Formant structure estimation using vocal tract length normalization for CALL systems

Yasushi Tsubota¹, Tatsuya Kawahara¹ and Masatake Dantsuji²

¹Graduate School of Informatics, Kyoto University
²Academic Center for Computing and Media Studies/Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto, 606–8501 Japan

(Received 2 September 2002, Accepted for publication 11 November 2002)

Keywords: CALL, English by Japanese students, Vowel, Formant structure estimation, VTLN

PACS number: 43.70.Kv, 43.72.Ct

1. Introduction

We are developing a system to instruct English vowel pronunciation to Japanese students. The number of English vowels is twelve, more than twice as much as that of the five vowels in Japanese, and two or more English vowels may be perceived as a single Japanese vowel by ordinary Japanese students. To evaluate the articulation of vowels, formant frequency is widely used. It must be normalized since absolute value is different for each speaker depending on the vocal tract length. Although correct samples are needed to estimate the normalization parameter, students' pronunciation of foreign language may be imperfect. Thus, we propose to estimate normalization parameter from samples of Japanese utterances and normalize English vowel with this parameter. We present two normalization methods and confirm their effect.

2. Evaluation of vowel articulation

Figure 1 is a block diagram to evaluate students’ speech and generate an instruction. To make a model of formant structure, we normalize the formant frequency of vowels in native speakers’ utterances. Then, we normalize the formant frequency of vowels in Japanese students’ speech and compare it with the model. For normalization, we use Japanese speech samples since English speech of Japanese students contain errors and are not appropriate as the samples of normalization [1].

To generate an instruction to students, we use the knowledge of articulatory phonetics. Formant frequencies are known to be related with three articulatory elements. Specifically, the first formant frequency is related with the aperture of mouth opening, the second formant frequency with the position of the tongue, and the higher formants with the amount of lip rounding. Thus, articulatory instruction is generated based on the comparison of formant structures. For formant measurement, the frames of central portion in a vowel is segmented and the smoothed spectrum of the frames is computed using PARCOR analysis. Formant frequency is measured by a peak picking method.

3. Formant structure estimation

We implement two normalization methods for the formant structure using Japanese speech by students: ML-VTLN and MMSE.

3.1. ML-VTLN

Maximum Likelihood Vocal Tract Length Normalization (ML-VTLN) is used to deal with speaker variability in speech recognition task. ML-VTLN performs spectral warping so that warped spectral feature $X'$ gives the maximum matching probability $\hat{a} = \arg \max \mathcal{P}(X'|W, A)$ for a transcription $W$ and a given acoustic model $A$. It is efficiently integrated with filter bank analysis. Spectral warping is performed in Eq. (1), where $minFreq$ and $maxFreq$ are the lower and upper frequency cutoffs in the filter bank analysis.

$$F_warp = \left\{ \begin{array}{ll}
\frac{(c_l - \alpha - minFreq)}{(cl - minFreq)} \times (F_{org} - minFreq) + minFreq & \text{if } F_{org} < cl = \frac{F_{cl} \times 2}{(1 + \alpha)} \\
\alpha \times F_{org} & \text{if } cl \leq F_{org} \leq cu = \frac{F_{cu} \times 2}{(1 + \alpha)} \\
\frac{((maxFreq - cu \times \alpha)}{(maxFreq - cu)} \times (F_{org} - cu) + \alpha \times cu & \text{if } F_{org} > cu = \frac{F_{cu} \times 2}{(1 + \alpha)}
\end{array} \right. \tag{1}
$$

The warping parameter $\alpha$ is chosen between 0.70 and 1.30 with a step of 0.02. We use the $\hat{a}$ as normalization parameter for formant frequency.

3.2. MMSE

We also implement a method based on Minimum Mean Square Error (MMSE) criterion. The warping parameter $\alpha$ is optimized by minimizing the following equation, where $M_{i,j}$ is the $j$-th mel-scale formant frequency of vowel $i$ by the speaker and $M_{i,j}$ is the mean of the formant frequency among the native speakers. In this equation, we only use the first
formant and the second formant, since they are most closely related with articulations of vowels. The method is intended to minimize the mean of the deviations for all vowels.

\[ E = \frac{\sum_{i} \sum_{j} (M_{i,j} - \bar{M}_{i,j})^2}{\# \text{vowels}} \]

(2)

4. Evaluation of formant frequency normalization

To verify the normalization methods, we collect speech samples from four bi-lingual speakers of English and Japanese. They utter words in a context of /b-V-t/ for English and /k-V-r-u/ for Japanese, where V means one of English and Japanese vowels. Speech samples are listed in Table 1. Vowel segments are aligned by the phone model. As the phone model, we use monophone HMM of 16 mixture components. In ML-VLTN, we can change the scaling function by setting lower and upper frequency bounds. An example of scaling function, where \( \alpha \) is set to 0.8, 1.0 and 1.2, is plotted in Fig. 2.

We test the warping with piecewise linear transformation with three sets of cutoff values: [100,3000], [100,2000] and [300,3000].

Distribution of estimated values of the parameter \( \alpha \) for a speaker is shown in Figs. 3–5. By using concatenated vowel segments, the normalized score was improved and the cutoff condition of warping did not affect the estimation. Thus, we use concatenated vowel segments in the following experiments.

We evaluate the formant structure estimation by the mean deviation given by Eq. (2). The result is shown in Table 2. The distribution of formant frequency before and after normalization are plotted in Figs. 6 and 7, respectively. It is

<table>
<thead>
<tr>
<th>Table 1</th>
<th>List of speech samples.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japanese</td>
<td>English</td>
</tr>
<tr>
<td>Uttered words</td>
<td>karu kiru kuru</td>
</tr>
<tr>
<td></td>
<td>keru koru</td>
</tr>
</tbody>
</table>

Fig. 1 The flow of generating instruction of articulation.

Fig. 2 Frequency scaling function. An example of scaling function, where \( \alpha \) equals 0.8, 1.0, 1.2, is shown in Fig. 2.

Fig. 3 The values of parameter \( \alpha \) are plotted against the normalized log-likelihood in Fig. 3. \( \alpha \) is estimated form English and Japanese words, respectively. Condition I means its cutoff values is [100,3000].
The values of parameter $\alpha$ are plotted against the normalized log-likelihood in Fig. 4. $\alpha$ is estimated from the concatenated vowels from English sentence. Condition 1, 2, 3 is condition of cutoff values: $[100,3000]$, $[100,2000]$ and $[300,3000]$, respectively.

The values of parameter $\alpha$ are plotted against the normalized log-likelihood in Fig. 5. $\alpha$ is estimated from the concatenated vowels from Japanese sentence. Condition 1, 2, 3 is condition of cutoff values: $[100,3000]$, $[100,2000]$ and $[300,3000]$, respectively.

### Table 2  Effect of normalization (Hz).

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>I</th>
<th>e</th>
<th>æ</th>
<th>A</th>
<th>a</th>
<th>Æ</th>
<th>U</th>
<th>U</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>before</td>
<td>91.3</td>
<td>77.6</td>
<td>75.5</td>
<td>85.9</td>
<td>82.5</td>
<td>94.8</td>
<td>86.9</td>
<td>111.8</td>
<td>164.3</td>
<td><strong>96.7</strong></td>
</tr>
<tr>
<td>ML-VTLN</td>
<td>104.4</td>
<td>79.4</td>
<td>74.6</td>
<td>84.1</td>
<td>75.0</td>
<td>86.1</td>
<td>91.3</td>
<td>92.5</td>
<td>150.2</td>
<td><strong>93.1</strong></td>
</tr>
<tr>
<td>MMSE</td>
<td>72.5</td>
<td>66.4</td>
<td>58.3</td>
<td>63.9</td>
<td>80.9</td>
<td>73.4</td>
<td>82.2</td>
<td>85.3</td>
<td>160.9</td>
<td><strong>82.3</strong></td>
</tr>
</tbody>
</table>

Standard deviations of vowels about the first and the second formant frequency are shown in Table 2. Each column means normalization methods: no normalization, ML-VTLN and MMSE.

Fig. 6 A formant chart for nine vowels without normalization is shown in Fig. 6.

Fig. 7 A formant chart for nine vowels with MMSE normalization is shown in Fig. 7.
confirmed that ML-VTLN decreases the mean of vowels deviation by 3 Hz. The MMSE method decreases the mean of vowels deviation by 14 Hz and the overlapped areas significantly decreased, which is critical in making correct instruction in a CALL system.

5. Conclusion
We have presented and compared formant structure normalization methods. ML-VTLN method was not so effective partly because it does not directly minimize the deviation and estimation is inaccurate with the short length of utterances. We confirmed that the MMSE method significantly decreases the deviation of the formant frequency for generating effective instructions of vowel articulation.

References