話者変換音声を対象とした音声–調音マッピングに関する実験的検討 *

☆内田秀継, 齋藤大輔, 峯松信明, 広瀬啓吉（東大）

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、発話者の調音運動を提示することによってトレーニングがより効果的になることが報告されている [1][2]。しかし、調音運動情報を得るためには特別な調音観測システムが必要となり、その測定コストが調音運動情報を利用する上で大きな障害となっている。そこで、調音観測システムを用いずに音響情報から調音運動を推定する手法が検討されている [3]。

音声から調音運動を推定する手法として、音声–調音運動パラレルデータと統計的手法によって音声–調音運動の関係をモデル化し、音声–調音変換を行う手法が検討されている [4]。しかし、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。

先行研究 [4] で用いられた統計的変換技術は，本来話者性の変換を目的として構築されたものである。そこで、本稿では、話者変換によって、入力音声をモデル話者に近づけることで、変換モデルの話者依存性の緩和を試みた。

2 話者変換音声を対象とした GMM による音声-調音マッピング法

2.1 話者変換と音声ー調音運動変換の統合

今、特定話者（以下モデル話者）の音声–調音運動のパラレルデータから音声–調音変換モデルが構築されている場合を考える。本研究では、モデル話者以外の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、別の発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。

2.2 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、発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 $\boldsymbol{z}^{(xy)} = [\boldsymbol{x}, \boldsymbol{y}]$ および $\boldsymbol{z}^{(sa)} = [\boldsymbol{y}^{(s)}, \boldsymbol{y}^{(a)}]$ の確率密度を以下の二つの混合ガウス分布（GMM）で表す;

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

$$P(\boldsymbol{z}^{(sa)}; \boldsymbol{\lambda}^{(sa)}) = \sum_{n=1}^{N} \beta_n \mathcal{N}(\boldsymbol{z}^{(sa)}; \boldsymbol{\mu}_{\boldsymbol{n}}^{(sa)}, \boldsymbol{\Sigma}_n^{(sa)}) \quad (2)$$

$$\boldsymbol{\mu}_m^{(xy)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix} \boldsymbol{\Sigma}_m^{(xy)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(x,x)} & \boldsymbol{\Sigma}_m^{(x,y)} \\ \boldsymbol{\Sigma}_m^{(y,x)} & \boldsymbol{\Sigma}_m^{(y,y)} \end{bmatrix} \quad (3)$$

$$\boldsymbol{\mu}_n^{(sa)} = \begin{bmatrix} \boldsymbol{\mu}_n^{(s)} \\ \boldsymbol{\mu}_n^{(a)} \end{bmatrix} \boldsymbol{\Sigma}_n^{(sa)} = \begin{bmatrix} \boldsymbol{\Sigma}_n^{(s,s)} & \boldsymbol{\Sigma}_n^{(s,a)} \\ \boldsymbol{\Sigma}_n^{(a,s)} & \boldsymbol{\Sigma}_n^{(a,a)} \end{bmatrix} \quad (4)$$

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, N は正規分布の混合数、α, β は各混合成分の重みパラメータである。

発話者の音声を調音運動に変換する場合は、まず、発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。

$$\hat{\boldsymbol{y}}^{(s)} = \arg\max_{\boldsymbol{y}^{(s)}} P(\boldsymbol{y}^{(s)} | \boldsymbol{x}^{(s)}; \boldsymbol{\lambda}^{(xy)}) \quad (5)$$

この $\hat{\boldsymbol{y}}^{(s)}$ は、発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。

$$\hat{\boldsymbol{y}}^{(a)} = \arg\max_{\boldsymbol{y}^{(a)}} P(\boldsymbol{y}^{(a)} | \hat{\boldsymbol{y}}^{(s)}; \boldsymbol{\lambda}^{(sa)}) \quad (6)$$

この手法では、発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。

2.3 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、発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 (1),(2) 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）、各混合成分におけるモデル話者の音声特徴量の平均ベクトルと分散共分散行列をそれぞれ同一のものと見なすことができる [5]。そして、モデル話者の音声特徴量を周辺化することで、

* An experimental study of applying statistical acoustics-to-articulatory mapping to voice-converted utterances, by Hidetsugu UCHIDA, Daisuke SAITO, Nobuaki MINEMATSU, Keikichi HIROSE (University of Tokyo)

発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。

発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。

$$\hat{\boldsymbol{y}}^{(a)} = \arg\max_{\boldsymbol{y}^{(a)}} P(\boldsymbol{y}^{(a)}|\boldsymbol{x}^{(s)};\boldsymbol{\lambda}) \tag{7}$$

上式を、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、

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

$$= \arg\max_{\boldsymbol{y}^{(a)}} \sum_{m=1}^{M} P(m|\boldsymbol{x}^{(s)}) \times \int_{\mathcal{Y}^{(s)}} P(\boldsymbol{y}^{(a)}|\boldsymbol{y}^{(s)},m) P(\boldsymbol{y}^{(s)}|\boldsymbol{x}^{(s)},m) d\boldsymbol{y}^{(s)} \tag{9}$$

となる。$\boldsymbol{y}^{(s)}$ について全積分を行うことで、

$$\hat{\boldsymbol{y}}^{(a)} = \arg\max_{\boldsymbol{y}^{(a)}} \sum_{m=1}^{M} P(m|\boldsymbol{x}^{(s)}) \mathcal{N}(\boldsymbol{y}^{(a)}|\boldsymbol{E}_m, \boldsymbol{D}_m) \tag{10}$$

を得る。ここで、

$$\boldsymbol{E}_m = \boldsymbol{\mu}_m^{(a)} + \boldsymbol{\Sigma}'_m(\boldsymbol{x}^{(s)} - \boldsymbol{\mu}_m^{(x)}) \tag{11}$$

$$\boldsymbol{D}_m = \boldsymbol{\Sigma}_m^{(a,a)} - \boldsymbol{\Sigma}'_m \boldsymbol{\Sigma}_m^{(x,x)} \boldsymbol{\Sigma}'^{\mathrm{T}}_m \tag{12}$$

$$\boldsymbol{\Sigma}'_m = \boldsymbol{\Sigma}_m^{(a,s)} \boldsymbol{\Sigma}_m^{(y,y)^{-1}} \boldsymbol{\Sigma}_m^{(y,x)} \tag{13}$$

である。式 (10) を変換式とすることで、話者変換音声を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 (13) の形で記述されていることがわかる。

3 実験

3.1 実験条件

本実験では、MOCHA データベース [6] を用いて各変換モデルを構築した。データベースには、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、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) の計 7 点の測定点に関する 2 次元位置情報である。音声特徴量と調音運動特徴量は、文献 [1] と同様に求めた。

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統合モデルを構築し、音声-調音マッピングを行った。さらに、比較のために、男性話者の音声–調音変換モデルを構築し、男性話者の音声を入力した場合 (以下 reference) の調音運動及び、女性話者の音声を直接入力した場合 (以下 baseline) の調音運動を求めた。各モデルの学習と評価は、データベース内の発話に対して、1/5 cross-validation を用いた。

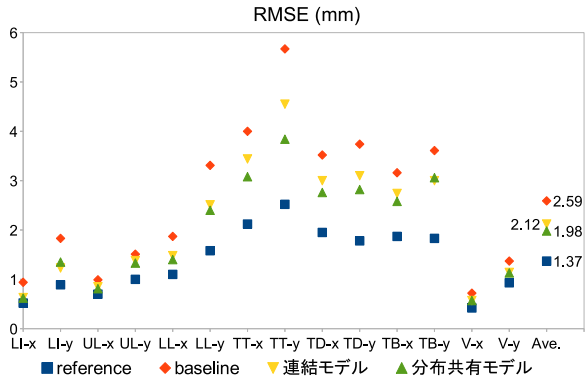

Fig. 1 RMSE of each measurement point

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、各手法で得られた調音運動の推定誤差を求めた結果を Fig.1 に示す。推定誤差は、全発話に対する二乗平均平方根誤差 (RMSE) とし、各測定点の水平方向 (*-x) および垂直方向 (*-y) について算出した。

Fig.1 の reference に注目すると、舌上の測定点に対する推定誤差が他の測定点と比べて大きく、baseline の推定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、推定精度が大きく改善されている。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、全測定点における推定誤差の平均 (Ave.) において、分布共有モデルが連結モデルよりも、 0.14 mm 上回っている。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。

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。その結果、統合モデルを用いることで、発話者とモデル話者の不一致による推定精度の悪化を軽減できることが示された。

参考文献

[1] A. Wrench, *et al.*, In *Proc. ICSLP-2002*, pp.965-968, 2002.

[2] A. Suemitsu, *et al.*, In *Proc. Acoustic society of Japan Autumn Meeting 2013*, pp.427-428, 2013.

[3] S. Hiroya, *et al.*, *IEEE Trans. Speech and Audio Processing*, vol. 12, pp. 175-185, 2004.

[4] T. Toda, *et al.*, *Speech Commun*, vol. 50, pp.215-227, 2008.

[5] Y. Ohtani, *et al.*, *IEICE Technical Report*, SP2008-140, pp.85-90, 2009

[6] http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html