Important Regions in the Articulator Trajectory

G. Ananthakrishnan, Olov Engwall
Centre for Speech technology, KTH (Royal Institute of Technology)
E-mail: agopal@kth.se, engwall@kth.se

Abstract

This paper deals with identifying important regions in the articulatory trajectory based on the physical properties of the trajectory. A method to locate critical time instants as well as the key articulator positions is suggested. Acoustic-to-Articulatory Inversion using linear and non-linear regression is performed using only these critical points. The accuracy of inversion is found to be almost the same as using all the data points.

1 Introduction

Acoustic-to-articulatory inversion has commonly been performed applying an inversion-by-synthesis method, in which an articulatory model is first used to build a code-book containing combinations of articulatory parameter values and acoustic features by synthesizing sounds from the entire articulatory space of the model [1, 10, 3]. A first problem with this approach is that the synthesis step will include articulations that never occur in human speech production. A second, is that the same acoustic features could have been generated by different combinations of articulatory parameters. These problems have been tackled by optimizing the code-book to avoid forbidden articulatory combinations and applying rules to ensure smoothness.

Recently, statistical inversion methods, relying on large databases of simultaneous acoustic and articulatory recordings of continuous speech, gathered with for e.g., Electromagnetic Articulography (EMA) [11, 8, 5] have gained ground. The advantage with the such methods is that only allowed articulatory configurations appear in the database. The articulatory trajectories indicate how the transition between phoneme targets is made. Statistical analysis and machine learning algorithms can then be applied to the databases to learn the acoustic-articulatory relationship.

A problem that is encountered in such methods is that, the acoustic to articulatory mapping is non-unique. However, the combination of the articulatory features corresponding to one acoustic feature is not arbitrary. For each phoneme, the position of some articulators is more critical. The positions of other articulators are not random, but governed by co-articulation between adjacent phonemes. For example, consider the word “cooler” /kuləɹ/. For /k/, the high dorsum position is critical, while the configurations of the lips and tongue tip are non-restricted. The lips hence already start rounding and the tip of the tongue may start rising, in anticipation of /u/ and /l/ respectively. Sometime during the pronunciation of /u/, the lips attain the maximum rounding state and start retracting to reach the spread position of /ə/. The raising of the tongue tip ends sometime in the middle of /l/ and then it starts moving downwards. The tip does not return to the prototypic position for /ə/, because of the effect of /ɹ/. The above example illustrates two aspects. The first is that the position of certain articulators is more critical than others, for each phoneme. The non-critical articulators are in the process of moving towards their next critical position in the utterance. The second is that the timing is important. Automatically identifying when an articulatory feature is critical could therefore improve the estimation and evaluation of the statistical inversion methods.

This paper attempts to identify the important regions of the articulatory trajectory in an utterance, and separate them from those that are just incidental, due to co-articulation with adjacent phonemes. Further, the paper tests acoustic-to-articulatory inversion performed by learning the linear or non-linear transformation using only the important regions of the articulator trajectory and their corresponding acoustic vectors rather than the entire data-set.

2 Obtaining the critical points

There are many ways a critical point in the articulatory trajectory can be defined. There are three pos-

sible parameters that need to be specified for criticality. They are: which articulator, where it is located and when it is located at the said place. The place of articulation which is used to define the various phonemes in IPA can give a useful clue about these critical points. However, they do not specify the timing of the criticality. Secondly, more than one articulator position may be critical in producing a particular acoustic cue. Thirdly, co-articulation in continuous speech has an effect on this criticality, since it can result in a large variability in the place of articulation. It thus becomes necessary to define new measures of criticality.

Jackson and Singampalli suggested a statistical approach to measure the criticality of the articulators [4], in which the Kullback-Leibler distance between the distributions of different articulators was used to classify articulators as critical, dependent or redundant. Recasens [7] used phonetic invariance in the articulatory space to explain critical articulators, while Blandon [2] explained the same phenomenon using articulatory resistance for the phoneme /l/.

The above studies have been able to explain criticality in terms of the position of the articulators, while explaining which articulators are important for the pronunciation of a particular phoneme. The method used in this paper, endeavors to associate criticality to the point in the articulatory trajectory where there is a change in direction of the trajectory or a minimum in velocity. The motivation behind using this simple physical means is the assumption that every articulator is moved in a series of critical positions. Consecutive critical positions may occur several phonemes later. If the articulator is not critical for a particular phoneme, then it is on its way to reach the next critical position. Thus, when there is a drop in velocity or change in angle, then the articulator has reached the critical position and the velocity increases again (probably with a change in angle) to reach the next critical location.

For an utterance with 'T' articulatory samples, the Importance, '$I_a(t)$', for articulator a at time t, is

$$I_a(t) = \frac{\theta_a(t)}{\max\limits_{1\leq i\leq T} \theta_a(i)} - \frac{v_a(t)}{\max\limits_{1\leq i\leq T} v_a(i)} \qquad (1)$$

where $v(t)$ is the velocity and $\theta_a(t)$ is the angle made by the trajectory of articulator a at time t.

A critical point is defined as a local maximum in this importance function, $I_a(t)$. However, since the articulatory trajectories are not completely smooth, minor perturbation may be misrepresented as local maxima and hence critical points. It is therefore necessary to define a window in which only one local maximum must be considered. A variable length window is used in this case. If the trajectory has large changes due to high velocity, then the window is made smaller, and when the velocity is lower, then the window is made larger.

At instant t, the window is calculated based on the parameter μ, which indicates the minimum movement of the articulator that can be called significant. This value must depend on the articulator itself and also on the location that the articulator is in. Their values need to be determined experimentally based on the impact on the acoustics, but this is beyond the scope of this paper. In this formulation, μ is empirically set to be a fraction between 0.05 to 0.1 of the entire range of the articulator in the utterance. If the fraction μ is larger, then fewer points will be selected as critical.

Let $\gamma_a(t)$ be the pair of x and y sampled positions of EMA coil 'a' at time instant 't'. The velocity $v_a(t)$ is calculated between the positions $\gamma_a(t)$ and $\gamma_a(t-1)$ for all time instants 2 to T. The starting time frame of the window at time t is

$$S_a(t) = \arg\min_{1\leq i\leq t}(|\sum_{i\leq j\leq t}(v_a(j)) - \mu_a|) \qquad (2)$$

and the ending time frame is

$$E_a(t) = \arg\min_{t\leq i\leq T}(|\sum_{t\leq j\leq i}(v_a(j)) - \mu_a|) \qquad (3)$$

for $2 \leq t \leq T$. The angle $\theta_a(t)$ is the acute angle made between the line segments $[\gamma_a(S_a(t)), \gamma_a(t)]$ and $[\gamma_a(t), \gamma_a(E_a(t))]$. A point '$t$' with the articulator positions $\gamma_a(t)$ on the trajectory of an articulator a is a critical point if

$$I_a(t) > I_a([S_a(t), t-1 \cup t+1, E_a(t)]) \qquad (4)$$

i.e., if the importance value at time t is higher than anywhere in the window. Thus we have a sliding window and the point with the highest importance in that window is the critical point.

3 Acoustic-articulatory data

Experiments on critical point identification and articulatory inversion were carried out using the MOCHA-TIMIT database [9], with one female speaker (fsew0) uttering 460 British English sentences. The positions of 7 EMA coils on the tongue tip (TT), tongue blade (TB), tongue dorsum (TD), velum (V), upper lip (UL), lower lip (LL) and lower jaw (LJ) along the mid-sagittal plane were recorded. The original sampling frequency of the recordings were 500 Hz. They were low-pass filtered and downsampled to 125 Hz.

Figure 1 shows a typical trajectory of the tongue

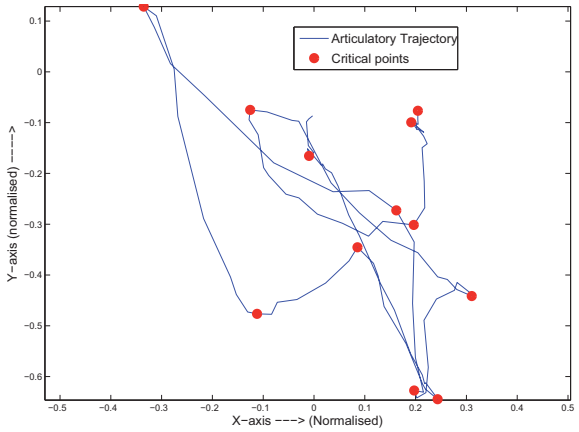

Figure 1: *The trajectory of the tongue tip during the utterance of the sentence 'He will allow a rare lie'.*

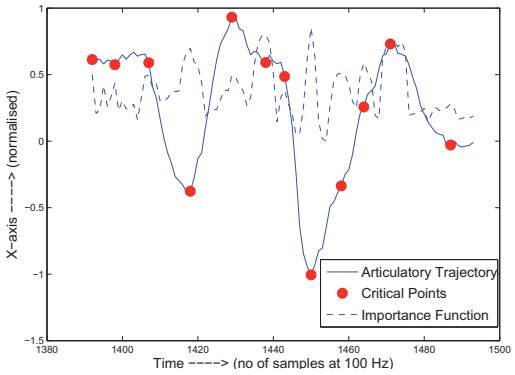

Figure 2: *The X-axis trajectory of the tongue tip along with the importance value '$I_{TT}(t)$' for the sentence 'He will allow a rare lie'.*

tip for the utterance of an English sentence. We can see that the important points usually correspond to the extremities of the trajectories. Figure 2 shows the projection of the trajectory of the tongue tip on the X-axis as a function of time along with the Importance function.

4 Experiments and Discussion

Since the formulation of the importance function is based on change of velocity and angle, the importance of the positions of the tongue is higher for the critical points in stop consonants and fricatives than in vowels. This is in accordance with the knowledge that the impact of the critical articulator on stop consonants is more pronounced.

Similarly, the variance of the location of the critical point also varies to a higher degree in vowels than in consonants. The variance for the critical articulator (related to the place of articulation) at the critical points, is often less than the variance of the entire trajectory during the pronunciation of the phoneme. On the other hand, the variance of the critical points for the non-critical articulators is almost as high as

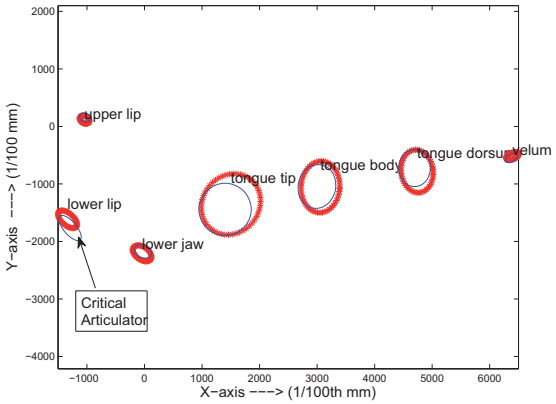

Figure 3: *The standard deviation of all the points for phoneme /p/ is the thin blue ellipse and the standard deviation for the critical points alone is the red thick ellipse. We can see that the mean is centered at the extreme end of the data points for the lips.*

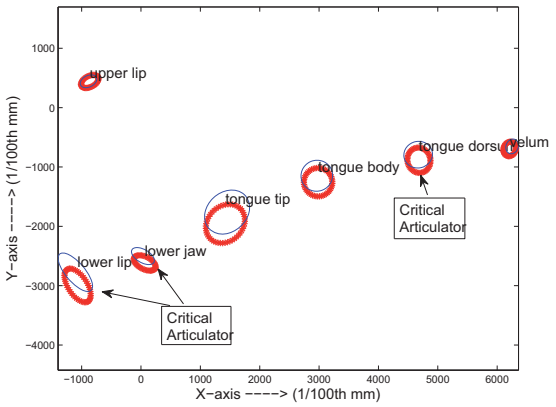

Figure 4: *The standard deviation of all the points for phoneme /a/ is the thin blue ellipse and the thick red ellipse for the critical points alone. We can see that the mean is centered at the extreme end of the data points for the jaws and lower lip.*

or sometimes higher than the variance for the entire trajectory. This aspect is illustrated in Fig. 3 and 4.

For the inversion experiments, the database was split into 5 equal parts and 4 of them were used for training and the remaining was used for testing. The test and training data-sets were rotated using the jackknife principle. The acoustic features were 16 Mel Frequency Cepstral Coefficients (MFCC) extracted from 30 ms length windows shifted with the same frame rate as the articulatory data (125 Hz or 8 ms shift).

The linear regression was used as described by Yehia *et al.* [11]. Non-linear regression was performed using the MATLAB artificial neural networks toolkit (ANN) [6]. Regression for each articulator was trained using, at first, the entire training data (i.e., 80 % of the corpus) and then only the articulatory and acoustic features of the frames correspond-

Table 1: *Table showing the results of linear and non-linear regression using Artificial Neural Networks (ANN) on the entire training data-set as against selecting only the critical points from the training data to perform the training.*

Method	mean % of samples for training	μ	mean RMS Error	Correlation Coeff.
Linear Regression (All data)	80%	-	2.28 mm	0.437
Linear Regression (Critical points)	7.40 %	0.05	2.3 mm	0.432
Linear regression (Critical points)	5.27 %	0.1	2.29 mm	0.434
ANN Regression (All data)	80%	-	2.1 mm	0.544
ANN Regression (Critical points)	7.40%	0.05	2.12 mm	0.543
ANN Regression (Critical points)	5.27 %	0.1	2.13 mm	0.544

ing to the critical points of the articulator among this 80%. The input to the regression was the 16 MFCCs, and the output consisted of the x- and y-coordinates. A hidden layer consisting of 16 neurons was used for the ANN, the same number as the input features.

The testing was done on the entire testing data-set, regardless of whether the samples were critical or not. The results presented in Table 1 are the mean results across the different cross validations. We can see that the linear and non-linear regression using only the critical points perform almost as well as the regression using all the points in the data-set. The reason for this can be attributed to the fact that the non-critical points can be interpolated from the critical points by the linear or non-linear regression.

5 Conclusions and Future Work

A method to find the important regions in the articulator trajectory has been proposed. The method relies on the physical parameters of the trajectory and in that way, is different from most other methods proposed, while being quite simple and intuitive. The method, not only finds the critical articulators and their positions, but also the time instant that the trajectory reaches the critical point. Since the use of critical points leads to a reduction of the training set of up to 65 %, the method can speed up the training stage of the inversion substatially, while maintaining almost the same estimation performance as when using the entire data-set.

These critical points could be used to find a suitable method for evaluation of the inversion process. The inversion method should be evaluated based on whether these critical points are detected with sufficient accuracy and without a time lag. It could also be used to suggest a data driven co-articulation model based on physical constraints of the articulator. It would be based on applying acceleration and jerk constraints on the trajectory between two critical points for the trajectory. Finally, the method suggested could also be generalized for obtaining visual parameters for audio-visual data processing.

6 Acknowledgements

The authors acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Sixth Framework Programme for Research of the European Commission, under FET-Open contract no. 021324".

References

[1] S. Atal, J. Chang, J. Mathews, and W. Tukey. Inversion of articulatory-to-acoustic information in the vocal tract by a computer-sorting technique. *J Acoust Soc Am*, 63:1535–1555, 1978.

[2] R. Bladon and A. Al-Bamerni. Coarticulation Resistance in English/1/. *Journal of Phonetics*, 4(2):137–150, 1976.

[3] J. Dang and K. Honda. Estimation of vocal tract shapes from sound with a physiological articulatory model. *Journal of Phonetics*, 30:511–532, 2002.

[4] P. Jackson and V. Singampalli. Statistical identification of critical, dependent and redundant articulators. *J Acoust Soc Am*, 123(5):3321, 2008.

[5] A. Katsamanis, G. Papandreou, and P. Maragos. Audiovisual-to-articulatory speech inversion using active appearance models for the face and hidden markov models for the dynamics. In *Proc. ICASSP*, 2008.

[6] M. Nørgaard. *Neural Networks for Modelling and Control of Dynamic Systems: A Practitioner's Handbook*. Springer, 2000.

[7] D. Recasens, D. Pallarés, and J. Fontdevilla. A model of lingual coarticulation based on articulatory constraints. *J Acoust Soc Am*, 102:544–561, 1997.

[8] K. Richmond. A trajectory mixture density network for the acoustic-articulatory inversion mapping. In *Proc. Interspeech*, pages 2465–2468, 2006.

[9] A. Wrench. The MOCHA-TIMIT articulatory database. *Queen Margaret University College, Tech. Rep*, 1999.

[10] Q. Xue, Y. Hu, and P. Milenkovic. Analyses of the hidden units of the multi-layer perceptron and its application in acoustic-to-articulatory mapping. In *Proc. ICASSP*, pages 869–872, 1990.

[11] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1-2):23–43, 1998.