Estimation of the movement trajectories of non-crucial 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-crucial articulators in speech production, which provides a systematic way to study the relationship between the behaviors of crucial and non-crucial articulators; crucial articulators are those essential for realizing a speech task. The underlying assumption of our model is that non-crucial articulatory movements are governed by the physiological constraints in relation to the corresponding crucial articulators as well as by the contextual constraint from the nearest crucial time of the non-crucial 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-crucial articulatory positions well in both neutral and emotional speech, significantly better than the simple interpolation of crucial points.

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

1. Introduction

Speech articulators can be categorized as crucial and non-crucial 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-crucial categorization of articulators, the movements of crucial articulators are governed by the present phonemic target (linguistic constraint), resulting in less postural variability than non-crucial articulators [1, 2, 3]. However, the constraints and control factors related to the movements of the non-crucial 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-crucial 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-crucial articulators better by improving this preliminary model in the present paper. Each trajectory of the non-crucial articulator is modeled as a function of (i) the position of the non-crucial 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. Anannthakrishan 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 WAVE 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.*

Neutrality	Anger	Happiness	Sadness	Fear	Total
40	41	32	42	45	200

(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 the knight is a monster; (7) Pam said bat that fat cat at that mat; (8) Hickory dickory dock, the mouse ran up the clock, hickory 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 constriction degree metric [10], or speech-production theoretic 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 constriction 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. $X_i = [x_i(1), x_i(2), \cdots, x_i(N)]$ 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. $\theta_i = [\theta_i(1), \theta_i(2), \cdots, \theta_i(N)]$ is the sequence of the angles of the i-th articulator, where $\theta_i(t)$ is the acute angle of three points $[x_i(t-T), x_i(t), x_i(t+T)]$. T is a time lapse (30 msec is used as a default). $S_i = [S_i(1), S_i(2), \cdots, S_i(N)]$ 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 cruciality score of i-th articulator at time t, computed as follows:

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

Phn	B	CH	D	DH	F	G	H	JH	K	L	M	N
Arti	LL	TD	TT	TT	LL	TD	None	TD	TD	TT	LL	TD
Phn	NG	P	R	S	SH	T	TH	V	W	Y	Z	ZH
Arti	TD	LL	TT	TT	TD	TT	TT	TD	LL	TD	TT	TD

$$C_i(t) = -\frac{\theta_i(t) - \min(\theta_i)}{\max(\theta_i) - \min(\theta_i)} - \frac{S_i(t) - \min(S_i)}{\max(S_i) - \min(S_i)} \quad (1)$$

$C_i(t)$ 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^v be the vertical position of x_i. t_i^p and t_i^f denote the nearest preceding and following local extrema time points from t in x_i^v, respectively. Also, $Y_{max}(t)$ and $Y_{min}(t)$ denotes $\max(x_i^v(t_i^p), x_i^v(t_i^f))$ and $\min(x_i^v(t_i^p), x_i^v(t_i^f))$, 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:

$$E_i(t) = \frac{x_i^v(t) - Y_{min}(t)}{Y_{max}(t) - Y_{min}(t)} \quad (2)$$

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

$$CS_i(t) = C_i(t) \times |E_i(t) - 1| \quad (3)$$

$$OS_i(t) = C_i(t) \times E_i(t) \quad (4)$$

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 [**?**]. 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; The bottom line is 5 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 $CS_i(t)$ (for consonants) or weighted $OS_i(t)$ (for vowels). Figure 1 shows $CS_i(t)$ and $OS_i(t)$ for the phrase "nine one five." 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 $\hat{f}_i(t)$ 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:

$$\mathcal{J} = \sum_{t=1}^{M} \left| f_i(t) - \hat{f}_i(t) \right|^2 \quad (5)$$

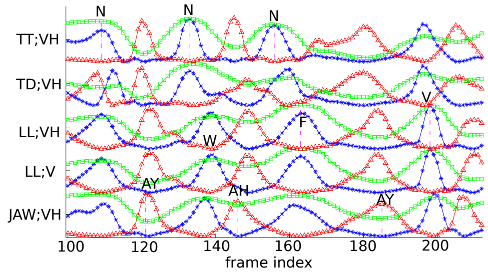

Figure 1: $OS_i(t)$ (red triangle line), $CS_i(t)$ (blue asterisk line) and vertical trajectory (green square line) of crucial articulators for "nine one five." Vertical dash-dot line (magenta) indicates the aligned crucial time point for each phone. 'VH' indicates that both vertical and horizontal trajectories are used for computing $CS_i(t)$ and $OS_i(t)$. 'V' indicates that only vertical trajectory is used.

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. $f_i(t_c)$ is the position of the i-articulator at the nearest crucial time point. $\hat{f}_i^p(t)$ 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 $\hat{f}_i^p(t)$, as follows:

$$\hat{f}_i(t) = f_i(t_c)K_i(t) + \hat{f}_i^p(t)(1 - K_i(t)) \quad (6)$$

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

$$K_i(t) = \begin{cases} \max(0, \min(1, (\eta \times \lambda_i(t) + \xi))) \\ \dfrac{1}{1 + \exp(-\eta(\lambda_i(t) - \xi))} \end{cases} \quad (7)$$

where η and ξ 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 $t' = \{k; k = \operatorname{argmin}|k' - t|, k' \in \{t_p, t_f\}\}$. Then, $\lambda_i(t)$ is:

$$\lambda_i(t) = \begin{cases} \left|\dfrac{t - t_j}{t_p - t_f}\right| & \text{if } t_p \text{ or } t_f \text{ is the point of } i\text{-th sensor} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

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 $K(t) \in [0, 1]$ in this model.

$\hat{f}_i^p(t)$ 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, $\hat{f}_i^p(t)$ is mathematically represented as the following:

$$\hat{f}_i^p(t) = \sum_{\substack{l=1 \\ l \neq i}}^{N_C(t)} (\alpha_{i,l} f_l(t)) + \beta_i \quad (9)$$

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

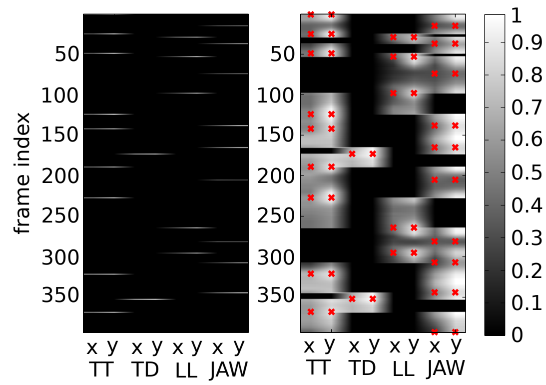

Figure 2: An example plot of crucial time points (non-zero values) for each articulatory trajectory (the left panel) and crucial articulatory data used for computing $\hat{f}_i^p(t)$ (the right panel). Red cross marks in the right panel denote the crucial points.

Figure 2 illustrates the crucial time points (the left panel) and the crucial articulatory data used for training the model. The sentence is "nine one five, two six nine, five one six two." The algorithm estimates all zero-valued regions in the left panel by using (i) positions and times of the non-zero samples of the estimating articulatory trajectory in the left panel, and (ii) the positions of the non-zero samples in the right panel, except of the estimating articulator.

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 $\alpha_{i,l}$ and β_i of f_i^p in (9) depends on the combination. Although leave-one-sentence-out setup can generalize the modeling power better across different sentences, this setup can not 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 E_{RMSE}, and the mean of the correlation coefficient, denoted by E_{CORR}, 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 of non-crucial articulatory trajectory is performed in two ways: using only four primary crucial articulators's data or using all six articulators' data.

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 (E_{RMSE} = 0.07, and E_{CORR} = 0.87), although the result of the best bounded linear function is very close to that. Figure 3 shows an example of the estiamted vertical trajectory when the best sigmoid function with all articuators' data is used. The sentence is "nine one five, two six nine, five one six two."

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

		RMSE				CORR			
#Arti	System	Mean	STD	η	ξ	Mean	STD	η	ξ
4	I_{Linear}	0.19	0.06			0.56	0.25		
	I_{Spline}	0.25	0.09			0.45	0.26		
	B-Linear	0.09	0.02	1.2	-0.35	0.78	0.09	1.2	-0.14
	Sigmoid	0.09	0.02	6	0.6	0.78	0.09	8	0.5
6	I_{Linear}	0.13	0.08			0.73	0.28		
	I_{Spline}	0.19	0.11			0.61	0.32		
	B-Linear	0.08	0.04	1.8	-1	0.87	0.16	2	-1
	Sigmoid	0.07	0.04	16	0.8	0.87	0.16	16	0.8

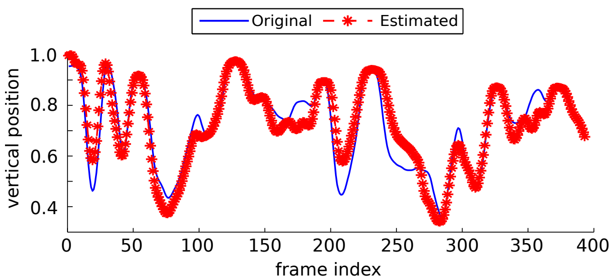

Figure 3: An example plot of the original and estimated vertical trajectories of the tongue tip.

In terms of both E_{RMSE} and E_{CORR} our model shows satisfactory estimation performance, significantly better than the two simple interpolation methods. The sigmoid kernel and the bounded linear kernel with their best hyper-parameters are similar each other; RMSE between the two kernel functions is 0.0948 for the input $[0, 0.01, \cdots, 1]$. Hence, the best kernel function suggests that the relative influence of the contextual constraint is almost linearly decaying roughly until the half time from the crucial point of the articulator to the nearest (preceding or following) crucial time point. After the half, the influence of the physical constraint is dominant. Note that $K_i(t)$ is a function of $\lambda_i(t)$ that is linear defined in $[t_p, t_f]$, where t_p and t_f are the closest crucial time points and $t_p \le t < t_f$ and t_p and t_f are the closest crucial time points, hence the linearity until the half holds.

For more detailed analysis, we computed RMSE and the correlation coefficient for each dimension. 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 uppwer lip and the lower lip are not anatomically constrained, but constrainted 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.

Table 4: *Estimation error for each dimension in terms of RMSE and correlation coefficient (CORR). 'STD' is standard deviation. The result of the best performing sigmoid kernel function is reported. '#Arti' denotes the number of articulators used.*

	#Arti		TT	TB	TD	UL	LL	Jaw	Mean	STD
RMSE	4	X	0.10		0.08		0.09	0.13	0.10	0.02
		Y	0.09		0.10		0.07	0.07	0.08	0.02
	6	X	0.05	0.04	0.05	0.18	0.11	0.07	0.08	0.05
		Y	0.06	0.05	0.04	0.08	0.08	0.04	0.06	0.02
CORR	4	X	0.69		0.76		0.75	0.62	0.71	0.06
		Y	0.82		0.80		0.89	0.88	0.85	0.04
	6	X	0.93	0.94	0.90	0.22	0.65	0.93	0.76	0.29
		Y	0.94	0.95	0.98	0.76	0.83	0.95	0.90	0.09

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.*

#Arti		Neutrality	Anger	Happiness	Sadness	Fear
4	E_{RMSE}	0.09	0.10	0.09	0.08	0.11
	E_{CORR}	0.78	0.76	0.77	0.80	0.69
6	E_{RMSE}	0.07	0.07	0.07	0.07	0.07
	E_{CORR}	0.87	0.84	0.85	0.85	0.85

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} with 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, suggeseting 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 IIS-1116076 and NIH DC007124. Special thanks to Mary Francis for her devotion and help in all SAIL research efforts.

8. References

[1] P. J. Jackson and V. D. Singampalli, “Statistical identification of articulation constraints in the production of speech,” *Speech Communication*, vol. 51, no. 8, pp. 695 – 710, 2009.

[2] G. Papcun, J. Hochberg, T. Thomas, F. Laroche, J. Zacks, and S. Levy, “Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data,” *Journal of the Acoustical Society of America*, vol. 92(2), pp. 688 – 700, 1992.

[3] J. Frankel and S. King, “ASR - Articulatory speech recognition,” in *Proceedings of Eurospeech*, vol. 1, 2001, pp. 599 – 602.

[4] C. A. Fowler and E. Saltzman, “Coordination and coarticulation in speech production,” *Language and Speech*, vol. 36(2,3), pp. 171 – 195, 1993.

[5] C. P. Browman and L. Goldstein, “Articulatory phonology: An overview,” *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[6] ——, “Articulatory gestures as phonological units,” *Haskins Laboratories Status Report on Speech Research*, vol. SR-99/100, pp. 69 – 101, 1989.

[7] G. Ananthakrishnan and O. Engwall, “Important regions in the articulator trajectory,” in *8th International Seminar on Speech Production*, 2008, pp. 305–308.

[8] T. Kato, S. Lee, and S. Narayanan, “An analysis of articulatory-acoustic data based on articulatory strokes,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, April 2009, pp. 4493–4496.

[9] P. Jackson and V. Singampalli, “Statistical identification of critical, dependent and redundant articulators,” *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3321, 2008.

[10] D. Recasens, M. D. Pallars, and J. Fontdevila, “A model of lingual coarticulation based on articulatory constraints,” *Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 544 – 561, 1997.

[11] E. Saltzman and J. Kelso, “Skilled actions: A task dynamic approach,” *psychological Review*, vol. 94, pp. 84 – 106, 1987.

[12] E. Saltzman and K. Munhall, “A dynamical approach to gestural patterning in speech production,” *Ecological Psychology*, vol. 1, pp. 333–382, 1989.