Statistical Acoustic-to-Articulatory Mapping Unified with Speaker Normalization Based on Voice Conversion

Hidetsugu Uchida, Daisuke Saito, Nobuaki Minematsu, Keikichi Hirose

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

{uchida,dsk_saito,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract

This paper proposes a model of speaker-normalized acoustic-to-articulatory mapping using statistical voice conversion. A mapping function from acoustic parameters to articulatory parameters is usually developed with a single speaker's parallel data. Hence the constructed mapping model can work appropriately only for this specific speaker, and applying this model to other speakers degrades the performance of acoustic-to-articulatory mapping. In this paper, two models of speaker conversion and acoustic-to-articulatory mapping are implemented using Gaussian Mixture Models (GMM), and by integrating these two models, we propose two methods of speaker-normalized acoustic-to-articulatory mapping. One is concatenating these models sequentially, and the other integrates the two models into a unified model, where acoustic parameters of a speaker can be converted directly to articulatory parameters of another speaker. Experiments show that both methods can improve the mapping accuracy and that the latter method works better than the former method. Especially in the case of velar stop consonants, the mapping accuracy is higher by 0.6 mm.

Index Terms: acoustic-to-articulatory mapping, Gaussian mixture model, voice conversion, speaker normalization

1. Introduction

A new type of pronunciation training system using information of articulatory movements has been studied [1]. The training system provides learners with visual feedback of articulatory movements both of the learners and teachers and it can improve the learners' pronunciation more effectively than a conventional system that uses only audio features [2].

To implement such systems, it is necessary to establish a method to estimate learners' articulatory movements from their utterances. Articulatory movements can be measured directly when special instruments are available, e.g. electromagnetic articulography [3] and electropalatography [4][5]. Since they are large-scale instruments, however, it is surely difficult to incorporate them into personalized pronunciation training. These days, methods to estimate articulatory movements from voices have been studied using a parallel corpus of acoustic and articulatory observation. In these studies, several mapping models have been proposed, e.g. Gaussian Mixture Model (GMM) [6], Hidden Markov Model (HMM) [7], and Deep Neural Network (DNN) [8]. A mapping model developed with acoustic-articulatory parallel data of a single speaker has to be a speaker-dependent model. If voices of a different speaker are input to the mapping model, acoustic and articulatory mismatch between the input and the model causes large errors in mapping. The problem can be avoided by constructing a model for each individual with his/her own parallel data. However, this solution is impractical because simultaneous recording of acoustic and articulatory data is expensive. We need a method that can estimate articulatory movements of arbitrary speakers only by using a specific speaker's acoustic-to-articulatory mapping model.

Statistical mapping methods are widely used in the voice conversion field [9]. Speaker conversion, which can be viewed as applied voice conversion, is a technique which can convert speaker identity of input voices to a target speaker without changing linguistic content. In this paper, a speaker conversion technique is used to conduct speaker normalization, where speaker identity of input voices is changed to the speaker used for constructing the acoustic-to-articulatory mapping model. For open speakers, mapping accuracy will be improved.

In this study, we investigate two methods which can estimate articulatory movements from arbitrary speakers' voices. One is concatenating a GMM-based speaker conversion model and a GMM-based acoustic-to-articulatory mapping model sequentially. The other integrates the models into a unified model.

2. Speaker normalization based on voice conversion

2.1. GMM-based parameter mapping

In this paper, both models of speaker conversion and acoustic-to-articulatory mapping are implemented as GMM-based parameter conversion, which is briefly described below.

Let $\boldsymbol{x} \in \mathcal{R}^{d_x}$ and $\boldsymbol{y} \in \mathcal{R}^{d_y}$ be source and target parameter vectors whose dimensions are d_x and d_y, respectively. $\boldsymbol{z}$ denotes a joint vector consisting of source and target parameters as $\boldsymbol{z} = [\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$. The probability density of the joint vector is modeled by using a GMM as follows:

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

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

$\boldsymbol{\lambda}$ is model parameters. $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a normal distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. M is the total number of mixture components and α is a weight parameter. The parameter mapping function using the GMM is derived from

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

A GMM-based speaker conversion model is developed with acoustic-acoustic parallel data, where source and target parameters are speech features of source and target speakers, respectively. A GMM-based acoustic-to-articulatory mapping model is developed with acoustic-articulatory parallel data, where a source parameter is a speech feature and a target parameter is an articulatory parameter. In this paper, the speaker used for the acoustic-to-articulatory mapping model is called *model* speaker henceforth. Then, GMM-based speaker conversion is done to change speaker identity of input voices to the model speaker.

Since GMM-based conversion is a frame-to-frame mapping, however, it should be noted that the temporal structure of input utterances is copied onto output utterances[1]. Considering this technical fact, the converted speech can be regarded

[1] Even if ML-based estimation is employed, the number of frames is the same between the input and the output [6].

as speech generated by the model speaker with a speaking style similar to the input speaker's style. If the speaker-normalized voices are input to the acoustic-to-articulatory mapping model, the output is the model speaker's articulatory movements generated in a speaking style similar to the input speaker's style. In other words, we can regard the output as the model speaker's imitation of input utterances. In the case of pronunciation training, we can assume that an input speaker is a learner and the model speaker is a teacher. Therefore, the estimated articulatory movements are regarded as the teacher's articulation when he/she imitates the learner's utterances. A teacher often imitates a learner's utterances to show clearly what is wrong with the utterances. This imitation is effective to increase learners' awareness of pronunciation. Our proposed method is expected to work as teacher, where the feedback of imitated articulatory movements will be effective for pronunciation training.

2.2. Integrating GMM-based speaker conversion and acoustic-to-articulatory mapping

In this study, we integrate the speaker conversion model and the acoustic-to-articulatory mapping model when two types of parallel data are prepared: one is acoustic-articulatory parallel data of the model speaker, and the other is acoustic-acoustic parallel data between that model speaker and an arbitrary speaker. For integrating these two models, two methods are proposed: a concatenation method and a unification method.

2.3. Concatenation method

This method concatenates two models of speaker conversion and acoustic-to-articulatory mapping sequentially. Let $\boldsymbol{x}^{(s)}$, $\boldsymbol{y}^{(s)}$, and $\boldsymbol{y}^{(a)}$ be a speech vector of the input speaker, that of the model speaker, and an articulatory vector of the model speaker, respectively. The probability density functions of two joint vectors $\boldsymbol{z}^{(xy)}=[\boldsymbol{x}^{(s)\top}, \boldsymbol{y}^{(s)\top}]^\top$ and $\boldsymbol{z}^{(sa)}=[\boldsymbol{y}^{(s)\top}, \boldsymbol{y}^{(a)\top}]^\top$ are characterized respectively as $P(\boldsymbol{z}^{(xy)};\boldsymbol{\lambda}^{(xy)})=\sum_{m=1}^{M}\alpha_m\mathcal{N}(\boldsymbol{z}^{(xy)};\boldsymbol{\mu}_{\boldsymbol{m}}^{(xy)},\boldsymbol{\Sigma}_m^{(xy)})$ and $P(\boldsymbol{z}^{(sa)};\boldsymbol{\lambda}^{(sa)})=\sum_{n=1}^{N}\beta_n\mathcal{N}(\boldsymbol{z}^{(sa)};\boldsymbol{\mu}_{\boldsymbol{n}}^{(sa)},\boldsymbol{\Sigma}_n^{(sa)})$.

When applying the resulting model to acoustic-to-articulatory mapping, $\boldsymbol{x}^{(s)}$ is firstly converted to $\hat{\boldsymbol{y}}^{(s)}$ by Eq. 3, where $\boldsymbol{x}=\boldsymbol{x}^{(s)}$ and $\boldsymbol{y}=\boldsymbol{y}^{(s)}$. Then, $\hat{\boldsymbol{y}}^{(s)}$ is converted to articulatory movements by Eq. 3 again, where $\boldsymbol{x}=\boldsymbol{y}^{(s)}$ and $\boldsymbol{y}=\boldsymbol{y}^{(a)}$.

2.4. Unification method

We propose another method that can integrate the two models into a unified model, where input voices are converted directly to articulatory movements of the model speaker.

In the concatenation method, one can find that two probability densities of $P(\boldsymbol{z}^{(xy)};\boldsymbol{\lambda}^{(xy)})$ and $P(\boldsymbol{z}^{(sa)};\boldsymbol{\lambda}^{(sa)})$ include a common distribution of $\boldsymbol{y}^{(s)}$[2]. Practically speaking, distribution of $\boldsymbol{y}^{(s)}$ in $P(\boldsymbol{z}^{(xy)};\boldsymbol{\lambda}^{(xy)})$ and that in $P(\boldsymbol{z}^{(sa)};\boldsymbol{\lambda}^{(sa)})$ may be different when parallel data of $(\boldsymbol{x}^{(s)},\boldsymbol{y}^{(s)})$ and those of $(\boldsymbol{y}^{(s)},\boldsymbol{y}^{(a)})$ are obtained by reading different sentences, for example. By assuming that the two kinds of distributions of $\boldsymbol{y}^{(s)}$ can be characterized by a single model even in such a case, the total number of parameters that have to be estimated for acoustic-to-articulatory mapping can be reduced [10]. Furthermore, by marginalizing the parameters of $P(\boldsymbol{y}^{(s)})$, the resulting mapping model becomes a unified model that can convert input voices to articulatory movements of the model speaker directly.

Input voices of speaker x are denoted as $\boldsymbol{x}^{(s)}$. In the unification method, articulatory movements are obtained by

$$\hat{\boldsymbol{y}}^{(a)} = \arg\max_{\boldsymbol{y}^{(a)}} P(\boldsymbol{y}^{(a)}|\boldsymbol{x}^{(s)};\boldsymbol{\lambda}). \tag{4}$$

The above equation can be approximated in the following way:

$$\begin{aligned}\hat{\boldsymbol{y}}^{(a)} &= \arg\max_{\boldsymbol{y}^{(a)}} P(\boldsymbol{y}^{(a)}|\boldsymbol{x}^{(s)}) && (5)\\ &\approx \arg\max_{\boldsymbol{y}^{(a)}} \sum_{m=1}^{M} P(m|\boldsymbol{x}^{(s)})\times \\ &\quad \int_{\mathcal{Y}^{(s)}} P(\boldsymbol{y}^{(a)}|\boldsymbol{y}^{(s)},m)P(\boldsymbol{y}^{(s)}|\boldsymbol{x}^{(s)},m)d\boldsymbol{y}^{(s)} && (6)\\ &= \arg\max_{\boldsymbol{y}^{(a)}} \sum_{m=1}^{M} P(m|\boldsymbol{x}^{(s)})\mathcal{N}(\boldsymbol{y}^{(a)}|\boldsymbol{E}_m,\boldsymbol{D}_m), && (7)\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{E}_m &= \boldsymbol{\mu}_m^{(a)} + \boldsymbol{\Sigma}'_m(\boldsymbol{x}^{(s)} - \boldsymbol{\mu}_m^{(x)}), && (8)\\ \boldsymbol{D}_m &= \boldsymbol{\Sigma}_m^{(a,a)} - \boldsymbol{\Sigma}'_m\boldsymbol{\Sigma}_m^{(x,x)}\boldsymbol{\Sigma}'^{\top}_m, && (9)\\ \boldsymbol{\Sigma}'_m &= \boldsymbol{\Sigma}_m^{(a,s)}\boldsymbol{\Sigma}_m^{(y,y)^{-1}}\boldsymbol{\Sigma}_m^{(y,x)}. && (10)\end{aligned}$$

By using Eq. 7, we can estimate articulatory movements of the model speaker from the input speaker's voices without converting them to the model speaker's voices explicitly.

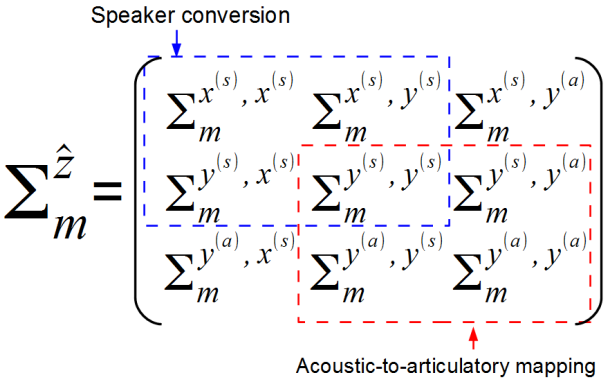

Figure 1: The covariance matrix of a component of GMM for the joint vector $\hat{\boldsymbol{z}}$

2.5. Implementation of the unification method

The unification method requires that each GMM component of $\boldsymbol{y}^{(s)}$ should be the same between the speaker conversion model and the acoustic-to-articulatory mapping model. To satisfy this condition, we first build a GMM of a new joint vector of $\hat{\boldsymbol{z}} = [\boldsymbol{x}^{(s)\top}, \boldsymbol{y}^{(s)\top}, \boldsymbol{y}^{(a)\top}]^\top$. Fig. 1 illustrates a covariance matrix in the GMM of $\hat{\boldsymbol{z}}$, where the covariance matrix of $\boldsymbol{z}^{(xy)}$ and that of $\boldsymbol{z}^{(sa)}$ coexist, between which the distribution of $\boldsymbol{y}^{(s)}$ is shared. Namely, the two models of speaker conversion and acoustic-to-articulatory mapping share the distribution of $\boldsymbol{y}^{(s)}$.

The simplest method to build a GMM of $\hat{\boldsymbol{z}}$ is to estimate its parameters by using samples of $(\boldsymbol{x}^{(s)},\boldsymbol{y}^{(s)},\boldsymbol{y}^{(a)})$. In the context of pronunciation training, however, this method is not practical. The speaker conversion model has to be made by a parallel data of the same language, namely, a learner's native language[3]. The acoustic-articulatory mapping model should be made by a parallel data of the target language of learning. These conditions require the model speaker to be a bilingual speaker

[2] $\boldsymbol{z}^{(xy)}=[\boldsymbol{x}^{(s)\top}, \boldsymbol{y}^{(s)\top}]^\top$ and $\boldsymbol{z}^{(sa)}=[\boldsymbol{y}^{(s)\top}, \boldsymbol{y}^{(a)\top}]^\top$

[3] Learners are not good speakers of the target language.

of both languages and under these conditions, simultaneous observation of $\boldsymbol{x}^{(s)}$, $\boldsymbol{y}^{(s)}$, and $\boldsymbol{y}^{(a)}$ is practically impossible.

To build a GMM of $\hat{\boldsymbol{z}}$ by using two non-overlapped parallel databases of $(\boldsymbol{x}^{(s)}, \boldsymbol{y}^{(s)})$ and $(\boldsymbol{y}^{(s)}, \boldsymbol{y}^{(a)})$, we have to derive a method to build a GMM by using $(\boldsymbol{x}^{(s)}, \boldsymbol{y}^{(s)}, \boldsymbol{y}^{(a)})$, where either of $\boldsymbol{x}^{(s)}$ or $\boldsymbol{y}^{(a)}$ is always missing. For derivation, joint vector $\hat{\boldsymbol{z}}$ is re-introduced as $\boldsymbol{z}_t{}^{(xyy')}=[\boldsymbol{x}_t{}^{(s)\top}, \boldsymbol{y}_t{}^{(s)\top}, \boldsymbol{y}'^{(a)\top}_t]^\top$ when $t = 1, \ldots, T_1$ and $\boldsymbol{z}_t{}^{(s'sa)}=[\boldsymbol{x}'^{(s)\top}_t, \boldsymbol{y}_t{}^{(s)\top}, \boldsymbol{y}_t{}^{(a)\top}]^\top$ when $t = T_1+1, \ldots, T_2$. Here, $\boldsymbol{y}'^{(a)}_t$ is a missing articulatory vector of the model speaker and $\boldsymbol{x}'^{(s)}_t$ is a missing speech vector of the input speaker. The E-step of the general EM algorithm for GMM of $\hat{\boldsymbol{z}}$ is shown below.

$$\gamma_{m,t}^{(xyy')} = \frac{\pi_m \mathcal{N}(\boldsymbol{z}_t^{(xyy')} | \boldsymbol{\mu}_m{}^{(\hat{z})}, \boldsymbol{\Sigma}_{\boldsymbol{m}}{}^{(\hat{z})})}{\sum_{j=1}^{M} \pi_j \mathcal{N}(\boldsymbol{z}_t^{(xyy')} | \boldsymbol{\mu}_j{}^{(\hat{z})}, \boldsymbol{\Sigma}_{\boldsymbol{j}}{}^{(\hat{z})})} \quad (11)$$

$$\gamma_{m,t}^{(s'sa)} = \frac{\pi_m \mathcal{N}(\boldsymbol{z}_t^{(s'sa)} | \boldsymbol{\mu}_m{}^{(\hat{z})}, \boldsymbol{\Sigma}_{\boldsymbol{m}}{}^{(\hat{z})})}{\sum_{j=1}^{M} \pi_j \mathcal{N}(\boldsymbol{z}_t^{(s'sa)} | \boldsymbol{\mu}_j{}^{(\hat{z})}, \boldsymbol{\Sigma}_{\boldsymbol{j}}{}^{(\hat{z})})} \quad (12)$$

Since $\boldsymbol{y}'^{(a)}_t$ and $\boldsymbol{x}'^{(s)}_t$ are missing, they have to be replaced by some appropriate terms. Here, their expected values are used.

$$\boldsymbol{y}'_t{}^{(a)} \leftarrow \sum_{m=1}^{M} P(m|\boldsymbol{\mu}_m^{(xy)}, \boldsymbol{\Sigma}_m^{(xy)}) E_{m,t}^{(y^{(a)}|x^{(s)},y^{(s)})}. \quad (13)$$

$$\boldsymbol{x}'_t{}^{(s)} \leftarrow \sum_{m=1}^{M} P(m|\boldsymbol{\mu}_m^{(sa)}, \boldsymbol{\Sigma}_m^{(sa)}) E_{m,t}^{(x^{(s)}|y^{(s)},y^{(a)})}. \quad (14)$$

$\boldsymbol{\mu}_m^{(xy)}$ and $\boldsymbol{\Sigma}_m^{(xy)}$ are estimated from samples of joint vector $(\boldsymbol{x}^{(s)}, \boldsymbol{y}^{(s)})$. Similarly, $\boldsymbol{\mu}_m^{(sa)}$ and $\boldsymbol{\Sigma}_m^{(sa)}$ are estimated from samples of joint vector $(\boldsymbol{y}^{(s)}, \boldsymbol{y}^{(a)})$. $E_{m,t}^{(y^{(a)}|x^{(s)},y^{(s)})}$ and $E_{m,t}^{(x^{(s)}|y^{(s)},y^{(a)})}$ are obtained as follows:

$$E_{m,t}^{(y^{(a)}|x^{(s)},y^{(s)})} = \boldsymbol{\mu}_m^{(a)} + \boldsymbol{\Sigma}_{\boldsymbol{m}}^{(a,xy)} \boldsymbol{\Sigma}_m^{(xy,xy)^{-1}} (\boldsymbol{z}_t^{(xy)} - \boldsymbol{\mu}_m^{(xy)}) \quad (15)$$

$$E_{m,t}^{(x^{(s)}|y^{(s)},y^{(a)})} = \boldsymbol{\mu}_m^{(s)} + \boldsymbol{\Sigma}_m^{(x,sa)} \boldsymbol{\Sigma}_m^{(sa,sa)^{-1}} (\boldsymbol{z}_t^{(sa)} - \boldsymbol{\mu}_m^{(sa)}). \quad (16)$$

$\boldsymbol{\mu}_m$, $\boldsymbol{\Sigma}_m$, and π_m are updated in the M-step below.

$$\boldsymbol{\mu}_m = \frac{1}{\gamma_m} \left(\sum_{t=1}^{T_1} \gamma_{m,t}{}^{(xyy')} \boldsymbol{z}_{\boldsymbol{t}}^{(xyy')} + \sum_{t=T_1+1}^{T_2} \gamma_{m,t}{}^{(s'sa)} \boldsymbol{z}_{\boldsymbol{t}}^{(s'sa)} \right) \quad (17)$$

$$\boldsymbol{\Sigma}_m = \frac{1}{\gamma_m} \left(\sum_{t=1}^{T_1} \gamma_{m,t}{}^{(xyy')} \left\{ (\boldsymbol{z}_{\boldsymbol{t}}^{(xyy')} - \boldsymbol{\mu}_m)(\boldsymbol{z}_{\boldsymbol{t}}^{(xyy')} - \boldsymbol{\mu}_m)^\top - \hat{\boldsymbol{D}}_m^{(xyy')} \right\} \right.$$
$$\left. + \sum_{t=T_1+1}^{T_2} \gamma_{m,t}{}^{(xyy')} \left\{ (\boldsymbol{z}_{\boldsymbol{t}}^{(xyy')} - \boldsymbol{\mu}_m)(\boldsymbol{z}_{\boldsymbol{t}}^{(xyy')} - \boldsymbol{\mu}_m)^\top - \hat{\boldsymbol{D}}_m^{(xyy')} \right\} \right) \quad (18)$$

$$\pi_m = \frac{\gamma_m}{\sum_{m=1}^{M} \gamma_m} \quad (19)$$

where $\gamma_m = \Sigma_{t=1}^{T_1} \gamma_{m,t}^{(xyy')} + \Sigma_{t=T_1+1}^{T_2} \gamma_{m,t}^{(s'sa)}$. In the M-step, Eq.15 and Eq.16 are substituted again for missing $\boldsymbol{y}'^{(a)}_t$ and $\boldsymbol{x}'^{(s)}_t$. $\hat{\boldsymbol{D}}_m^{(xyy')}$ and $\hat{\boldsymbol{D}}_{\boldsymbol{m}}^{(s'sa)}$ in Eq.18 are obtained as

$$\hat{\boldsymbol{D}}_m^{(xyy')} = \begin{bmatrix} \boldsymbol{0}^{(d_1,d_1)} & \boldsymbol{0}^{(d_1,d_2)} \\ \boldsymbol{0}^{(d_2,d_1)} & \boldsymbol{D}_{\boldsymbol{m}}^{(y^{(a)}|x^{(s)},y^{(s)})} \end{bmatrix} \quad (20)$$

$$\hat{\boldsymbol{D}}_m^{(s'sa)} = \begin{bmatrix} \boldsymbol{D}_{\boldsymbol{m}}^{(x^{(s)}|y^{(s)},y^{(a)})} & \boldsymbol{0}^{(d_3,d_4)} \\ \boldsymbol{0}^{(d_4,d_3)} & \boldsymbol{0}^{(d_4,d_4)} \end{bmatrix}, \quad (21)$$

where

$$\boldsymbol{D}_m^{(y^{(a)}|x^{(s)},y^{(s)})} = \boldsymbol{\Sigma}_m^{(a,a)} - \boldsymbol{\Sigma}_m^{(a,xy)} \boldsymbol{\Sigma}_m^{(xy,xy)^{-1}} \boldsymbol{\Sigma}_m^{(xy,a)}, \quad (22)$$

$$\boldsymbol{D}_m^{(x^{(s)}|y^{(s)},y^{(a)})} = \boldsymbol{\Sigma}_m^{(x,x)} - \boldsymbol{\Sigma}_m^{(x,sa)} \boldsymbol{\Sigma}_m^{(sa,sa)^{-1}} \boldsymbol{\Sigma}_m^{(sa,x)}. \quad (23)$$

$0^{(m,l)}$ is an $m \times l$ zero matrix, and d_1, d_2, d_3, and d_4 are dimensions of $\boldsymbol{z}^{(xy)}$, $\boldsymbol{y}^{(a)}$, $\boldsymbol{z}^{(x)}$, and $\boldsymbol{z}^{(sa)}$, respectively.

3. Experimental evaluations

3.1. The database used for experiments

To evaluate the performance of our proposal of speaker-normalized acoustic-to-articulatory mapping model, we conducted experimental evaluations using MOCHA database [11]. This database includes acoustic-articulatory parallel data of one male speaker and one female speaker. In the experiments, either of the two was used as model speaker and the other was an input speaker. Articulatory data are measured by an electromagnetic articulography, where its sensors are placed at lower incisor (LI), upper and lower lips (UL and LL), tongue tip (TT), tongue body (TB), tongue dorsum (TD) and velum (V) in the med-sagittal plane. The articulatory data of each point are two-dimensional data of horizontal (x-axis) and vertical (y-axis) directions. The parallel data were divided into 5 parts, where 4 parts were used for training and the other was used for testing.

3.2. Conditions

For the experiments, three kinds of features were extracted, which are related to acoustic and articulatory observation at time t: 1) 0-24$^{\text{th}}$ MFCCs and their deltas, 2) the first 75 principal components obtained from PCA of 11 frames of MFCCs centering at time t [6], and 3) 28-dimensional data consisting of 14-dimensional articulatory parameters (2-dimensional data obtained from the 7 sensors) and their deltas. It should be noted that, by following [6], acoustic-to-articulatory conversion was implemented from 2) to 3), not from 1) to 3).

In the concatenation method, firstly, conversion of MFCCs of an input speaker to MFCCs of the model speaker was carried out. After applying PCA, acoustic-to-articulatory mapping was done. On the other hand in the unification method, joint vectors of PCA-based features of an input speaker, those of the model speaker, and articulatory parameters of the model speaker were formed to estimate their GMM, where some values are missing as explained in section 2.5. Then, the PCA-based features of the input speaker was converted directly to the articulatory parameters of the model speaker. For each conversion process, ML-based estimation using dynamic features was adopted. For time alignment between two sequences, we employed DTW.

For comparison, we built another acoustic-to-articulatory mapping model using MOCHA's acoustic-articulatory parallel data of the model speaker. This model is called as reference model. We conducted mapping experiments in four conditions:

1. reference: the reference model is used, where both the input and output speakers are the model speaker.
2. baseline: the reference model is used again but the input speaker is different from the model (output) speaker.
3. concatenation: the speaker-normalized model based on concatenation is used, where the input speaker is different from the model (output) speaker.
4. unification: the speaker-normalized model based on unification is used, where the input speaker is different from the model (output) speaker.

Five-fold cross-validation tests were employed. For that, MOCHA database of two speakers was divided into several

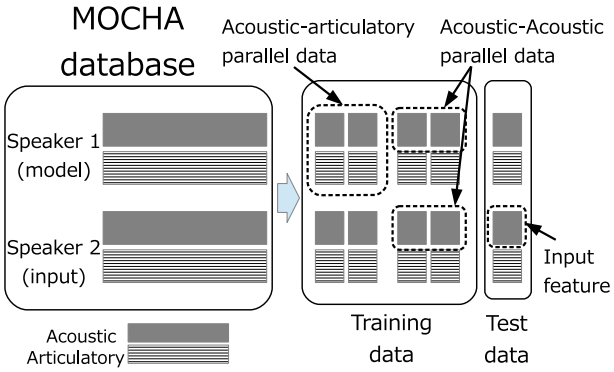

Figure 2: Illustration of how MOCHA database was used in the experiments

parts and, as shown in Fig. 2, they were used well to realize cross-validation. The training data of both speakers was divided into 2 partitions. One was used as acoustic-acoustic parallel data between the two speakers and the other was as acoustic-articulatory parallel data within each speaker. As told in section 2.5, the acoustic-acoustic data should be in L1 of a learner and the acoustic-articulatory data should be in the target language of learning. In this paper, however, to assess our proposal only from a technical viewpoint, not from an educational one, the above condition was adopted, where acoustic-acoustic sentences and acoustic-articulatory sentences were in the same language but not overlapped. To assess our proposal from an educational viewpoint, we are developing another database using bilingual speakers.

Since both speakers have their articulatory observations in the database, the root-mean-square error (RMSE) of each of the seven positions from LI to V was calculated in the four conditions of reference, baseline, concatenation, and unification. Here, the estimated movements have the same temporal structure of input utterances. To cancel time alignment gaps, DTW was conducted again. The resulting RMSE values were averaged for both model speakers.

3.3. Results and discussion

Fig. 3 shows RMSE of each of all the seven articulatory parameters and their average for each condition. We can see that, when mapping is done in the condition of baseline, the mapping accuracy is deteriorated especially in the case of tongue-related parameters such as TT, TB, and TD. The average of RMSE for those parameters is 3.77 mm. On the other hand, the averaged RMSEs of TT, TB, and TD in the conditions of concatenation and unification are 2.97 mm and 2.77 mm, respectively. It means that speaker dependency of baseline is alleviated by introducing speaker normalization. Further, the unification method is better than the concatenation method by 0.12 mm on average. This is considered to be because the concatenation method inevitably accumulates estimation errors.

Fig. 4 shows RMSE of each phoneme, where RMSE is averaged over all the articulatory parameters. We can see that, compared with the baseline condition, the unification method improves the mapping accuracy in all the cases. On the other hand, the concatenation method degrades the mapping accuracy in /k/ and /g/. To investigate the mapping accuracy in those two phonemes, we show their RMSEs of each parameter in Fig. 5. We can see that the mapping accuracy of the concatenation method is worse than that of baseline for TD-x and TD-y. On the other hand, the unification method improves the mapping accuracy even in those parameters. /k/ and /g/ are velar stop consonants, which are produced by having the back of the tongue touch the soft palate. When building a pronunciation assessment system, it is undesirable at all if the system has difficulty in assessing specific types of phonemes.

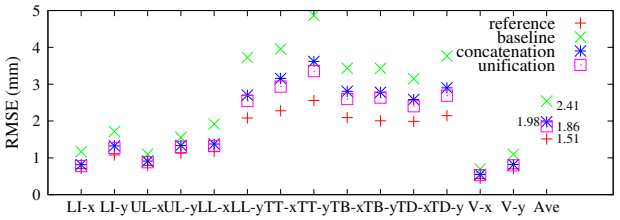

Figure 3: RMSE measured at each observation point

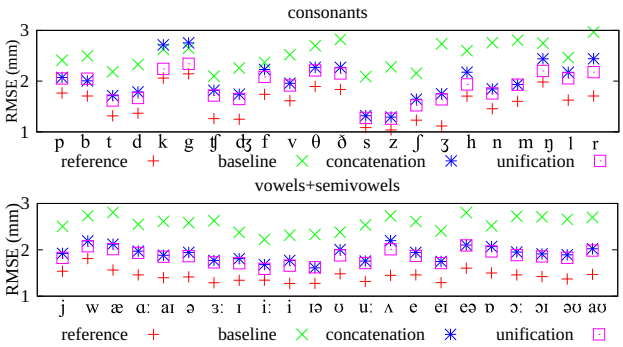

Figure 4: RMSE of each phoneme

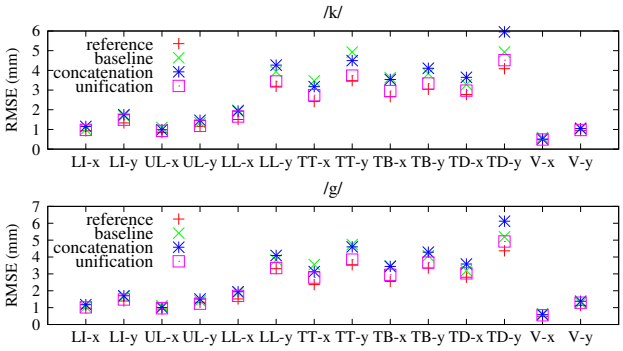

Figure 5: RMSE measured at each point for /k/ and /g/

4. Conclusions

In this paper, we proposed a speaker-normalized acoustic-to-articulatory mapping model based on speaker conversion. To implement the model, two methods were proposed and tested; the concatenation method and the unification method. To evaluate mapping accuracy of our proposed model, we conducted acoustic-to-articulatory mapping experiments using MOCHA database. As a result, it was shown that our proposed model can alleviate speaker dependency of acoustic-to-articulatory mapping, and that the unification method can estimate articulatory movements better than the concatenation method. In this paper, the experiments were conducted on the condition that acoustic-acoustic parallel data and acoustic-articulatory parallel data were of the same language. For future work, to assess our proposal in a more realistic situation, we will conduct experiments using parallel data collected using different languages and bilingual speakers.

5. References

[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. 1-10, 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.* 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, pp. 428-443, 2005

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

[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*, pp. 965-968, 2002.

[6] 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.

[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] 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*, 2011.

[9] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, No. 2, pp. 131-142, 1998.

[10] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Many-to-many eigenvoice conversion with reference voice," In *Proc. INTERSPEECH*, pp. 1623-1626, 2009

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