Estimation of the movement trajectories of non-cruclal articulators based on the detection of crucial moments and physiological constraints

Jangwon Kim, Sungbok Lee and Shrikanth S. Narayanan

Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, USA
jangwon@usc.edu, sungbokl@usc.edu, shri@sipi.usc.edu

Abstract
This study develops a mathematical model that estimates the movements of (linguistically) non-cruclal articulators in speech production, which provides a systematic way to study the relationship between the behaviors of crucial and non-cruclal articulators; crucial articulators are those essential for realizing a speech task. The underlying assumption of our model is that non-cruclal articulatory movements are governed by the psychological constraints in relation to the corresponding crucial articulators as well as by the contextual constraint from the nearest crucial time of the non-cruclal articulator. These constraints have been generally assumed in the speech production literature, but they have not been incorporated directly into articulatory models. The crucial articulatory moments in an utterance are automatically determined by a novel forced-alignment algorithm for articulatory trajectories, which uses the inherent physical properties of crucial articulatory movements. Experimental results suggest that the proposed algorithm is capable of estimating non-cruclal articulatory positions well in both neutral and emotional speech, significantly better than the simple interpolation of crucial points.

Index Terms: non-cruclal articulators, articulatory modeling, emotional speech

1. Introduction
Speech articulators can be categorized as crucial and non-cruclal articulators, based on their relative levels of linguistic contribution and relevance for producing target speech sounds. For example, the movement of the tongue tip is more crucial than that of the lower lip for producing the /t/ sound. Since a speech utterance consists of a series of phones, the crucial articulators change over time depending on the lexical content of the utterance. For the binary crucial/non-cruclal categorization of articulators, the movements of crucial articulators are governed by the present phonemic target (linguistic constraint), resulting in less postural variability than non-cruclal articulators [1, 2, 3]. However, the constraints and control factors related to the movements of the non-cruclal articulators still remains as open research questions. The hypothesized constraint factors in the articulatory control can be many, e.g., physiological, as well as linguistic gestural and contextual constraints.

The present study aims to develop a model for the non-cruclal articulatory trajectories from the perspective that they are not directly controlled, but indirectly by the controls of the crucial articulators and the vocal tract constraints. The motivation for this work is to examine the validity of this assumption, and eventually to understand the control mechanism of non-cruclal articulators better by improving this preliminary model in the present paper. Each trajectory of the non-cruclal articulator is modeled as a function of (i) the position of the non-cruclal articulator at the previous and following crucial time points of the articulator, and time distance to the crucial time points, and (ii) the positions of the present crucial articulators. The factor (i) corresponds to the contextual constraint, the degree of which changes as a function of time distance between the present time and the nearest crucial time point. The factor (ii) corresponds to the physiological constraint which is assumed to be static over time. For example, the physiological constraint between the lower lip and the jaw is high, because the lower lip is anchored to the jaw (mandible). The linguistic gestural constraint refers to the coordinated articulatory actions for producing lexical units [4, 5, 6]. The linguistic gestural constraint is another important factor, although it is not considered in this study.

For training our model, it is required that the crucial articulatory points for each phone in an utterance are pre-selected. Anamthakrishnan and Engwall have proposed an algorithmic way of determining crucial points in the articulatory trajectory based on their physical properties [7], although the number of resulting crucial points can be many or none for each phone. The algorithm is built based on the assumption that crucial points exhibit relatively greater change in the movement direction and smaller speed than the other points. Also, Kato et al. reported that the time point of the constriction formation of the crucial articulator corresponds to the moment of the local minima of velocity and of the local maxima of acceleration [8]. This indicates that the constriction formation point shows the change of movement direction in short time and low articulatory speed. Motivated by these previous studies, the present paper proposes a novel forced-alignment algorithm for determining the best crucial points, i.e., the time points of formation of articulatory constriction and largest opening, in articulatory trajectories, at most one for each crucial articulator in each phone. It should be noted that this algorithm does not require any phonetic label on the articulatory trajectories for training any model, because it determines the optimal time points by solving the optimization problem at utterance-level trajectory.

2. Dataset
An electromagnetic articulography dataset collected at the University of Southern California with the NDI WA VE system is used for the experiments. The dataset contains both articulatory data and simultaneously recorded speech waveform. The articulatory data comprise 3-dimensional coordinates of six sensors that were attached near the tongue tip (TT), the tongue blade (TB), the tongue dorsum (TD), the upper lip (UL), the lower lip (LL) and the lower incisor (JAW) of a female native speaker of American English.

The speaker was asked to start speaking after immersing herself into one of the five emotions – neutrality, anger, happiness, sadness and fear. The stimuli consists of eight sentences:
Table 1: The number of utterances for each emotion label.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Neutrality</th>
<th>Anger</th>
<th>Happiness</th>
<th>Sadness</th>
<th>Fear</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40</td>
<td>41</td>
<td>32</td>
<td>42</td>
<td>45</td>
<td>200</td>
</tr>
</tbody>
</table>

(1) Nine one five, two six nine, five one six two; (2) Ma ma ma, ma ma ma, ma ma ma ma; (3) John bought five black cats at the store; (4) The leopard, skunk, and peacock are wild animals; (5) Charlie did you think to measure the tree?; (6) The queen said to the king, "I want to go to the garden store; (4) The leopard, skunk, and peacock are wild animals; (5) Dickory dock.

These sentences were repeated five times in a randomized order. The sampling rates of the articulatory data and speech audios are 100 Hz and 22050 Hz, respectively.

The articulatory data is aligned to the occlusal plane of the speaker, followed by interpolation for missing articulatory frames. Each sensor trajectory was smoothed by a 9th-order Butterworth low pass filter with a cutoff frequency of 20 Hz. Only the position data in the horizontal direction (denoted by X) and the vertical direction (denoted by Y) are used in this study. Each dimension of articulatory data of each emotion is scaled to the range of [0,1] for the fair evaluation. The initial and final silence region in each utterance is excluded for analysis.

Each utterance has the emotion label that was obtained by a perceptual evaluation of the emotion quality by 11 native speakers of American English. The emotion label is determined by majority voting. Table 1 shows the number of utterances for each emotion.

3. Crucial/Non-crucial articulators

In the literature, crucial articulators for each phone have been determined by data-driven approaches, e.g., based on the postural variability metric [9] or the conduction degree metric [10], or speech-production theoretical approaches, e.g., using articulatory phonology [5,6] and Task dynamics [11,12]. The present study follows the theoretic approach, and adopts the description of crucial articulators involved in each task variable from articulatory phonology. For more detailed description of the crucial articulators for each phone in American English, see the manual of the Task Dynamics model [11] and "Fig.1" in [5]. Although the number of crucial articulators for one phone can be more than one, only one (primary) articulator is used for forced alignment for algorithmic simplicity and better alignment accuracy. The crucial time points of the other crucial articulators can be found easily given the crucial time points of the primary articulators. Jaw is selected as the primary crucial articulator for vowels, because the vertical movement of the jaw reflects the consonant-vowel-consonant transition nicely, compared to other articulators, such as the tongue blade, the tongue dorsum, and the lips. Table 2 shows the list of the primary crucial articulator of each consonant defined in this study.

4. Forced alignment of crucial points

This section discusses an algorithm we propose, by which the sequence of crucial points for each phone in an utterance is aligned on the articulatory trajectories. This algorithm uses conduction score $CS_i(t)$ and opening score $OS_i(t)$ driven from the physical properties inherent in the articulatory movements.

Assume $N$ is the number of frames of an utterance, $X_i = [x_i(1), x_i(2), ..., x_i(N)]$ is the sequence of the 2-dimensional position vectors of the $i$-th articulator for the utterance, where $x_i(t)$ is the position vector at time $t$. $θ_i = [θ_i(1), θ_i(2), ..., θ_i(N)]$ is the sequence of the angles of the $i$-th articulator, where $θ_i(t)$ is the acute angle of three points

Table 2: The selection of the crucial articulators for each consonant in the dataset. "Phn" is phone, "Arti" is articulator.

<table>
<thead>
<tr>
<th>Phn</th>
<th>B</th>
<th>CH</th>
<th>D</th>
<th>DH</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>JH</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arti</td>
<td>LL</td>
<td>ID</td>
<td>1T</td>
<td>1T</td>
<td>ID</td>
<td>LL</td>
<td>ID</td>
<td>None</td>
<td>ID</td>
<td>ID</td>
<td>1T</td>
<td>LL</td>
</tr>
</tbody>
</table>

$x_i(t−T), x_i(t), x_i(t+T)$. $T$ is a time lapse (30 msec is used as a default). $S_i = [S_i(1), S_i(2), ..., S_i(N)]$ is the sequence of the tangential speed of the $i$-th articulator, where $S_i(t)$ is the tangential speed at time $t$. $C_i(t)$ is the tangential score of $i$-th articulator at time $t$, computed as follows:

$$C_i(t) = \frac{θ_i(t) - min(θ_i)}{max(θ_i) - min(θ_i)}$$

$C_i(t)$ is similar to the "importance" score in a previous study [7], but it is slightly modified for the purpose of the forced-alignment in our experimental setup.

Let $x_{i}^p$ be the vertical position of $x_i$, $t_{i}^p$ and $t_{i}^o$ denote the nearest preceding and following local extrema time points from $t$, respectively. Also, $Y_{max}(t)$ and $Y_{max}(t)$ denotes max($x_{i}^p(t_{i}^p)$), $x_{i}^p(t_{i}^o)$, and min($x_{i}^p(t_{i}^p)$), $x_{i}^p(t_{i}^o)$, respectively. $E_i(t)$ is the normalized local excursion score of the $i$-th articulator, which reflect the degree of articulatory opening between the two extrema points:

$$E_i(t) = \frac{x_{i}^p(t) - Y_{min}(t)}{Y_{max}(t) - Y_{min}(t)}$$

Finally, $CS_i(t)$ and $OS_i(t)$ are represented as functions of $C_i(t)$ and $E_i(t)$ as follows.

$$CS_i(t) = C_i(t) \times |E_i(t) - 1|$$

$$OS_i(t) = C_i(t) \times E_i(t)$$

Finally, the optimal crucial time points, one point for each phone, are determined by maximizing the sum of $CS_i(t)$ or $OS_i(t)$ for each phone using the Viterbi algorithm. We assume that the crucial time points are located at the local maxima of $CS_i(t)$ for consonants (closure/constriction gesture) and the local maxima of $OS_i(t)$ for vowels (opening gesture). So, $CS_i(t)$ and $OS_i(t)$ are chosen for consonants and vowels, respectively. In order to prevent alignment error due to short pauses or speech production error, $CS_i(t)$ and $OS_i(t)$ are loosely weighted based on acoustic phonetic labels obtained using an adaptive speech-text alignment tool, SailAlign [7]. For the weighting, $CS_i(t)$ or $OS_i(t)$ is multiplied by a trapezoid window. The center of the window is at the middle of the phone boundaries; The top line is 3 times of phone duration. In the Viterbi decoding, the state transition is one-way (left-to-right) and uniformly weighted. The object likelihood for each phone is weighted loosely based on acoustic phonetic labels obtained using an adaptive speech-text alignment tool, SailAlign [7].

In Figure 1, the local maxima of $CS_i(t)$ and $OS_i(t)$ represent the crucial time points of each articulator well.

5. Estimation of the trajectories of non-crucial articulators

5.1. Model description

Let $f_i(t)$ and $F_i(t)$ be the true and estimated trajectory of $i$-th (non-crucial) articulator at time $t$, respectively. The optimal $\hat{f}_i(t)$ is found by minimizing $\mathcal{J}$, where $\mathcal{J}$ is:

$$\mathcal{J} = \sum_{t=1}^{M} |f_i(t) - \hat{f}_i(t)|^2$$
where $M$ is the number of articulatory frames (in the train set). $t_c$ denotes the nearest crucial time point from time $t$ for the $i$-th articulator. $\tilde{f}_i(t_c)$ is the position of the $i$-articulator at the nearest crucial time point. $\tilde{f}_i(t)$ denotes (estimated) physiologically constrained motion of the $i$-articulator. Then, $\hat{f}_i(t)$ is modeled by convex combination of $f_i(t_c)$ and $\tilde{f}_i(t)$, as follows:

$$\hat{f}_i(t) = f_i(t_c)K_i(t)+\tilde{f}_i(t)(1-K_i(t)) \tag{6}$$

$K_i(t)$ is a weighting function on the contextual constrained motion. We consider a bounded linear kernel function and a sigmoid kernel function in this study:

$$K_i(t) = \frac{\max(0, \min(1, (\eta \times \lambda_i(t) + \xi)))}{1 + \exp(-\eta(\lambda_i(t) - \xi))} \tag{7}$$

where $\eta$ and $\xi$ are hyper-parameters and $\lambda_i(t)$ represents the time-varying influence from the crucial time points of $i$-th articulator. $T_{CP}$ denotes the set of all crucial time points of all articulators in the utterance. Let $t_p$ and $t_f$ be the preceding and following crucial time point from $t$ in $T_{CP}$. Also, let $t'=\{k; k = \arg\min|k_l - t_i|, k_l \in \{t_p, t_f\}\}$. Then, $\lambda_i(t)$ is:

$$\lambda_i(t) = \begin{cases} \frac{t - t_p}{t_p - t_f} & \text{if } t_p < t_f \text{ is the point of } i\text{-th sensor} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $t_j$ is the one of non-$i$-th articulator between $t_p$ and $t_f$. Note that $\lambda_i(t) \in [0, 1]$ and $K_i(t) \in [0, 1]$ in this model.

$\tilde{f}_i(t)$ is modeled simply by a linear regression of the positions of all crucial articulators at time $t$ in the present study, although other modeling methods can be applied. An underlying assumption is that the effect of the physiological constraint among articulators can be represented by a linear transformation. For example, the physiological influence from the position of the jaw to the movement of the lower lip can be computed by rotation and translation, i.e., linear transformation. In this study, $f_i(t)$ is mathematically represented as the following:

$$\tilde{f}_i(t) = \frac{N_C(t)}{N_i(t)} \sum_{i} (\alpha_{i}f_i(t_i)) + \beta_i \tag{9}$$

where $N_C(t)$ is the number of the crucial articulators, except the $i$-th articulator, at $t$; $\alpha_{i}$ and $\beta_i$ are the coefficients of the linear regression model; $f_i(t)$ is the position of the $i$-th (crucial) articulator at time $t$. Note that the crucial articulators’ data used for representing $f_i(t)$ do not include the data of the $i$-th articulator itself.

Figure 2: An example plot of crucial time points (non-zero values) for each articulatory trajectory (the left panel) and crucial articulatory data used for computing $\tilde{f}_i(t)$ (the right panel). Red cross marks in the right panel denote the crucial points.

5.2. Experimental setup

Our model is trained in leave-one-utterance-out setup for each emotion and each combination set of crucial articulators, except the estimating articulator, because $\alpha_{i,j}$ and $\beta_i$ of $f_i(t)$ in (9) depends on the combination. Although leave-one-sentence-out setup can generalize the modeling power better across different sentences, this setup cannot be used in our dataset. The reason is that the combination of crucial articulators of 7 sentences do not contain all combination of crucial articulators of the remaining sentence. The best hyper-parameters for each kernel function are obtained on the test data, by which the nature of the weighting time function for the contextual constraint effect can be studied. The estimation performance of our model is evaluated in terms of the mean of the root-mean-squared-error (RMSE), denoted by $ERM_{RMSE}$, and the mean of the correlation coefficient, denoted by $ECORR$, between the true trajectory and the estimated trajectory of all utterances. The linear interpolation, denoted by $I_{Linear}$, and the spline interpolation, denoted by $I_{Spline}$, are tested as the baseline systems. Only for the baseline systems, the data of the initial and final frames are included on top of data of only crucial points (in the left-most panel in Figure 2) in order to perform interpolation over all frames. The estimation performance of each system with the best hyper-parameters. The best performance is achieved by the sigmoid kernel function with all six articulators’ data ($ERM_{RMSE} = 0.07$, and $ECORR = 0.87$), although the result of the best bounded linear function is very close to that. Figure 3 shows an example of the estimated vertical trajectory when the best sigmoid function with all articulators’ data is used. The sentence is “nine one five, two six nine, five one six two.”

5.3. Results and Discussion

Table 3 shows the estimation performance of each system with the best hyper-parameters. The best performance is achieved by the sigmoid kernel function with all six articulators’ data ($ERM_{RMSE} = 0.07$, and $ECORR = 0.87$), although the result of the best bounded linear function is very close to that. Figure 3 shows an example of the estimated vertical trajectory when the best sigmoid function with all articulators’ data is used. The sentence is “nine one five, two six nine, five one six two.”
For more detailed analysis, we computed RMSE and the correlation coefficient for each dimension in terms of RMSE. Table 4 shows the result. Overall, the RMSEs of TT, TD and JAW are smaller when all articulatory data are used (#Arti = 6) than when only primary crucial articulatory data are used (#Arti = 4). It is speculated that using the position of the tongue blade improves the estimation accuracy for the tongue tip and the tongue dorsum because of their highly correlated motions overall. However, this is not the case for the lower lip. It seems that the information of the upper lip does not improve the estimation accuracy for the lower lip in the modeling framework of this paper. One possible reason is that the upper lip and the lower lip are not anatomically constrained, but constrained by the linguistic gesture. The current model does not consider the gestural difference, i.e., different coordinated actions of the lips depending on phone. For example, the movements of the upper lip and the lower lip are highly (negatively) associated for bilinguals, while they are not for labio-dentals. The worst estimation performance is shown for the upper lip, presumably due to the lack of anatomical constraints of the upper lip to other sensors.

Finally, we compared the estimation performance among each emotion’s data to evaluate its estimation accuracy in emotional speech. Table 5 shows $E_{RMSE}$ and $E_{CORR}$ of the best kernel function for each emotion. In terms of both $E_{RMSE}$ and $E_{CORR}$, the estimation accuracy is better when all articulatory data are used than when only primary crucial articulatory data are used. For the case of all data used, the estimation accuracy is not much different among different emotions, suggesting that the estimation performance of our model is not degraded significantly by emotion. For the case of four articulators’ data, fear shows the worse estimation accuracy in terms of both $E_{RMSE}$ and $E_{CORR}$. Recall that the range of each articulatory dimension is normalized to $[0,1]$ for each emotion data. This result suggests that additional control factors (other than the physiological constraints and contextual constraints of our model with the four articulators) may play a more important role to control non-crucial articulators for fear than for the other emotions.

### 6. Conclusions and future work

The present study introduces an algorithm for estimating the trajectories of linguistically non-crucial points by using the physiological and contextual constraints on the crucial point data. Experimental results suggest that a simple interpolation may not be good enough for estimating the trajectories of the non-crucial points. Results also suggest that the proposed algorithm is capable of estimating the trajectories of the non-crucial points well, considerably better than the simple interpolation methods. The estimation accuracy is also consistent for different emotion when all sensors’ data are used.

The proposed model in this work has rooms for improvement. As mentioned in the Introduction, linguistic constraints are not considered in the model for better estimation, especially for the upper lip (and the lower lip). Different articulatory gestures, which can affect the relationship between the motions of the crucial and non-crucial articulators, should be considered in the modeling framework. Evaluation on a larger and phonetically balanced dataset is needed for such a model. Also, in order to generalize the estimation power of the model, a larger dataset with redundant lexical data is needed. Finally, incorporating the physical constraint among non-crucial articulators in their estimation process can also be useful.

### 7. Acknowledgements

This work was supported by NSF IB-1116076 and NIH DC007124. Special thanks to Mary Francis for her devotion and help in all SAIL research efforts.

### Table 3: The evaluation results of the estimated non-crucial articulatory trajectories. ‘STD’ is standard deviation. ‘B-Linear’ and ‘Sigmoid’ denote the bounded linear kernel and the sigmoid kernel, respectively. ‘#Arti’ denotes the number of articulators whose data are used for the experiment.

<table>
<thead>
<tr>
<th>#Arti</th>
<th>System</th>
<th>RMSE</th>
<th>CORR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>STD</td>
<td>ζ</td>
</tr>
<tr>
<td>4</td>
<td>Linear</td>
<td>0.19</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>Spline</td>
<td>0.25</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>B-Linear</td>
<td>0.09</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Sigmoid</td>
<td>0.09</td>
<td>0.02</td>
</tr>
<tr>
<td>6</td>
<td>Linear</td>
<td>0.13</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>Spline</td>
<td>0.19</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>B-Linear</td>
<td>0.08</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>Sigmoid</td>
<td>0.07</td>
<td>0.04</td>
</tr>
</tbody>
</table>

### Table 5: The evaluation results with the best kernel function in terms of $E_{RMSE}$ and $E_{CORR}$ for all utterances. ‘#Arti’ denotes the number of articulators used.

<table>
<thead>
<tr>
<th>#Arti</th>
<th>Neutrality</th>
<th>Anger</th>
<th>Happiness</th>
<th>Sadness</th>
<th>Fear</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>E_{RMSE}</td>
<td>0.78</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>E_{CORR}</td>
<td>0.63</td>
<td>0.58</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td>6</td>
<td>E_{RMSE}</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>E_{CORR}</td>
<td>0.87</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
</tr>
</tbody>
</table>
8. References


