'_m (\boldsymbol{x}^{(s)} - \boldsymbol{\mu}_m^{(x)}) \tag{14}$$

$$\boldsymbol{D}_m = \boldsymbol{\Sigma}_m^{(a,a)} - \boldsymbol{\Sigma}'_m \boldsymbol{\Sigma}_m^{(x,x)} \boldsymbol{\Sigma}'^{\,\mathrm{T}}_m \tag{15}$$

$$\boldsymbol{\Sigma}'_m = \boldsymbol{\Sigma}_m^{(a,s)} \boldsymbol{\Sigma}_m^{(y,y)^{-1}} \boldsymbol{\Sigma}_m^{(y,x)} \tag{16}$$

である。式 (13) を変換式とすることで、話者変換音声を明示的には経由せずに入力音声を調音運動に変換することができ、中間のモデルの影響が式 (16) の形で記述されていることがわかる。

2.5 分布共有モデルの構築法

前節で述べた分布共有モデルを実装するためには、話者変換モデルと音声–調音変換モデルの間で共通する特徴量であるモデル話者音声の平均ベクトルと分散共分散行列が、各混合成分おいて一致するように二つの変換モデルを構築する必要がある。しかし、一部のパラメータが一致するように二つのモデルを個別に構築することは困難である。そこで、本稿では、話者変換モデルと音声–調音変換モデルを個別に構築するのではなく、一つの GMM によって構築する方法について検討する。今、入力話者の音声特徴量およびモデル話者の音声特徴量と調音特徴量（三種類の特徴量）の結合ベクト

ル $\hat{\boldsymbol{z}} = [\boldsymbol{x}^{(s)\top}, \boldsymbol{y}^{(s)\top}, \boldsymbol{y}^{(a)\top}]^\top$ を考える。この結合ベクトルの GMM は、$P(\hat{\boldsymbol{z}}; \boldsymbol{\lambda}^{(\hat{z})}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\hat{\boldsymbol{z}}; \boldsymbol{\mu}_{\boldsymbol{m}}^{(\hat{z})}, \boldsymbol{\Sigma}_m^{(\hat{z})})$ で表され、各混合における平均ベクトル及び分散共分散行列は、 図 1 のようになる。図 1 に示すように、この GMM では話者変換を表す部分空間と音声–調音変換を表す部分空間がモデル話者の音声特徴量の部分空間を共有した形で構成されている。

今、a) 入力話者の音声とモデル話者の音声のパラレルデータ、b) モデル話者の音声と調音運動のパラレルデータを用いて、上で述べた三種類の特徴量の結合ベクトルの GMM を構築することを考える。a) と b) のパラレルデータにそれぞれ固有の読み上げテキストが存在するものとする。GMMの構築には、入力話者の音声特徴量およびモデル話者の音声特徴量と調音特徴量（三種類の特徴量）の結合ベクトル ($\hat{\boldsymbol{z}} = [\boldsymbol{x}^{(s)\top}, \boldsymbol{y}^{(s)\top}, \boldsymbol{y}^{(a)\top}]^\top$) が必要となるが、a) のパラレルデータに固有の読み上げテキストに対応するモデル話者の調音運動は観測されていない（欠損値）。同様に、b) のパラレルデータに固有な読み上げテキストに対しては、入力話者の音声が欠損値となる。

つまり、二つのパラレルデータそれぞれに固有な読み上げテキストから得られる結合ベクトル、$\boldsymbol{z}_t^{(xy)} = [\boldsymbol{x}_t^{(s)\top}, \boldsymbol{y}_t^{(s)\top}]^\top$ および、$\boldsymbol{z}_t^{(sa)} = [\boldsymbol{y}_t^{(s)\top}, \boldsymbol{y}_t^{(a)\top}]^\top$ に対して、擬似的な三種類の特徴量の結合ベクトル、$\boldsymbol{z}_t{}^{(xyy')} = [\boldsymbol{x}_t{}^{(s)\top}, \boldsymbol{y}_t{}^{(s)\top}, \boldsymbol{y}'^{(a)\top}_t]^\top (t = 1 \sim T_1)$,$\boldsymbol{z}_t{}^{(s'sa)} = [\boldsymbol{x}'^{(s)\top}_t, \boldsymbol{y}_t{}^{(s)\top}, \boldsymbol{y}_t{}^{(a)\top}]^\top (t = T_2 \sim T_3)$ を考えた場合、$\boldsymbol{y}'^{(a)}, \boldsymbol{x}'^{(s)}$ がそれぞれ欠損値となる。この欠損値を含む二つの結合ベクトルを時間軸方向に並べたベクトル、$\hat{\boldsymbol{z}} = \{\boldsymbol{z}_1^{(xyy')}, \cdots, \boldsymbol{z}_{T_1}^{(xyy')}, \boldsymbol{z}_1^{(s'sy)}, \cdots, \boldsymbol{z}_{T_2}^{(ss'a)}\}$ について、以下のように GMM を構築する。

E ステップ

$$\gamma_{m,t}^{(xyy')} = \frac{\pi_m \mathcal{N}(\boldsymbol{z}_t^{(xyy')} | \boldsymbol{\mu}_m{}^{(\hat{z})}, \boldsymbol{\Sigma_m}^{(\hat{z})})}{\sum_{j=1}^{M} \pi_j \mathcal{N}(\boldsymbol{z}_t^{(xyy')} | \boldsymbol{\mu}_j{}^{(\hat{z})}, \boldsymbol{\Sigma_j}^{(\hat{z})})} \tag{17}$$

$$\gamma_{m,t}^{(s'sa)} = \frac{\pi_m \mathcal{N}(\boldsymbol{z}_t^{(s'sa)} | \boldsymbol{\mu}_m{}^{(\hat{z})}, \boldsymbol{\Sigma_m}^{(\hat{z})})}{\sum_{j=1}^{M} \pi_j \mathcal{N}(\boldsymbol{z}_t^{(s'sa)} | \boldsymbol{\mu}_j{}^{(\hat{z})}, \boldsymbol{\Sigma_j}^{(\hat{z})})} \tag{18}$$

M ステップ

$$\boldsymbol{\mu}_m = \frac{1}{\gamma_m} \left(\sum_{t=1}^{T_1} \gamma_{m,t}{}^{(xyy')} \boldsymbol{z_t}^{(xyy')} + \sum_{t=T_2}^{T_3} \gamma_{m,t}{}^{(s'sa)} \boldsymbol{z_t}^{(s'sa)} \right) \tag{19}$$

$$\Sigma_m = \frac{1}{\gamma_m} \left(\sum_{t=1}^{T_1} \gamma_{m,t}{}^{(xyy')} \left\{ (\boldsymbol{z_t}^{(xyy')} - \boldsymbol{\mu}_m)(\boldsymbol{z_t}^{(xyy')} - \boldsymbol{\mu}_m)^T - \hat{D}_m^{(xyy')} \right\} + \sum_{t=T_2}^{T_3} \gamma_{m,t}{}^{(xyy')} \left\{ (\boldsymbol{z_t}^{(xyy')} - \boldsymbol{\mu}_m)(\boldsymbol{z_t}^{(xyy')} - \boldsymbol{\mu}_m)^T - \hat{D}_m^{(xyy')} \right\} \right) \tag{20}$$

$$\pi_m = \frac{\gamma_m}{\sum_{m=1}^{M} \gamma_m} \tag{21}$$

ここで、$\gamma_m = \Sigma_{t=1}^{T_1} \gamma_{m,t}^{(xyy')} + \Sigma_{t=1}^{T_2} \gamma_{m,t}^{(s'sa)}$ である。

E ステップにおける欠損値 $\boldsymbol{y'}_t{}^{(a)}, \boldsymbol{x'}_t{}^{(s)}$ は、$\boldsymbol{y}_t{}^{(a)}, \boldsymbol{x}_t{}^{(s)}$ の推定値である。各推定値は、測定値の結合ベクトル $\boldsymbol{z}_t^{(xy)}, \boldsymbol{z}_t^{(sa)}$ と GMM の部分空間を用いて以下の式によって求めることが出来る。

$$\boldsymbol{y'}_t{}^{(a)} = \sum_{m=1}^{M} P(\boldsymbol{z}_t^{(xy)} | \boldsymbol{\mu}_m^{(xy)}, \boldsymbol{\Sigma}_m^{(xy)}) E_{m,t}^{(y^{(a)}|x^{(s)}, y^{(s)})} \tag{22}$$

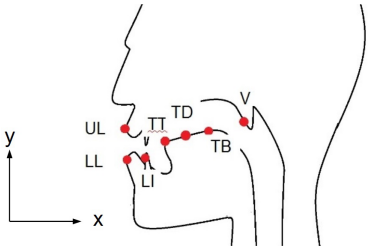

図 2　調音データの測定点

Fig. 2　Measurement points of articulatory data

$$\boldsymbol{x'}_t{}^{(s)} = \sum_{m=1}^{M} P(\boldsymbol{z}_t^{(sa)} | \boldsymbol{\mu}_m^{(sa)}, \boldsymbol{\Sigma}_m^{(sa)}) E_{m,t}^{(x^{(s)}|y^{(s)}, y^{(a)})} \tag{23}$$

ここで、$\boldsymbol{\mu}_m^{(xy)}, \boldsymbol{\Sigma}_m^{(xy)}$ は、結合ベクトル $\boldsymbol{z}^{(xy)}$ の平均ベクトルと分散共分散行列、$\boldsymbol{\mu}_m^{(sa)}, \boldsymbol{\Sigma}_m^{(sa)}$ は、結合ベクトル $\boldsymbol{z}^{(sa)}$ に関する平均ベクトルと分散共分散行列である。また、

$$E_{m,t}^{(y^{(a)}|x^{(s)}, y^{(s)})} = \boldsymbol{\mu}_m^{(a)} + \Sigma_m^{(a,xy)} \Sigma_m^{(xy,xy)^{-1}} (\boldsymbol{z}_t^{(xy)} - \boldsymbol{\mu}_m^{(xy)}) \tag{24}$$

$$E_{m,t}^{(x^{(s)}|y^{(s)}, y^{(a)})} = \boldsymbol{\mu}_m^{(s)} + \Sigma_m^{(x,sa)} \Sigma_m^{(sa,sa)^{-1}} (\boldsymbol{z}_t^{(sa)} - \boldsymbol{\mu}_m^{(sa)}) \tag{25}$$

である。ここで、$\Sigma_m^{(a,xy)}, \Sigma_m^{(x,sa)}$ は、それぞれ $\boldsymbol{y}^{(a)}$ と $\boldsymbol{z}^{(xy)}$、$\boldsymbol{y}^{(x)}$ と $\boldsymbol{z}^{(sa)}$ の分散共分散行列、$\Sigma_m^{(xy,xy)}, \Sigma_m^{(sa,sa)}$ は、それぞれ $\boldsymbol{z}^{(xy)}$,$\boldsymbol{z}^{(sa)}$ の分散共分散行列である。

M ステップにおける欠損値 $\boldsymbol{y'}_t{}^{(a)}, \boldsymbol{x'}_t{}^{(s)}$ は、それぞれ式 (24),(25) で表される平均ベクトルである。また、式 (20) における $\hat{D}_m^{(xyy')}, \hat{D}_m^{(s'sa)}$ は、以下の行列である。

$$\hat{D}_m^{(xyy')} = \begin{bmatrix} \mathbf{0}^{(d_1,d_1)} & \mathbf{0}^{(d_1,d_2)} \\ \mathbf{0}^{(d_2,d_1)} & D_m^{(y^{(a)}|x^{(s)}, y^{(s)})} \end{bmatrix} \tag{26}$$

$$\hat{D}_m^{(s'sa)} = \begin{bmatrix} D_m^{(x^{(s)}|y^{(s)}, y^{(a)})} & \mathbf{0}^{(d_3,d_4)} \\ \mathbf{0}^{(d_4,d_3)} & \mathbf{0}^{(d_4,d_4)} \end{bmatrix} \tag{27}$$

ここで、

$$D_m^{(y^{(a)}|x^{(s)}, y^{(s)})} = \Sigma_m^{(a,a)} - \Sigma_m^{(a,xy)} \Sigma_m^{(xy,xy)^{-1}} \Sigma_m^{(xy,a)} \tag{28}$$

$$D_m^{(x^{(s)}|y^{(s)}, y^{(a)})} = \Sigma_m^{(x,x)} - \Sigma_m^{(x,sa)} \Sigma_m^{(sa,sa)^{-1}} \Sigma_m^{(sa,x)} \tag{29}$$

である。また、$0^{(m,l)}$ は m 行 l 列の零行列である。d_1, d_2, d_3, d_4 は、それぞれ $\boldsymbol{z}^{(xy)}, \boldsymbol{y}^{(a)}, \boldsymbol{z}^{(x)}, \boldsymbol{z}^{(sa)}$ の次元数と等しい値である。

3. 実　　験

3.1 実験条件

本実験では、MOCHA データベース [12] を用いて各変換モデルを構築した。データベースには、英国人の男女各 1 名ずつの音声–調音運動パラレルデータが収録されている。調音運動データは、三次元磁気センサシステムによって計測されたもので、正中矢状面上の下前歯（LI）・上下の口唇（UL・LL）・舌上の三点（TT・TD・TB）・軟口蓋（V）の計 7 点の測定点の運動が収録されている。各測定点の大まかな配置は図 2 に示すとおりである。また、調音データは、それぞれ水平方向（x 軸）と垂直方向 (y 軸）の二次元データである。

データベースの男性話者をモデル話者、女性話者を発話者として連結モデルと分布共有モデルをそれぞれ構築した。このとき、話者変換モデルの音声特徴量として、25 次元の MFCC(パワー成分を含む）と Δ MFCC を合わせた 50 次元の特徴量を用いた。また、音声–調音変換モデルの音声特徴量は、文献 [8] を

参考して、MFCC 系列の当該フレームから前後 5 フレームの特徴量を主成分分析し、得られた主成分の上位 75 成分から構成される 75 次元の特徴量を用いた。連結モデルにおいて入力音声から調音運動への変換を行う際には、話者変換モデルで出力された MFCC を男性話者の主成分に変換した値を音声–調音変換モデルの入力とした。各モデルの構築の際に必要となる男性話者と女性話者の音声特徴量の結合ベクトルは、MFCC 系列に対して DTW (dynamic time warping) を用いて求めた。

また、調音特徴量は、各測定点のフレームごとの座標データに、速度成分を加えた 28 次元の特徴量である。各特長量のフレーム長は 10 ms である。

今回は、比較のために連結モデルと分布共有モデルに加え、男性話者（モデル話者）自身の音声–調音変換モデルを構築した。そして、以下の 4 つ条件において音声–調音マッピング実験を行った。

（1） モデル話者（男性話者）の音声–調音変換モデルに、モデル話者の音声を入力する。(reference)

（2） モデル話者の音声–調音変換モデルに、発話者（女性話者）の音声を入力する。(baseline)

（3） 連結モデルに、発話者（女性話者）の音声を入力する。(連結モデル)

（4） 分布共有モデルに、発話者（女性話者）の音声を入力する (分布共有モデル)

reference 条件は、モデル話者と入力音声の話者が一致する状態、つまり、変換モデルの話者依存性の問題が生じない理想的な状態である。一方、baseline 条件は、変換モデルにモデル話者以外の音声がそのまま入力される状態、つまり、話者依存性の問題が生じる状態である。本実験では、連結モデルと分布共有モデルを用いることで、変換精度が baseline に比べ、同程度改善されるかを検証する。

各モデルの学習と評価は、データベース内の全 460 発話に対して、5-fold cross-validation を用いて行った。

3.2 結　　果

実測された男性話者の調音運動を正解データとして、各条件で得られた調音運動の変換誤差を求めた。しかし、reference 以外の条件で得られる調音運動は、入力音声となる女性話者の音声の時間構造を持つため、男性話者の調音運動と直接比較することは出来ない。そこで、男性話者と女性話者の音声男特徴量の結合ベクトルを求めたときと同様に、MFCC 系列に対する DTW を用いて時間構造を一致させた。変換誤差は、全発話に対する二乗平均平方根誤差（RMSE) とし、各測定点ごとに水平方向および垂直方向について算出した。

各測定点ごとの変換誤差の平均値を図 3 に示す。図 3 の reference に注目すると、舌上の測定点に対する変換誤差が他の測定点と比べて大きく、baseline の変換精度の悪化の度合いも大きい。このような測定点では、二つの提案手法のどちらにおいても baseline と比べ、変換精度が大きく改善されている。また、変換誤差 (RMSE) について F 検定（両側検定、有意水準 95 %) を行ったところ、連結モデルにおいて、全ての測定点で、baseline から有意な誤差の減少が認められた。分布共有モデルについても同様に、全ての測定点で、baseline から有意な誤差の減少が認められた。このことから、話者変換を利用することで変換モデルの話者依存性が緩和されることがわかる。

二つの提案手法を比較した場合、全測定点における変換誤差の平均 (Ave.) において、分布共有モデルが連結モデルよりも、0.25 mm 上回っている。また、F 検定の結果、全ての測定点において、連結モデルに比べ分布共有モデルの変換誤差が有意に

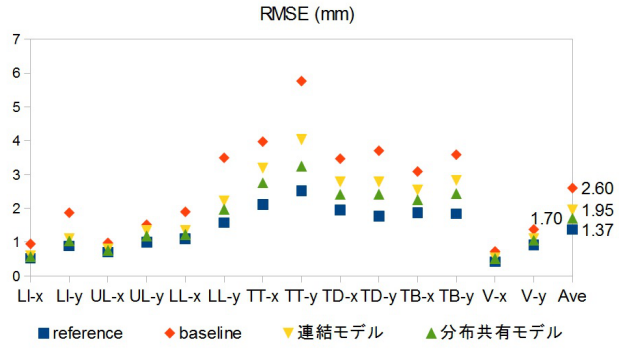

図 3　測定点ごとの変換誤差

Fig. 3　RMSE measured at each observation point

減少していることが認められた。これは、入力音声から話者変換音声を経由せずに調音運動へ直接変換することで、多段の変換による誤差の蓄積が避けられた結果だと考えられる。

また、図 4 に、各音素ごとの変換誤差を示す。ここでの変換誤差は、各音素について、全ての調音点の水平方向・垂直方向の変換誤差を平均したものである。reference 条件に注目すると、子音に比べ、母音半母音は変換精度が安定していることが分かる。音響的エネルギーが大きい母音の方が安定した変換が可能であることが示されている。母音に対する変換精度は連結モデル・分布共有モデルを用いることで、baseline に比べ変換精度が向上することが分かる。一方、子音に注目すると、軟口蓋閉鎖音/k//g/では、連結モデルを用いても変換精度が baseline から改善されていない。しかし、分布共有モデルを用いることで、変換精度が向上し、reference とほぼ変わらない精度で推定可能であることが分かる。F 検定の結果においても、分布共有モデルは/k//g/を含めた全ての音素について、baseline に対して有意な誤差の減少が認められた。

図 5 に、/k//g/に対する調音点ごとの変換誤差を示す。舌上のセンサー、特に後方の 2 つのセンサー（TD,TB) に注目すると、連結モデルは baseline からの改善がほとんど見られないが、分布共有モデルは reference と同等の精度が得られていることがわかる。音素/k/に関して、分布共有モデルに注目して F 検定を行った結果、全ての測定点について、baseline に対して有意な誤差の減少が認められ、連結モデルと比較した場合は、LL-x,TT-y 以外の測定点において有意な誤差の減少が認められた。一方、音素/g/に関して、分布共有モデルに注目して F 検定を行った結果、LI-x,LL-x,TB-x,V-y 以外の測定点について、baseline に対して有意な誤差の減少が認められ、連結モデルと比較した場合は、LI-x,LI-y,LL-x,TT-y 以外の測定点において有意な誤差の減少が認められた。

/k//g/は、どちらも軟口蓋閉鎖音であるため、調音点に近い TD・TB において、分布共有モデルを用いることで良好な変換精度が得られるということは、発音トレーニングなどへの応用を検討する上で非常に重要である。

4. おわりに

本稿では、音声-調音マッピングにおける話者依存性の緩和のために、話者変換を利用した変換モデルを検討した。入力音声を話者変換を用いて、音声–調音変換モデルのモデル話者音声に近づけることによって、入力音声とモデル音声の音響的差異を取り除き、変換モデルの話者依存性の軽減を試みた。連結モデルと分布共有モデルという二つのモデルを提案し、音声–調音変換実験を行った。その結果、提案法を用いることで、変換モデルの話者依存性による変換精度の悪化を軽減できることが示された。また、分布共有モデルは、連結モデルに比べ高い精

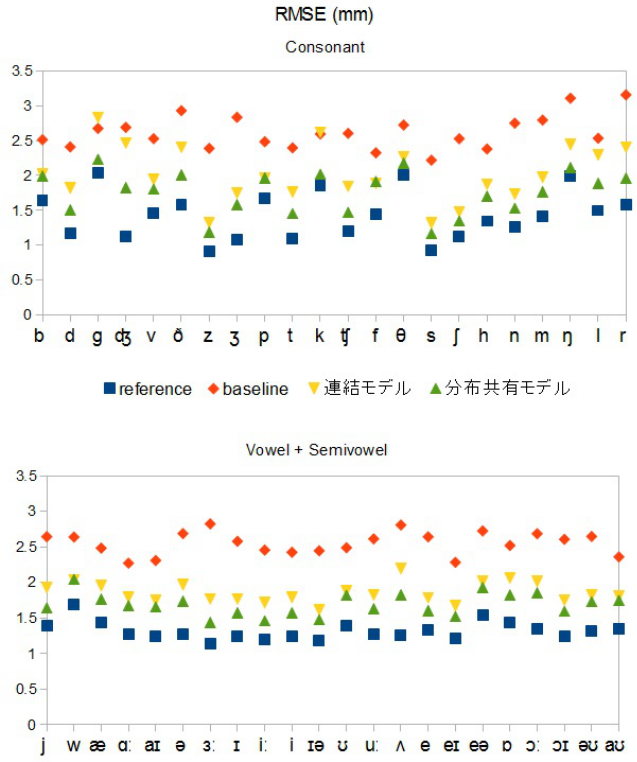

図 4　音素ごとの変換誤差

Fig. 4　RMSE of each phoneme

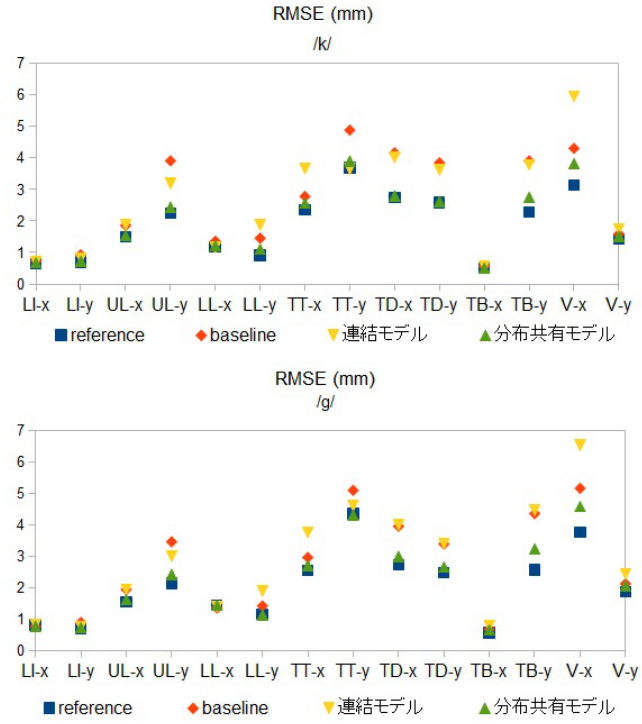

図 5　音素/k/および/g/における測定点ごとの変換誤差

Fig. 5　RMSE of /k/and /g/

度を得られることが明らかになった。

分布共有モデルの構築法として、欠損値を含むＥＭアルゴリズムを提案した。本稿の実験では、同一の読み上げテキストの音声パラレルデータと音声–調音パラレルデータから欠損値を設定し、変換モデルを構築した。これは、実験の妥当性を損なうものではないが、更なる検討として、音声パラレルデータと音声–調音パラレルデータおいて、読み上げテキストが異なる場合の実験を行う必要がある。

提案手法では，入力話者とモデル話者の音声パラレルデータの収録において，両者が同一の調音運動をしていることを前提としている。互いに同一言語を母語としている場合であっても，地方訛りなどの違いによって必ずしも同一の調音運動をしていない可能性もある。両者の調音運動に差が存在するまま音声パラレルデーターとして使った場合、発話者特有の喋りの特徴は話者変換によってモデル話者の喋りの特徴に変換されてしまい、音声–調音変換の結果として得られる調音運動は、発話者の特徴を反映したものにはならないと考えられる。発音トレーニングへの応用を考えた場合、推定するべき調音運動は発話者自身の喋りの特徴を十分に反映した調音運動である。したがって、本研究の枠組みにおいて、発話者の喋りの特徴を反映した調音運動を推定するためは、発話者の喋りの特徴を残したままモデル話者の音声に近づけるが可能な話者変換手法を検討する必要がある。あるいは，発話者の母語発声において，同一の調音運動をすると考えられるモデル話者を用意する必要がある。

文　　献

[1] P. Badin,A. Youssef, G. Bailly, F. Elisei, and T. Hueber, ”Visual articulatory feedback for phonetic correction in second language learning,” In *L2SW, Workshop on ”Second Language Studies: Acquisition, Learning, Education and Technology,”* pp. P1-10, 2010.

[2] A. Suemitsu, J. Dang, T. Ito, M. Tiede, ”A study on effect of real-time articulatory feedback presentation in American English pronunciation learning”, In *Proc. Acoustic society of Japan Autumn Meeting 2013*.pp. 427-428, 2013.

[3] T. Kaburagi, K. Wakamiya, and M. Honda, ”Three-dimensional electromagnetic articulography,” *Journal of the Acoustical Society of America*, vol. 118, pp428-443.

[4] W.Hardcastle, W.Jones and C.Knight, ”New developments in electropalatography: A state-of-the-art report,” Clinical Linguistics & Phonetics, 1989, vol. 3, No. 1, pp1-38.

[5] A. Wrench, F. Gibbon, A. M. McNeill, and S. Wood, ”An EPG therapy protocol for remediation and assessment of articulation disorders,”In *Proc. ICSLP 2002*, Denver, USA, pp. 965 -968, 2002.

[6] T. Takano, Y. Sakamot, and T. Kaburagi, ”A study on the inverse estimation from formant frequencies to the vocal-tract shape using a sensitivity funtion,” *IEICE Technical Report*, SP2011-73, pp. 25-30, 2011.

[7] S. Hiroya, and M. Honda, ”Estimation of articulatory movements from speech acoustics using an HMM-based speech production model,” *IEEE Trans. Speech and Audio Processing*, vol. 12, pp. 175-185, 2004.

[8] T. Toda, W. A. Black, and K. Tokuda, ”Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model”, *Speech Commun*, vol. 50,pp.215 -227, 2008.

[9] B. Uria, S. Renals, and K. Richmond, ”A deep neural network for acoustic-articulatory speech inversion,” In *Proc. NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*, Sierra Nevada, Spain, 2011.

[10] Y. Stylianou, O. Capp, and E. Moulines, ”Continuous probabilistic transform for voice conversion,” *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131142, Mar. 1998.

[11] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Many-to-many eigenvoice conversion with reference voice,” in *Proc. INTERSPEECH, Brighton, U.K., Sep. 2009*, pp. 16231626.

[12] MOCHA-TIMIT - Centre for Speech Technology Research, 5 June 2014, ”http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html”.