INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Prediction of the articulatory movements of unseen phonemes of a speaker using the speech structure of another speaker

Hidetsugu Uchida, Daisuke Saito, Nobuaki Minematsu

The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
{uchida,dsk_saito,mine}@gavo.t.u-tokyo.ac.jp

Abstract

In this paper, we propose a method to predict the articulatory movements of phonemes that are difficult for a speaker to pronounce correctly because those phonemes are not seen in the native language of that speaker. When one wants to predict the articulatory movements of those unseen phonemes, since he/she has difficulty to generate those sounds, the conventional acoustic-to-articulatory mapping cannot be applied as it is. Here, we propose a solution by using the speech structure of another reference speaker who can pronounce the unseen phonemes. Speech structure is a kind of speech feature that represents only the linguistic information by suppressing the non-linguistic information, e.g. speaker identity, of an input utterance. In the proposed method, by using the speech structure of those unseen phonemes and other phonemes as constraint, the articulatory movements of the unseen phonemes are searched for in the articulatory space of the original speaker. Experiments using English short vowels show that the averaged prediction error was 1.02 mm.

Index Terms: acoustic-to-articulatory mapping, Gaussian mixture model, speech structure, pronunciation training system

1. Introduction

Articulatory movements play an important role of creating phoneme characteristics in voices in a process of speech production. Therefore studies of articulatory movements cover many research topics both of engineering fields and scientific fields. Among these topics, speech training systems based on measuring or estimating articulatory movements have drawn researchers' attention as their potential is very high in language learning or speech therapy [1][2].

These training systems may provide learners with visual feedback of two kinds of articulatory movements, one is those of a learner or client and the other is those of his/her teacher or therapist. Users of these systems can correct mispronunciations intuitively by comparing their own articulatory movements with target articulatory movements [3]. In this situation, it would be ideal that those two articulatory movements should be visualized in the user's articulatory space for easy comparison. Here, the target articulatory movements are those which could be realized by the user himself when he improves his speaking skills. In language learning, for example, the target of speech training often includes phonemes that are not seen in the native language of a learner. Hence, the target articulatory movements are generally difficult to measure in a learner's articulatory space.

In this paper, we aim to predict those unseen articulatory movements in that learner's space. Acoustic-to-articulatory mapping has been studied to predict articulatory movements only from speech signals [4]. Generally speaking, the technique requires speech signals of a speaker of interest as inputs. Therefore, the conventional acoustic-to-articulatory mapping technique cannot be applied directly to unseen phonemes. In the proposed method, we use another speaker, teacher or therapist, who can pronounce those phonemes correctly. From that speaker, speech structure is extracted which includes those phonemes. Speech structure is a speech feature representing only the linguistic aspect of speech where its non-linguistic aspect such as speaker identity is effectively suppressed or removed. Here, a given utterance, i.e. a sequence of speech events such as phonemes, is represented only as distance matrix among those events, where event-to-event distance is measured as f-divergence. Since f-divergence is transform-invariant [5], the distances are regarded as speaker-invariant. In the proposed method, we extract a speech structure including the unseen phonemes from a reference speaker and use it as constraint when predicting the target articulatory movements in the original speaker's articulatory space.

2. Prediction of unseen articulatory movements

2.1. Speech structure [6]

Acoustic features of speech signals vary easily depending on their non-linguistic factors such as age and gender of the speaker and channel characteristics of transmission. Speech structure was proposed to remove those non-linguistic biases and extract only the linguistic aspect of speech signals [6]. Acoustic variation due to non-linguistic factors can be classified into two types. In the cepstrum domain, one is additive, $c' = c + b$, and the other is multiplicative, $c' = Ac$ [7]. Microphone difference is a good example for the former and vocal tract length difference is for the latter. Generally speaking, static and non-linguistic variation can be approximated as linear transformation of $c' = Ac + b$.

In structural analysis of speech, a speech sequence is converted at first to a sequence of feature distributions, which may correspond to speech events such as phonemes. Between any pair of events, Bhattachartyya distance (BD) is calculated and the resulting distances can form a distance matrix. This matrix is the speech structure of this utterance (see Figure 1).

Let $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ be probability density functions of two speech events in the cepstrum space. BD, which is one of f-divergences, between the two speech events is

$$BD(p_1, p_2) = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(\boldsymbol{x}) p_2(\boldsymbol{x})} d\boldsymbol{x}. \quad (1)$$

If both $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ follow Gaussian, $p_1(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $p_2(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, Eq.1 becomes

$$BD(p_1, p_2) = \frac{1}{8} \boldsymbol{\mu}_{12}^{\top} \boldsymbol{V}_{12}^{-1} \boldsymbol{\mu}_{1,2} + \frac{1}{2} \frac{|\boldsymbol{V}_{12}|}{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}} |\boldsymbol{\Sigma}_2|^{\frac{1}{2}}}. \quad (2)$$

http://dx.doi.org/10.21437/Interspeech.2016-1138

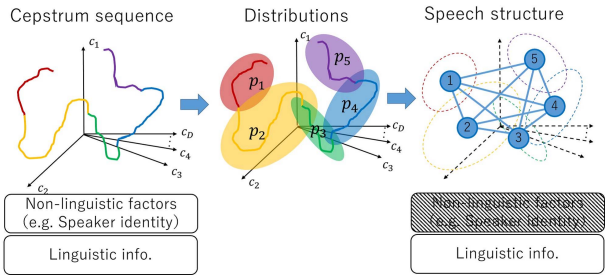

Figure 1: Speech structure

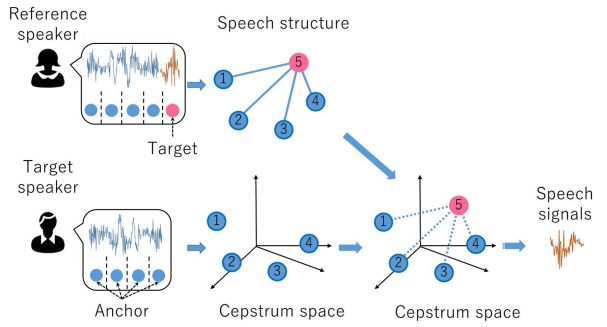

Figure 2: Speech generation using a speech structure

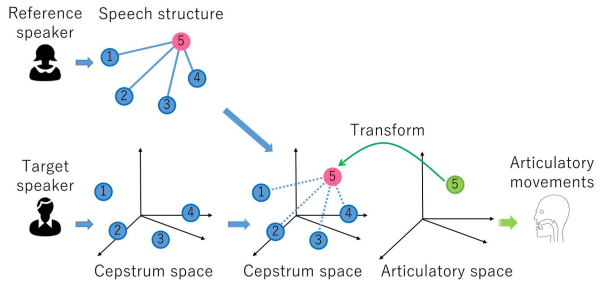

Figure 3: Prediction of articulatory movements using speech structure

Here, $\boldsymbol{\mu}_{12}$ is $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and $\boldsymbol{V}_{12}$ is $\frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}$. Mathematical properties of BD claim that BD between p_1 and p_2 is invariant against any kind of linear transform. Hence, the speech structure can be used as transform-invariant or speaker-invariant representation of speech. It should be noted in Figure.1 that, since a speech structure is just a distance matrix, it loses information about the positions of the events in the acoustic space where their structure is formed. In other words, the speech structure is a very abstract representation of speech.

2.2. Speech generation using a speech structure [8]

Speech generation derived from a given speech structure has been studied [8]. As told above, only with a structure, no event can be realized in an acoustic space as audio signals. To locate an event or a node of a given structure in the acoustic space, additional constraints have to be given. For example, if physical properties of a speaker are given, a speech structure may be able to be converted to that speaker's acoustic events, i.e. voices, of the linguistic message represented by the structure. In [8], however, instead of giving physical properties of a speaker, the absolute positions of some events are given in the acoustic space, and those of the remaining events were predicted. For example, in the speech structure of Figure 1, the positions of nodes 1, 2, 3, and 4 were given and, using these nodes as anchors and the distance matrix among the five nodes as constraints, the position of the 5-th node was predicted.

Suppose that we have a teacher and a learner. The teacher can pronounce all the five phonemes correctly but the learner can pronounce only phonemes 1 to 4. Since a speech structure excludes static speaker biases, by applying the teacher's speech structure to the learner's speech sounds of phonemes 1 to 4, the learner's speech sound of phoneme 5 can be predicted [9]. Here, the learner's speech sounds of phonemes 1 to 4 are used as anchors and the teacher's speech structure is used as constraint for prediction[1] (see Figure 2). Hereafter, a teacher is called reference speaker and a learner is called target speaker. Further, phonemes 1 to 4 are called anchor phonemes and phoneme 5 is a target phoneme. In [9], this prediction problem was mathematically formulated as minimization problem. The (position of) feature distribution of phoneme 5, $\hat{p}_5$, in the target speaker's acoustic space can be obtained as

$$\hat{p}_5 = \arg\min_{p_5} \sum_{n=1}^{4} \{BD^{(t)}(p_5, p_n) - a_{5,n}^{(r)}\}^2. \quad (3)$$

p_n is the feature distribution of phoneme n in the target speaker's acoustic space. $a_{5,n}^{(r)}$ is BD between phoneme 5 and phoneme n in the reference speaker's acoustic space.

2.3. Prediction of unseen articulatory movements of a speaker using the speech structure of another speaker

We apply our previous method [9] to the current problem of predicting the articulatory movements of unseen phonemes. The unseen articulatory movements are searched for by using the speech structure of a reference speaker and the acoustic anchors of a target speaker. Different from [9], what has to be predicted, articulatory movement, is not in the same space of the anchors. By using the articulatory-to-acoustic mapping function, which was estimated in advance by using a parallel data of the target speaker, an articulatory movement $\boldsymbol{x}$ of the target speaker can be mapped to its corresponding acoustic observation (see Figure 3). Here, the result of mapping is denoted as $\mathcal{F}(\boldsymbol{x})$. Using $\boldsymbol{x}$ and $\mathcal{F}(\boldsymbol{x})$, the current problem of predicting the articulatory movement $\hat{\boldsymbol{x}}$ of unseen phoneme 5 can be formulated as

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \sum_{n=1}^{4} \left\{ BD^{(t)}\left(\mathcal{N}(\mathcal{F}(\boldsymbol{x}), \Sigma), p_n\right) - a_{5,n}^{(r)} \right\}^2 \quad (4)$$

$$\left(= \arg\min_{\boldsymbol{x}} J(\boldsymbol{x}) \right). \quad (5)$$

Here, it is assumed that a common variance and covariance matrix is shared among all the kinds of sounds and the acoustic feature distribution follows Gaussian. It should be noted that *articulatory* variable $\boldsymbol{x}$ is varied to minimize the *structural* difference between the two speakers, whose structures are calculated in the *acoustic* spaces of both speakers. Mapping between the two spaces is modeled as function $\mathcal{F}$.

We employ Gaussian mixture model-based (GMM-based) articulatory-to-acoustic mapping [4] for transform function $\mathcal{F}$. Let $\boldsymbol{x} \in \mathcal{R}^{d_x}$ and $\boldsymbol{y} \in \mathcal{R}^{d_y}$ be articulatory and acoustic parameter vectors whose dimensions are d_x and d_y, respectively.

[1] For simplicity, we explain our approach by using 4 anchors and 1 target. But it can be generalized for any anchors and any targets.

$\boldsymbol{z}$ denotes a joint vector consisting of articulatory and acoustic 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:

$$P(\boldsymbol{z}; \boldsymbol{\lambda}^{(z)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}), \quad (6)$$

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(x,x)} & \boldsymbol{\Sigma}_m^{(x,y)} \\ \boldsymbol{\Sigma}_m^{(y,x)} & \boldsymbol{\Sigma}_m^{(y,y)} \end{bmatrix}. \quad (7)$$

$\boldsymbol{\lambda}$ 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

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\lambda}^{(z)}). \quad (8)$$

The MMSE-based mapping function is represented as follows:

$$\hat{\boldsymbol{y}} = \sum_{m=1}^{M} P(m|\boldsymbol{x}, \boldsymbol{\lambda}^{(\boldsymbol{z})})(\boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} {\boldsymbol{\Sigma}_m^{(xx)}}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_m^{(x)})). \quad (9)$$

Eq.(9) is employed for transform function $\mathcal{F}$ for Eq.(4). By approximating $P(m|\boldsymbol{x}, \boldsymbol{\lambda}^{(\boldsymbol{z})})$ of Eq.(9) by a constant value, Eq.(9) can be represented as a linear transformation of $\boldsymbol{x}$ and can be integrated simply to Eq.(4).

2.4. Normalization of speech structure

The proposed method uses $a_{5,n}^{(r)}$ and their corresponding BDs of the target speaker, $a_{5,n}^{(t)}$ $(1 \leq n \leq 4)$. It is also implicitly assumed that $a_{l,m}^{(r)}$ and $a_{l,m}^{(t)}$ $(1 \leq l, m \leq 4)$ are the same. According to [10], it is known that the structural features of a speaker and those of another can vary due to dialectal variation or accent variation. This may be the case in our study. In other words, some differences could be found between $a_{l,m}^{(r)}$ and $a_{l,m}^{(t)}$. To cancel these differences, accent normalization should be introduced between the reference speaker and the target speaker. This normalization process should be introduced also to $a_{5,n}^{(r)}$.

Let $\boldsymbol{A}^{(r)} = \{a_{i,j}^{(r)}\}_{1 \leq i,j \leq 4}$ and $\boldsymbol{A}^{(t)} = \{a_{i,j}^{(t)}\}_{1 \leq i,j \leq 4}$ be the distance matrices of the reference speaker and the target speaker, respectively. As told above, some differences could be found between them. Normalization for $\boldsymbol{A}^{(r)}$ is done by modifying $\boldsymbol{A}^{(r)}$ into $\boldsymbol{S}\boldsymbol{A}^{(r)}\boldsymbol{S}$, where $\boldsymbol{S}$ is a diagonal matrix $\mathrm{diag}\{s_1, s_2, \cdots, s_4\}$. The (i, j) element of $\boldsymbol{S}\boldsymbol{A}^{(r)}\boldsymbol{S}$ is $s_i s_j a_{i,j}^{(r)}$. $\boldsymbol{S}$ is determined based on the following:

$$\hat{\boldsymbol{S}} = \arg\min_{\boldsymbol{S}} \sum_{i=1}^{3} \sum_{j=i+1}^{4} (s_i s_j a_{i,j}^{(r)} - a_{i,j}^{(t)})^2 \quad (10)$$

$$= \mathrm{diag}\{\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_4\}. \quad (11)$$

The above minimization is done by using $a_{i,j}^{(r)}$ and $a_{i,j}^{(t)}$ in the range of $1 \leq i, j \leq 4$. If this normalization is applied also to $a_{5,n}^{(r)}$, then, it is modified to $\hat{s}_n a_{5,n}^{(r)}$.

3. Experiments

3.1. Conditions

To evaluate the performance of our proposed prediction method, 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 a target speaker and the other was a reference speaker. Articulatory data were measured by an electromagnetic articulography, where its sensors were placed at 7 points of articulators (lower incisor, upper and lower lips, 3 points on a tongue, velum) in the mid-sagittal plane. The articulatory data of each point are two-dimensional data of horizontal and vertical directions. The sentences read by the two speakers are 460 sentences extracted from TIMIT. The temporally-detailed transcriptions are included.

In the experiments, we focused on eight kinds of short vowels included in MOCHA (@, a, e, i, iy, o, u, and ah). One of those vowels was used as an unseen phoneme (target phoneme) and the others were used as anchors (anchor phonemes). To predict the articulatory movements of the target phoneme, the following conditions were adopted. The speech structure among the target phoneme and the anchor phonemes was extracted from acoustic data of the reference speaker. Here, each phoneme distribution was modeled as Gaussian distribution, a variance-covariance matrix of which was a diagonal matrix. In the same way, the distribution of each anchor phoneme of the target speaker was modeled using acoustic data of the target speaker. Articulatory-to-acoustic mapping model was developed with acoustic-articulatory parallel data of the target speaker. It should be noted, however, that segments of the target phoneme were eliminated from the parallel data.

To validate the articulatory movements predicted by our proposed method, their corresponding ground truth has to be prepared. For that, we assumed the Gaussian distribution for articulatory data of the target phoneme produced by the target speaker. For validation, only the mean vector was examined.

As for conditions of acoustic analysis, 24-dimensional mel-cepstrum (C1∼C24) was used as acoustic features and articulatory data which was orthogonalized via PCA was used as articulatory features.

When searching for the target movement, we employed the steepest descent method to minimize the cost function $J(\boldsymbol{x})$ of Eq.(5). This minimization procedure requires the initial value of $\boldsymbol{x}$. In the reference speaker's acoustic space, the target phoneme and all the anchor phonemes are present and the anchor phoneme that is the closest to the target phoneme can be detected. The detected anchor phoneme is also found in the *target* speaker's *articulatory* space. The mean vector of this detected phoneme was adopted as initial vector of $\boldsymbol{x}$. Our preliminary experiments showed that, if Eq.(5) was used as it is, since Eq.(5) did not constrain the region where $\boldsymbol{x}$ can vary, the optimal $\boldsymbol{x}$ can be a vector that is found outside the region of articulatory movement. To solve this, we introduced two constraint terms to $J(\boldsymbol{x})$ as follows:

$$J'(\boldsymbol{x}) = J(\boldsymbol{x}) + \alpha_1 (\boldsymbol{x} - \boldsymbol{x}_0)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{x}_0) + \alpha_2 (\boldsymbol{x} - \boldsymbol{x}_c)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{x}_c). \quad (12)$$

Here, $\boldsymbol{x}_0$ is the initial vector, $\boldsymbol{x}_c$ is the averaged vector over all the anchor phonemes in the articulatory space of the target speaker, and $\boldsymbol{\Sigma}$ is the shared variance-covariance matrix in the target speaker's articulatory space. The second term of $J'(\boldsymbol{x})$ restricts the space for searching only to the region around the closest target phoneme and the third term restricts it only to the region around the averaged vowel. α_1 and α_2 are weighting factors to the original cost function $J(\boldsymbol{x})$.

3.2. Results

Figure 4 shows the prediction error of each target phoneme. Since MOCHA has two speakers (male and female), the prediction experiment can run in four conditions and all the four

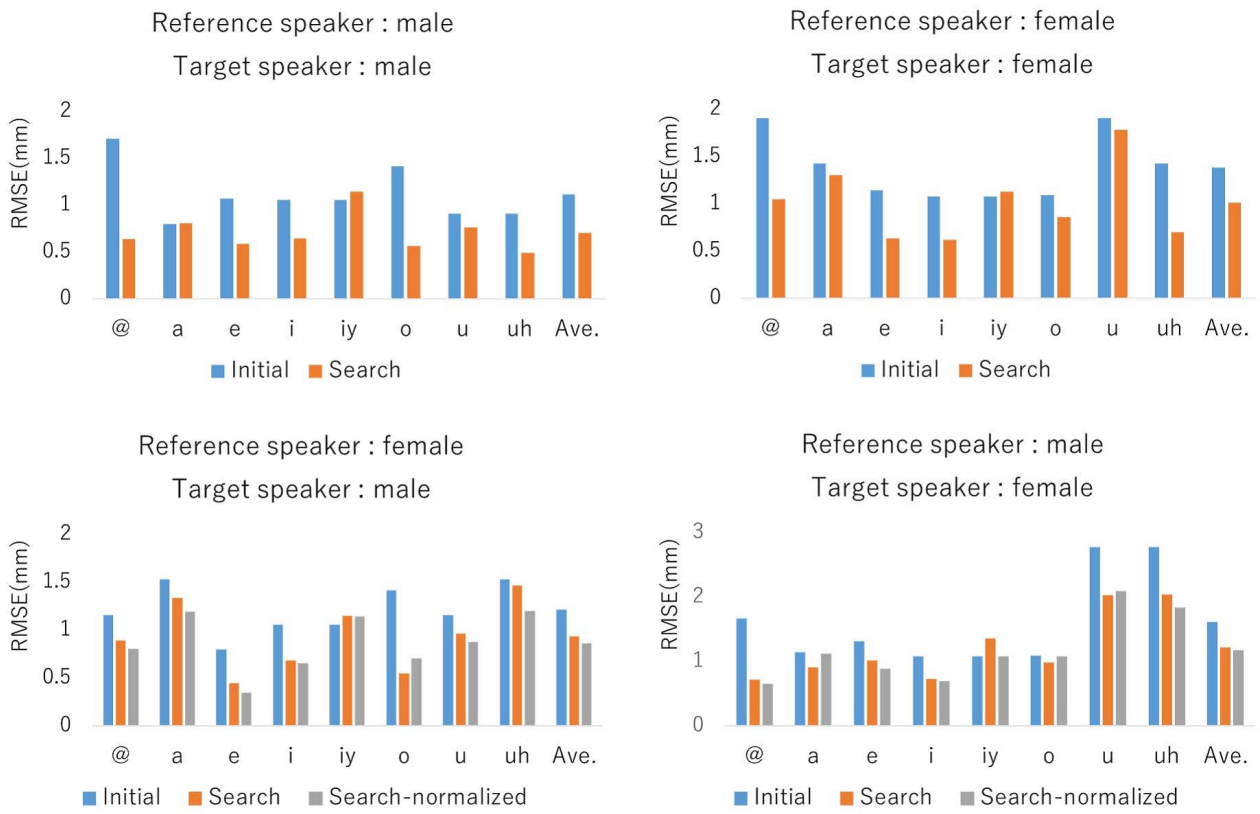

Figure 4: Prediction errors of each target phoneme in four different conditions

Table 1: RMSE(Ave.(mm))

Reference	Target	Search	Search-normalized
female	female	1.01	
male	female	1.22	1.18
male	male	0.70	
female	male	0.93	0.86

results are shown. In the figure, 'Search' shows the root-mean-square error (RMSE) between the predicted and measured articulatory movements. Here, RMSE is obtained by calculating the prediction error for each of the seven positions and averaging those errors over the seven positions. 'Ave.' means the average of RMSE for all target phonemes. For comparison, RMSEs between the initial vectors of $\boldsymbol{x}_0$ and measured articulatory movements are also shown as 'Initial'. Normalization of the speech structure was introduced only to the cases where the reference speaker and the target speaker were assigned to two different speakers. Results are shown as 'Search-normalized'.

If we focus only on Ave., we can see that prediction errors of 'Search' are lower than those of 'Initial' in all cases. These results clearly show that the predicted positions of the target phonemes in the target speaker's articulatory space is more valid rather than substituting the positions of the closest anchor phoneme.

There is one exception, that is /iy/. We can see that RMSEs of 'Search' are larger than those of 'Initial' in all cases. This is considered to be because of the third term of Eq.(12), which poses a penalty if $\boldsymbol{x}$ is located distant from the central vowel. Among the eight vowels, /iy/ is known to have the longest distance from the central vowel.

Table 1 shows the values of Ave. for each case. We can see in the column of 'Search' that prediction errors increase when the two speakers are assigned to reference and target. This will be due to the influence of difference of the speech structure between the target speaker and the reference speaker. In the column of 'Search-normalized', however, RMSEs are effectively reduced by structural normalization.

4. Conclusions

In this paper, we proposed the prediction method of the articulatory movements of unseen phonemes of a speaker using the speech structure of another speaker. In the proposal method, by using the speech structure, which is extracted from another speaker, among the unseen phonemes and other phonemes as constraint, the unseen articulatory movements are searched. To evaluate the performance of the proposal prediction method, we conducted experimental evaluations focusing on English short vowels. As a result, effectiveness of the proposed prediction method was shown. For future works, we will apply the proposed methods to continuous speech.

5. Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP15K12059.

6. 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. P1-10, 2010.

[2] 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.

[3] 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.

[4] 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.

[5] Yu Qiao and N. Minematsu, "A study on invariance of f-divergence and its application to speech recognition," *IEEE Trans. on Signal Processing,* vol.58, no.7, pp. 3884-3890, 2010

[6] N. Minematsu, S. Asakawa, and M. Suzuki, Y. Qiao, "Speech structure and its application to robust speech processing," *Journal of New Generation Computing*, vol.28, pp. 299-319, 2010

[7] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. Speech and Audio Processing,* vol. 13, pp. 930-944, 2005

[8] D. Saito, S. Asakawa, N. Minematsu and K. Hirose, " A fundamental study of structure-to-speech conversion," *TECHNICAL REPORT OF IEICE*, pp. 55-60, 2007.

[9] R. Mihara, D. Saito, N. Minematsu and K. Hirose, "Cross-speaker and cross-language voice conversion based on structural representation of speech," *TECHNICAL REPORT OF IEICE*, pp.55-60, 2009.

[10] S. Kasahara, S. Kitahara, N. Minematsu, H.-P. Shen, T. Makino, D. Saito, K. Hirose, "Improved and robust prediction of pronunciation distance for individual-basis clustering of world Englishes pronunciation," *IEEE International Conference on Acoustic, Speech and Signal Processing,* pp. 3241-3243, 2014.

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