When acquiring a second language (L2), learners have difficulty in achieving native-like production even if they receive instruction on how to position the speech articulators for correct production. A principal reason is that learners lack information on how to modify their articulation to produce correct L2 sounds. A visual feedback method using Electromagnetic Articulography (EMA) has been previously implemented for this application with some success [Levitt et al. (2010)]. However, because this approach provided tongue tip position only, it is unsuitable for vowels and many consonants. In this work we have developed a more general EMA-based articulatory feedback system that provides real-time visual feedback of multiple head movement-corrected sensor positions, together with target articulatory positions specific to each learner. We have used this system to improve the production of the unfamiliar vowel /æ/ for Japanese learners of American English. We predicted an appropriate speaker-specific /æ/ position for each Japanese learner using a model trained on previously collected kinematic data from 49 native speakers of American English, based on vowel positions for the overlapping /iy/, /aa/, and /uw/ vowels found in both languages. Results comparing formants pre- and post-feedback training will be presented to show the efficacy of the approach.
INTRODUCTION

In recent years, developments in Computer-Assisted Language Technologies have offered new opportunities facilitating second language (L2) learning [1]. However, learners continue to have difficulty in achieving native-like production even if they receive instruction on how to correctly position the speech articulators. A principal reason is that the learners lack appropriate information on how to modify their articulation to produce correct L2 sounds. In other words, it is difficult for them to know the state of their articulators and how to position the articulators for correct sounds.

To overcome the problem, an Electromagnetic Articulographic (EMA)-based visual feedback approach has been proposed by Levitt and Katz [2]. They showed that kinematic feedback with EMA facilitated the acquisition and maintenance of the Japanese flap consonant. However, because their approach provided tongue tip position only, it is unsuitable for vowels and many consonants.

In this study, we describe a more general EMA-based articulatory feedback system that can provide real-time visual feedback of multiple sensor positions corrected for head movement, together with target articulatory positions specific to each learner. We examine the efficacy of our proposed approach by using this system to improve the production of the unfamiliar vowel /ae/ for Japanese learners of American English (AE).

EMA-BASED ARTICULATORY VISUAL FEEDBACK SYSTEM

An overview of the developed system is shown in Fig. 1. This system displays real-time multiple sensor positions on the midsagittal plane in a learner’s coordinate system by rotating the recorded data to the occlusal plane and correcting for head movement relative to reference sensors. In addition, the target articulatory positions estimated by a prediction model and the previously recorded palate shape are superimposed on the display. Figure 2 shows an example of the real-time visual feedback display (x-axis: posterior-anterior direction; y-axis: inferior-superior direction), where the red color indicates the target articulatory positions, the cyan the actual sensor positions, the yellow the palate shape. Here the tongue surface contour is calculated by spline interpolation.

The prediction model was constructed from the acoustic and kinematic data of 49 native AE speakers (24 males and 25 females) from the University of Wisconsin X-ray microbeam (XRMB) speech production corpus [3] and our EMA corpus. Specifically, in order to predict an appropriate speaker-specific /ae/ position for each Japanese learner, we developed multiple linear regression models with stepwise selection using the F1 and F2 values and the x and y coordinate values of the tongue tip (TT), tongue blade (TB), tongue dorsum (TD), lower incisor...
(LI), upper lip (UL), and lower lip (LL) for the overlapping vowels as a predictor, based on vowel positions for the overlapping /aa/, /iy/, and /uw/ vowels found in both languages.

**METHODS**

The subjects were 3 male (A, B, C) and 2 female (D, E) native speakers of Japanese, aged between 22 and 35, with no self-reported hearing deficits or speech disorders. Six sensors were placed on the articulator positions TT, TB, TD, LI, UL, and LL for the articulatory visual feedback, and 4 additional sensors (the upper incisor, bridge of the nose, left and right mastoid processes behind the ears) were used as references to correct for head movement (Fig. 3). Articulatory movement and speech data were recorded simultaneously at sampling rates of 200 Hz and 16 kHz, respectively, using 3D EMA (Carstens AG500). In the real-time articulatory visual feedback process, all position data were processed at 20 Hz. For alignment across subjects, the occlusal plane was estimated using a biteplate with 3 additional sensors.

<table>
<thead>
<tr>
<th>label</th>
<th>location</th>
</tr>
</thead>
<tbody>
<tr>
<td>TT</td>
<td>tongue tip</td>
</tr>
<tr>
<td>TB</td>
<td>tongue blade</td>
</tr>
<tr>
<td>TD</td>
<td>tongue dorsum</td>
</tr>
<tr>
<td>LI</td>
<td>lower incisor</td>
</tr>
<tr>
<td>UL</td>
<td>upper lip</td>
</tr>
<tr>
<td>LL</td>
<td>lower lip</td>
</tr>
<tr>
<td>Ref</td>
<td>nose bridge</td>
</tr>
<tr>
<td></td>
<td>right and left ears</td>
</tr>
<tr>
<td></td>
<td>upper incisor</td>
</tr>
</tbody>
</table>

The experimental procedure consisted of four phases: preparation, pre-test, training, and post-test. In the preparation phase, articulatory positions and speech data were collected for prediction of the /ae/ position when the subject produced the isolated Japanese vowels /aa/, /iy/, and /uw/. The palatal shape and occlusal plane were also recorded. In the pre-test and post-test
phases, the subject was asked to produce a sustained vowel /ae/ and 3 CVC words (‘back’, ‘sad’, and ‘had’) in response to the speech sound of an AE speaker of the same gender as the subject, which was selected from the XRMB corpus. Each stimulus was presented 5 times in random order. Only in post-test phase, we instructed the subjects to reproduce the articulatory movement that was learned in the training phase. During the training phase, the real-time articulatory visual feedback and articulatory target for /ae/ were presented on a display. The subject was first asked to try to fit his/her tongue contour, UL, LL, and LI to the articulatory target without speech production for about 5 minutes while watching the display screen as in Fig. 2. Then, the subject was asked to match the articulators with the target as closely as possible on the display and then to produce the vowel /ae/, following the speech sound of the same speaker as the test phase. This task was repeated 20 times.

In order to assess the effect of articulatory training, the first (F1) and second (F2) formants of the vowel /ae/ were extracted from the acoustic measurements for pre-test and post-test phases using a 16th order LPC analysis over a 25 ms window with a 15 ms overlap, and averaged over 5 frames centered at the middle frame of a spectrally stable part of the /ae/ segment for each utterance. Taking the human hearing system and the anisotropy of the F1-F2 space into account, all formant values were converted to the Equivalent Rectangular Bandwidth (ERB) scale defined by

$$E = 21.4 \log_{10}(4.37f + 1),$$

where $E$ represents the number of ERBs and $f$ is frequency in kHz [4].

**RESULTS**

Figure 4 shows the distribution of the produced /ae/ sounds in the F1-F2 space for each subject before and after training, in which a purple ellipse represents the 95% confidence limits for the /ae/ distribution for the same gender AE speakers obtained from the XRMB and EMA corpus, and vertical and horizontal lines indicate the median of the F1 and F2 values of the native /ae/ distribution, respectively. From this figure, a tendency can be seen that the produced /ae/ sounds are distributed closer to the center of the native /ae/ distribution after articulatory training.

To statistically evaluate differences between pre-test and post-test, the median of the native /ae/ distribution was defined as a reference sound and the Euclidean distances between the produced and reference sounds were calculated in the F1-F2, F1, and F2 spaces, respectively. Figure 5 compares the average distance of the pre-test and post-test for all subjects, where error bars represent standard error of the mean ($n = 20$). A one-way within-subjects repeated measures ANOVA revealed a significant difference between pre-test and post-test in the F1-F2 ($F(1, 4) = 7.964, p < .05$) and F2 ($F(1, 4) = 17.650, p < .05$) spaces. This result shows that articulatory training using our system facilitates improvements in speech production learning of a non-native vowel /ae/ for Japanese learners. The reason for no improvement in F1 may be because there is little acoustic difference between Japanese vowel /aa/ and non-native vowel /ae/ for this formant.

**CONCLUDING REMARKS**

We have developed a more general EMA-based articulatory visual feedback system and demonstrated that in speech production learning of a non-native vowel /ae/, our system can help Japanese learners improve their speech production. This result suggests that this system might be useful in facilitating L2 speech production learning in other language contexts if there are some overlapping vowels between the native language and L2. Moreover, this system might be
FIGURE 4: Scatterplots of utterances for all subjects in the F1-F2 space (ERB scale).

FIGURE 5: Distribution of the Euclidean distance between the produced and reference sounds at the pre-test and post-test phases in (a) F1-F2, (b) F1, and (c) F2 spaces.

applicable to the rehabilitation of patients after oral surgery if speech therapists are able to provide support in terms of articulatory knowledge.

To further validate the efficacy of this system, it is necessary to investigate relationships between the achieved articulatory positions and target positions estimated by the prediction model. From a communication point of view, perceptual evaluation by native AE speakers will also be explored. The application of the proposed approach to other vowels and consonants as well as other languages is left for future work.
ACKNOWLEDGMENTS

This work was partly supported by the Institutional Program for Young Researcher Overseas Visits from the Japan Society for the Promotion of Science (JSPS).

REFERENCES


