Jerk Minimization for Acoustic-To-Articulatory Inversion

Avni Rajpal and Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT),
Gandhinagar-382007, India.
E-mail:{avni_rajpal,hemant_patil}@daiict.ac.in

Abstract

The effortless speech production in humans requires coordinated movements of the articulators such as lips, tongue, jaw, velum, etc. Therefore, measured trajectories obtained are smooth and slowly-varying. However, the trajectories estimated from acoustic-to-articulatory inversion (AAI) are found to be *jagged*. Thus, energy minimization is used as smoothness constraint for improving performance of the AAI. Besides energy minimization, jerk (i.e., rate of change of acceleration) is known for quantification of smoothness in case of human motor movements. Human motors are organized to achieve intended goal with smoothest possible movements, under the constraint of minimum accelerative transients. In this paper, we propose jerk minimization as an alternative smoothness criterion for frame-based acoustic-to-articulatory inversion. The resultant trajectories obtained are smooth in the sense that for articulator-specific window size, they will have minimum jerk. The results using this criterion were found to be comparable with inversion schemes based on existing energy minimization criteria for achieving smoothness.

Index Terms: Acoustic-to-Articulatory Inversion (AAI), Jerk Minimization, Electromagnetic Articulography (EMA).

1. Introduction

Articulatory parameters have complementary information as opposed to that provided by the acoustic features [1]. However, recording of articulatory movements is more difficult than the speech signal. Thus, attempts have been made to estimate articulator trajectories from the features obtained from the acoustic signal. This is called as acoustic-to-articulatory inversion (AAI). The nature of mapping between acoustic and articulatory parameters was found to be nonlinear and non-unique which makes AAI an ill-posed problem [2], [3]. Solution to the inversion problem is significant both for practical applications (such as speech synthesis, speech recognition, language acquisition, speaker verification, etc. [4], [5]) as well as for theoretical analysis (e.g., production-perception link [6]). Humans coordinate movements of their articulators in order to produce intelligible speech signal. This results in smooth and slowly-varying trajectories, which is observed in measured articulatory data. It is known that, the mapping between acoustic and articulatory features is non-unique, i.e., different articulator configurations can correspond to similar acoustic features. Moreover, the estimated trajectories are *jagged* [7], [8]. Thus, constraints have to be applied such that the estimated solution does not deviate much from the exact or true solution, i.e., the ground truth. This reduces non-uniqueness in mapping function and smoothed estimated trajectories are obtained. [9].

Most commonly, energy minimization and dynamic constraints are used as smoothness criteria for inversion. The energy minimization constraint aims at minimizing the energy of estimated trajectories in high frequency regions. As a result, the trajectories are smooth and lowpass in nature. The study presented in [10], [11] apply energy minimization constraint by using lowpass filter as a post-processing step. Other studies such as generalized smoothness criterion (GSC) [9] and sparse smoothing [8] propose objective function that combines articulator-specific energy minimization constraint with the mapping function. The dynamic constraints, on the other hand, are based on the information related to the continuity in the trajectory. This information is provided by first and second order derivatives that act as dynamic features. The study reported in [10] proposed an objective function that combines Gaussian mixture model (GMM)-based mapping with dynamic features. In [12], information from dynamic features has been utilized, while estimating trajectories using mixture density networks. All the above approaches showed, improved inversion performance on applying smoothness constraints.

In this paper, jerk minimization a physiological-based smoothness constraint is proposed for frame-based inversion. This is motivated from the Hogan's work who proposed minimum jerk trajectory model for human motor movements [13]. According to this model, human motor movements are organized to perform intended task such that they take smoothest path under the constraint that jerk is minimum [13]. Hence, this fact about human motor movements is used as a constraint to obtain smooth articulator trajectories after inversion. In [14], jerk minimization is exploited for smoothing trajectories obtained from gesture-based inversion. However, in our work, we aim to understand the usefulness of such physiological constraint for frame-based inversion task.

The rest of the paper is organized as follows. In Section 2, non-parametric regression is described. Section 3 discusses details of jerk minimization. Experimental setup is described in Section 4. Experimental results are presented in Section 5which is followed by summary and conclusions in Section 6.

2. Non-Parametric Regression

There are number of approaches available in literature to obtain inverse mapping function between acoustic and articulatory vectors [4], [11], [14], [15], [16], [17]. In this paper, non-parametric regression is used to estimate the articulator values from the acoustic feature vector. The acoustic and articulatory vectors are obtained from MOCHA database [18], the details of which are given in Section 4. The training set consists of M acoustic-articulatory vector pairs $\{(\lambda^i, \beta^i); 1 \leq i \leq M\}$, where λ^i is *14-D* MFCC vector and β^i is *14-D* articulatory vector for i^{th} frame. For a given test utterance with N frames, aim is to estimate *14-D* articulatory vector for each frame. The steps involved in regression are described in Table 1 [9], [19].

http://dx.doi.org/10.21437/SSW.2016-14

Table 1. *Steps to estimate articulatory trajectories*

1. **Input:** Acoustic-articulatory vector pairs $\{(\lambda^i, \beta^i); 1 \le i \le$ *M*, test acoustic feature vector *un;1≤n≤N.*
2. **Output:** *14-D* estimated articulatory vector $\{\hat{z}^n; 1 \le n \le$ *N.*
3. For n^{th} test frame u^n, find Euclidean distance with each acoustic vector in the training set
$$\gamma_{n,i} = \left\| u^n - \lambda^i \right\|, 1 \le i \le M, \qquad (1)$$
4. $\gamma_{n,i}$ is sorted to obtain top *K*, λ^i that have minimum distance with u^n. (In this paper, *K=200* is used).
5. For each of these *K* acoustic vectors, corresponding β^i are picked from the training set.
6. Take weighted mean of β^i, to obtain $\hat{z}^n$, which is given by
$$\hat{z}^n = \sum_{m=1}^{K} \beta_m^i w, \qquad (2)$$

where $w = \frac{\gamma_{n,i}^{-1}}{\sum_i \gamma_{n,i}^{-1}}$ is the weighting function which is *inversely* proportional to distance $\gamma_{n,i}$.

The articulatory trajectory estimated using non-parametric regression for each articulator is then smoothened using jerk minimization. The details of which are discussed in Section 3.

3. Jerk Minimization

Hogan proposed that human motor movements tend to perform a task such that the smoothness of motion is maximized [13]. In this model, jerk is used as a measure of smoothness, which is defined as rate of change of acceleration or third time derivative of position *z(t)* which is given by (3). Thus, smoothest path between initial and final target position is the one which gives minimum jerk cost. The jerk cost is defined as the mean square jerk along the trajectory and is given by (4):

$$\dddot{z}(t) = \frac{d^3 z(t)}{dt^3}, \qquad (3)$$

$$f(z(t)) = \frac{1}{2} \int_{t=t_i}^{t=t_f} \left(\dddot{z}(t)\right)^2 dt , \qquad (4)$$

where t_i and t_f are the starting and ending time of the trajectory. The trajectory which minimizes (4) will be the desired smoothest path and is called *minimum jerk trajectory*. The ½ factor in (4) is to make calculation easier; otherwise, it has no significance. The calculus of variation is used as optimization technique to find the minimum of (4), the details of which are discussed in Appendix A. The solution obtained is 5^{th} order polynomial which is given by [13]:

$$z(t) = a_o + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + a_5 t^5 \ . \qquad (5)$$

Now, if $\hat{z}_k(t)$ is the estimated trajectory of the k^{th} articulator, then the smooth version of the trajectory $\hat{z}_{sk}(t)$ is obtained by fitting minimum jerk trajectory which is given by [14]:

$$\hat{z}_{sk}(t) = \begin{bmatrix} 1 & t & t^2 & t^3 & t^4 & t^5 \end{bmatrix} \times \begin{bmatrix} \bar{\mathbf{1}} & \bar{\mathbf{t}} & \bar{\mathbf{t}}^2 & \bar{\mathbf{t}}^3 & \bar{\mathbf{t}}^4 & \bar{\mathbf{t}}^5 \\ \bar{\mathbf{0}} & \bar{\mathbf{1}} & 2\bar{\mathbf{t}} & 3\bar{\mathbf{t}}^2 & 4\bar{\mathbf{t}}^3 & 5\bar{\mathbf{t}}^4 \\ \bar{\mathbf{0}} & \bar{\mathbf{0}} & \bar{\mathbf{2}} & 6\bar{\mathbf{t}} & 12\bar{\mathbf{t}}^2 & 20\bar{\mathbf{t}}^3 \end{bmatrix}^{-1} \times \begin{bmatrix} \hat{z}_k(\bar{\mathbf{t}}) \\ d\hat{z}_k(\bar{\mathbf{t}}) \\ d^2\hat{z}_k(\bar{\mathbf{t}}) \end{bmatrix}, \qquad (6)$$

where $\bar{\mathbf{t}}$ is a column vector of time instances in the interval $[t - w_s,\ t + w_s]^T$ and $\bar{\mathbf{0}}$, $\bar{\mathbf{1}}$ and $\bar{\mathbf{2}}$ are the column vectors of length $2w_s+1$. Moreover, inverse taken in (6) is pseudo-inverse, and $\hat{z}_k(t)$ is estimated using non-parametric regression.

4. Experimental Setup

This Section briefly discusses about basic framework and features used for experiments presented in this paper.

4.1. MOCHA Database

The Multichannel Articulatory (MOCHA) database [18] consists of simultaneously recorded acoustic and articulatory data obtained from two speakers, i.e., male and female speaker. The corpus consists of *460* phonetically diverse British English TIMIT sentences, audio signal sampled at *16* kHz, the laryngographic signal sampled at *16* kHz, Electro Magnetic Articulography (EMA) data sampled at *500* Hz and EPG data sampled at *200* Hz. However, only EMA data is used as articulatory data for inversion in this work.

EMA data consists of X and Y coordinates of *9* receiver sensor coils attached to *9* points along the midsaggital plane, namely, the lower incisor or the jaw (*li_x, li_y*), upper lip (*ul_x, ul_y*), lower lip (*ll_x, ll_y*), tongue tip (*tt_x, tt_y*), tongue body (*tb_x, tb_y*), tongue dorsum (*td_x, td_y*), velum (*v_x, v_y*), upper incisor (*ui_x, ui_y*) and bridge of the nose (*bn_x, bn_y*). The upper incisor and bridge of the nose are used as reference coils. For our experiments, out of *460* utterances available in the database, *368* utterances (*80 %*) were used for training, *37* utterances (*8 %*) as the development (dev) set and the remaining *55* utterances (*12 %*) as the test set. The numbers of frames of the articulatory data used for experiments are shown in Table 2.

Table 2. *Count of articulatory data frames*

Speaker	Number of Articulatory Frames		
	Training Set	Dev Set	Test Set
Male	82591	8550	13920
Female	97939	10061	16254

4.2. Articulatory Features

Articulatory data obtained from *14* channels corresponding to first seven coils except the reference coils are used as articulatory features in our experiments. The following preprocessing steps similar to [9], [11] are used:

- First, EMA data from each channel is lowpass filtered with cutoff frequency of *35* Hz. The filtering process is "zero-phase" filtering to alleviate phase distortion.
- Filtered data is downsampled by a factor of *5* in order to match the frame rate of the acoustic features.
- In order to avoid articulators from taking any position in silence regions, the starting and ending portion of the silence from each utterance is removed manually.
- Next, the slowly-varying trends in EMA data are removed [11]. For this, first file-by-file mean for each articulator is calculated. Then this sequence of raw means is lowpass filtered. The file-specific mean returned after lowpass filtering is then subtracted from articulatory trajectories.
- Finally, global mean of each articulator is added.

4.3. Acoustic Features

Mel frequency cepstral coefficients (MFCC) were calculated for the speech data obtained from the MOCHA database. For inversion,*14-D* MFCC are used as acoustic features which are calculated using *20* ms Hamming window with shift of *10* ms duration.

4.4. Performance Measures

The performance of jerk-minimization based AAI was evaluated using two measures, namely, Root Mean Square Error (RMSE (ε)) and Pearson correlation (ρ). The RMSE is the measure which determines pointwise closeness of the estimated trajectory with the measured trajectory in terms of the distance. Pearson correlation (ρ), on the other hand indicates, how closely the shape of the estimated trajectory matches with the original trajectory. The RMSE (ε) (*mm*) and Pearson correlation (ρ) for k^{th} articulator is given by [9]:

$$\varepsilon = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\hat{z}_k^n - \hat{z}_{sk}^n\right)^2}\,, \qquad (7)$$

$$\rho = \frac{N\sum_n \hat{z}_k^n \hat{z}_{sk}^n - \sum_n \hat{z}_k^n \sum_n \hat{z}_{sk}^n}{\sqrt{N\sum_n\left(\hat{z}_k^n\right)^2 - \sum_n\left(\hat{z}_k^n\right)^2}\sqrt{N\sum_n\left(\hat{z}_{sk}^n\right)^2 - \left(\sum_n \hat{z}_{sk}^n\right)^2}}, (8)$$

where $\hat{z}_k^n$ is the measured trajectory and $\hat{z}_{sk}^n$ is the estimated trajectory of length *N* obtained after smoothing.

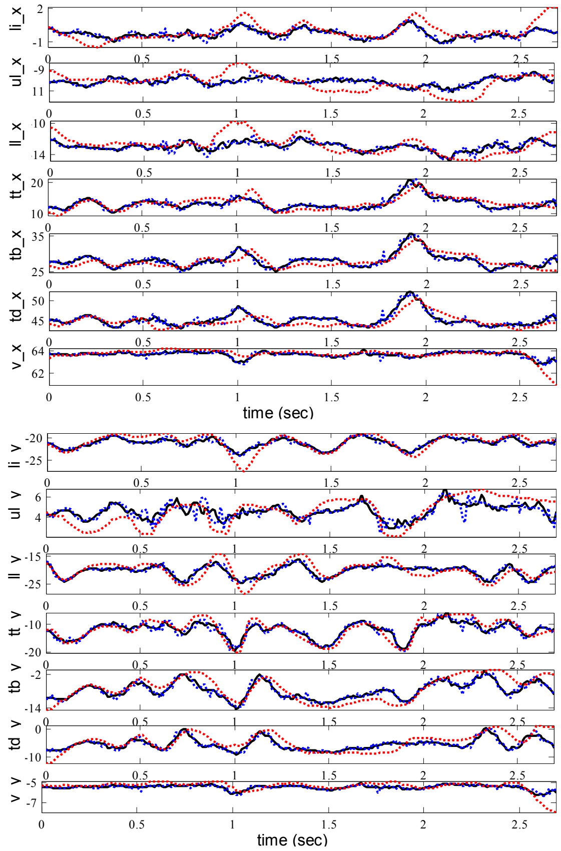

Figure1. *Measured trajectories (red --), estimated trajectories without smoothing (blue --)and estimated trajectory (black solid line) of female speaker of the utterance "Those who teach values first abolish cheating" with respect to time in seconds.*

5. Experimental Results

In our experiments, smooth trajectories are obtained using jerk minimization, for which tuning of w_s parameter for specific articulator, is done using dev set of MOCHA database. First, the trajectory for each articulator is estimated for the dev set, then it is smoothed for different set of values of w_s using (6). w_s value that gave minimum average ε was chosen. The similar experiment was repeated for all the articulators. The set of values of w_s taken for our experiments are {*40, 60, 80, 100, 120, 200, 300*}. Figure1 shows the estimated trajectories of all *14* articulators before and after jerk minimization, for female speaker of the randomly chosen utterance, namely, *"Those who teach values first abolish cheating"*, over laid on the measured trajectories of same utterance available from the database. The average ε and average ρ along with their standard deviation (SD) and w_s values for female speaker are shown in Table 3.

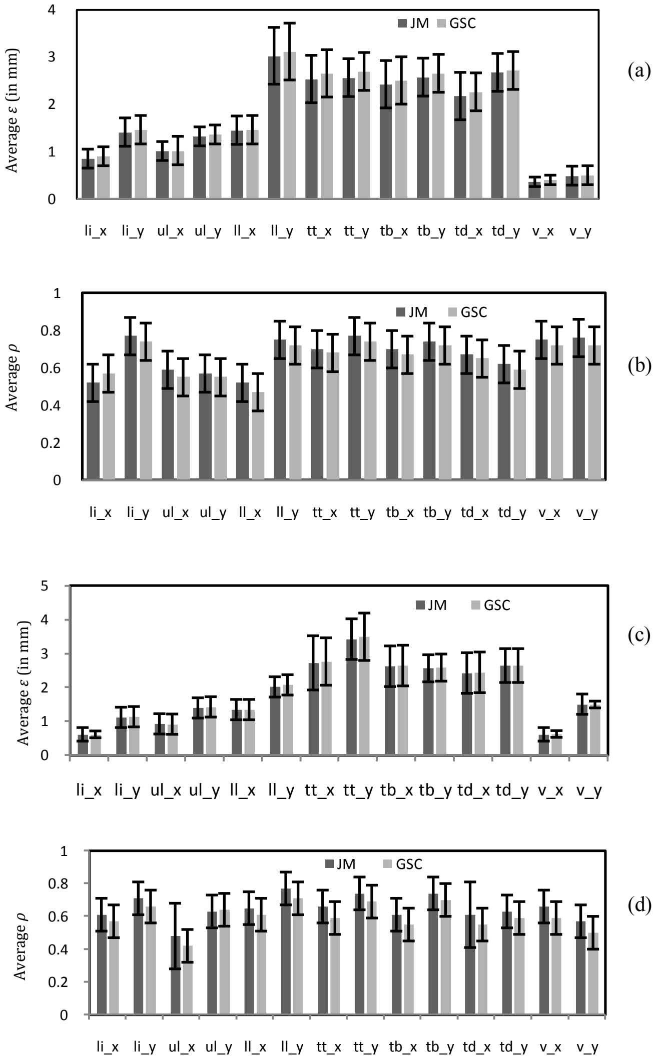

Figure 2. *Bar diagram of average ε and average ρ for jerk minimization (a)-(b) MOCHA female, (c)-(d) MOCHA male. Error bars indicate SD.*

.

Table 3. *Comparison of the performance of jerk minimization (JM) with GSC for female speaker, in term of average ε, average ρ (along with SD of ε and ρ is shown in bracket) and 95 % confidence interval. Table also shows, the parameters used for performing JM and GSC*

EMA Channel	Parameters			Average ε (SD) (mm)		Average ρ (SD)		*95%* Confidence Interval	
	JM	GSC		JM	GSC	JM	GSC	JM	GSC
	w_s (*ms*)	*C*	Cut-off frequency (γ_c) (in *Hz*)						
li_x	100	0.1	1.89	0.85 (0.2)	0.9 (0.2)	0.52 (0.1)	0.57 (0.1)	0.87±0.14	0.88±0.14
li_y	100	50	3.08	1.41 (0.3)	1.46 (0.3)	0.77 (0.1)	0.74 (0.1)	1.46±0.18	1.64±0.2
ul_x	100	10	2.29	1.01 (0.2)	1.02 (0.3)	0.59 (0.1)	0.55 0.1)	1.04±0.16	1.06±0.16
ul_y	100	5	2.68	1.32(0.2)	1.36 (0.2)	0.57 (0.1)	0.55 (0.1)	1.33±0.18	1.39±0.18
ll_x	80	10	2.68	1.45 (0.3)	1.46 (0.3)	0.52 (0.1)	0.47 (0.1)	1.48±0.19	1.51±0.19
ll_y	80	5	2.29	3.02 (0.6)	3.11 (0.6)	0.75 (0.1)	0.72 (0.1)	3.1±0.27	3.28±0.27
tt_x	80	5	2.29	2.53 (0.5)	2.65 (0.5)	0.7 (0.1)	0.68 (0.1)	2.58±0.24	2.73±0.25
tt_y	80	10	2.68	2.56 (0.4)	2.69 (0.4)	0.77 (0.1)	0.74 (0.1)	2.59±0.24	2.76±0.25
tb_x	80	50	2.29	2.42 (0.5)	2.5 (0.5)	0.7 (0.1)	0.67 (0.1)	2.48±0.24	2.64±0.25
tb_y	80	5	2.29	2.57 (0.4)	2.65 (0.4)	0.74 (0.1)	0.72 (0.1)	2.61±0.24	2.74±0.25
td_x	80	50	2.68	2.17 (0.5)	2.26(0.4)	0.67 (0.1)	0.65(0.1)	2.23±0.23	2.55±0.24
td_y	80	5	2.29	2.67 (0.4)	2.71(0.4)	0.62 (0.1)	0.59(0.1)	2.72±0.25	2.78±0.25
v_x	100	100	2.29	0.36 (0.1)	0.4 (0.1)	0.75 (0.1)	0.72(0.1)	0.37±0.09	0.48±0.11
v_y	100	50	3.87	0.49 (0.2)	0.5 (0.2)	0.76 (0.1)	0.72 (0.1)	0.52±0.11	0.53±0.11

Table 4. *Comparison of GSC and JM at phoneme-level, in terms average ε, average ρ (along with SD of ε and ρ is shown in bracket)*

Phoneme Class	MOCHA Female Database				MOCHA Male Database			
	Average ε (SD) (mm)		Average ρ (SD)		Average ε (SD) (mm)		Average ρ (SD	
	GSC	JM	GSC	JM	GSC	JM	GSC	JM
Vowels and Diphthongs	1.59 (0.19)	**1.47 (0.16)**	0.36 (0.14)	**0.42 (0.13)**	1.61(0.22)	**1.54 (0.21)**	0.36 (0.14)	**0.43 (0.13)**
Stop consonants	1.87 (0.28)	**1.73 (0.28)**	0.27 (0.05)	**0.37 (0.07)**	1.88 (0.39)	**1.79 (0.39)**	0.25 (0.07)	**0.37(0.05)**
Nasals and liquids	1.69 (0.23)	**1.59 (0.21)**	0.27 (0.05)	**0.37 (0.02)**	1.8 (0.32)	**1.73 (0.31)**	0.26 (0.08)	**0.36 (0.06)**
Fricatives	1.66 (0.23)	**1.53 (0.22)**	0.29 (0.12)	**0.35 (0.11)**	1.61(0.31)	**1.53 (0.29)**	0.31 (0.09)	**0.39 (0.08)**

5.1. Comparison with GSC

The results obtained using jerk minimization was found in similar range with the results reported in the literature using different smoothness constraints [8], [9], [10], [11]. In particular, in this paper GSC is chosen for comparison with our proposed JM because GSC uses non-parametric regression for finding mapping which is similar to what is being used in our proposed scheme [9]. Therefore, direct comparison of results can be made. However, difference between our approach and GSC lies in the smoothness constraint. In GSC, energy minimization constraint is used (i.e., articulator-specific highpass filter is chosen, that minimizes energy in the high frequency regions of the output) whereas in our proposed approach jerk is being minimized.

Since, the training and test samples of GSC were not known; it was implemented using same experimental setup, as it is used in this paper. The performance of the inversion using GSC for female speaker along with the parameters used to obtain these results is shown in Table 3. For comparison of the performance of two different smoothing constraints, we calculated *95 %,* confidence interval of framewise error for the test set. Table 3 indicates the reduction of average ε and increase in average ρ for jerk minimization as compared to GSC. However, in terms of confidence interval, the performance differences of both constraints are not statistically significant. The bar plots for average ε and average ρ along with error bars for both the speakers shown in Figure 2 also indicates the comparable performances of both constraints. The similar results are obtained for the male speaker. We also compared the performance of the constraints at phoneme-level. For this, utterances in the test set were divided based on the phonemes they belong to. Then, for each phoneme, the estimated articulatory trajectory obtained both from GSC and our proposed method was compared with the measured trajectory. The average RMSE and average correlation for different phoneme classes along with their standard deviation (SD) is shown in Table 4. The results clearly indicate that jerk minimization gives better estimates for individual phoneme trajectories as compared to GSC. Thus, experiments indicate that jerk minimization though did not give significant advantage over other smoothness constraints for frame-based inversion, yet show improved performance at phoneme-level. This motivates us to further investigate its implications in speech modification applications in future.

In order to modify speech signal in an understandable way, articulatory parameters should be manipulated rather than acoustic parameters. For this reason, researchers have incorporated articulatory parameters in a variety of speech modification problems. However, majority of the researchers use directly measured articulatory parameters rather than inverted parameters in applications such as voice transformation [20], foreign accent conversion [21], and flexible text-to-speech (TTS) synthesis [22]. Inversion problem is too complex to be solved without approximation which result in degraded synthetic speech quality. Hence, even though inverted articulatory features have been found useful for speech recognition [1], [23], however, very few work indicates the effectiveness of inverted articulatory features in speech modification [24], [25]. In our work, we intend to improve the existing inversion techniques by proposing physiologically motivated smoothness constraint. In future, our experiments will be directed to study the impact on synthesis quality when articulators predicted using jerk minimization are used in articulatory synthesis.

6. Summary and Conclusions

The literature related to AAI shows improvement in inversion performance when smoothness constraint for articulator trajectories is imposed. In this paper, articulatory trajectories were estimated using frame-based inversion, such that the output trajectories obtained are smooth under the constraint that jerk is minimum. This constraint is motivated from the minimum jerk trajectory model which is found to work well in human motor movements. The results obtained using minimum jerk criterion shows improvement in estimation accuracy over GSC. Moreover, average RMSE and average correlation of individual articulators are in range as compared to the state-of-the-art results reported in the literature. The effect of jerk minimization was also checked at phoneme- level. The results show that our proposed method gives better estimation of articulatory trajectories at phoneme-level. Thus, study in our paper indicates that jerk minimization can be used as an alternative smoothness constraint for frame-based AAI besides dynamic constraints and energy minimization constraint. In future, the effectiveness of different mapping techniques under jerk minimization constraint will be explored.

7. Acknowledgements

The authors would like to thank Dept. of Electronics and Information Technology (DeitY), Govt. of India, for sponsored project, Development of Text-to-Speech (TTS) System in Indian Languages (Phase-II) and the authorities of DA-IICT, Gandhinagar.

Appendix A

In order to find the functional *z*(*t*) that has the smoothest path, the functional that has minimum jerk cost is chosen. For this minimum of the function given by (A.1) is solved using calculus of variation.

$$f(z(t)) = \frac{1}{2} \int_{t=t_i}^{t=t_f} \left(\dddot{z}(t)\right)^2 dt \qquad \text{(A.1)}$$

Let $\beta(t)$be the variation, with following properties:

$$\begin{cases} \beta(t_i)=0 & \beta(t_f)=0 \\ \dot{\beta}(t_i)=0 & \dot{\beta}(t_f)=0 \\ \ddot{\beta}(t_i)=0 & \ddot{\beta}(t_f)=0 \end{cases}. \qquad \text{(A.2)}$$

Replace *z*(*t*) with $z(t) + \mathrm{e}\beta(t)$ in equation (A.1), and solve it by taking derivative with respect to the variation.

$$f(z+e\beta) = \frac{1}{2}\int_{t_i}^{t_f} \left(\dddot{z}+e\dddot{\beta}\right)^2 dt,$$

$$\frac{df(z+e\beta)}{de} = \int_{t_i}^{t_f} \left(\dddot{z}+e\dddot{\beta}\right)\dddot{\beta}\, dt,$$

$$\left.\frac{df(z+e\beta)}{de}\right|_{e=0} = \int_{t_i}^{t_f} \dddot{z}\dddot{\beta}\, dt\,.$$

Using integration by parts,

$$\int_{t_i}^{t_f} \dddot{z}\dddot{\beta} dt = \dddot{z}\dddot{\beta}\Big|_{t_i}^{t_f} - \int_{t_i}^{t_f} \dddot{\beta} z^{(4)} dt,$$

$$= -\int_{t_i}^{t_f} \ddot{\beta} z^{(4)} dt,$$

where $z^{(4)}$ is the *4*^*th*^ derivative of *z*(*t*). Continuing integration by parts,

$$-\int_{t_i}^{t_f} \ddot{\beta} z^{(4)} dt = -z^{(4)}\dot{\beta}\Big|_{t_i}^{t_f} + \int_{t_i}^{t_f} \dot{\beta} z^{(5)} dt = \int_{t_i}^{t_f} \dot{\beta} z^{(5)} dt,$$

$$\int_{t_i}^{t_f} \ddot{\beta} z^{(5)} dt = -z^{(5)}\dot{\beta}\Big|_{t_i}^{t_f} + \int_{t_i}^{t_f} \dot{\beta} z^{(6)} dt = \int_{t_i}^{t_f} \dot{\beta} z^{(6)} dt. \qquad \text{(A.3)}$$

Integral in equation (A.3) is the derivative of the functional

$$\left.\frac{df(z+e\beta)}{de}\right|_{e=0} = \int_{t_i}^{t_f} \beta z^{(6)} dt = 0, \qquad \text{(A.4)}$$

The equation (A.4) is valid for all β, and therefore, the functional that satisfies

$$z^{(6)} = 0\,, \qquad \text{(A.5)}$$

i.e., *6*^*th*^ derivative equal to zero will minimize jerk function. The general solution of (A.5) is *5*^*th*^ degree polynomial which is

$$z(t) = a_o + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + a_5 t^5. \qquad \text{(A.6)}$$

Thus, among all the functions, *5*^*th*^ polynomial has the minimum jerk cost.

8. References

[1] I. Zlokarnik, "Adding articulatory features to acoustic features for automatic speech recognition", *J. Acoust. Soc. Am.* 97, no.5, pp. 3246-3246, 1995.

[2] C. Qin and M. A. Carreira-Perpiñán, "An empirical investigation of the non-uniqueness in the acoustic-to-articulatory mapping," in Proc. INTERSPEECH, Antwerp, Belgium, 2007, pp. 74–77.

[3] D. Neiberg, G. Ananthakrishnan, and O. Engwall, "The acoustic-to- articulation mapping: non-linear or non-unique?," in Proc. INTERSPEECH, Brisbane, Australia, 2008, pp. 1485-1488.

[4] J. Schroeter and M. Sondhi, "Techniques for estimating vocal tract shapes from the speech signal," IEEE Trans. Speech Audio Processing, vol. 2,no. 1, pp. 133–150, 1994.

[5] M. Li, J. Kim, P. Ghosh, V. Ramanarayanan, and S. Narayanan, "Speaker verification based on fusion of acoustic and articulatory information," in Proc. INTERSPEECH, Lyon, France, 2013, pp. 1614– 1618.

[6] A. Liberman, F. Cooper, D. Shankerweiler and M. Studdert-Kennedy, "Perception of the speech code," Psychol. Rev., vol. 74, no. 6, pp. 431- 461, 1967.

[7] C. Qin and M. A. Carreira-Perpiñán, "A comparison of acoustic features for articulatory inversion," in Proc. INTERSPEECH, Antwerp, Belgium, 2007, pp. 2469–2472.

[8] P. Sudhakar, L. Jacques, and P. K. Ghosh, "A sparse smoothing approach for Gaussian mixture model based acoustic-to-articulatory inversion," in Proc. Int'l Conference on Acoustics, Speech, and Signal Processing (ICASSP), Florence, Italy, 2014, pp. 3032-3036.

[9] P. K. Ghosh and S. S. Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," *J. Acoust. Soc. Am.*, vol. 128, no. 4, pp. 2162–2172, 2010.

[10] T. Toda, A. W. Black, and K. Tokuda, "Acoustic-to-articulatory inversion mapping with Gaussian mixture model," in Proc. INTERSPEECH, Jeju, Korea, 2004, pp. 1129–1132.

[11] K. Richmond, "Estimating articulatory parameters from the acoustic speech signal," Ph.D. Dissertation, Edinburgh Univ., Centre Speech Technol. Res., Edinburgh, U.K., 2002.

[12] K. Richmond, "A trajectory mixture density neural network for the acoustic-articulatory inversion mapping," in INTERSPEECH, Pittsburgh, Pennsylvania, USA, September 2006, pp. 577–580.

[13] N. Hogan, "An organizing principle for a class of voluntary movements," J. Neuroscience, vol. 4, no. 11, pp. 2745-2754, 1984.

[14] G. Ananthakrishnan and O. Engwall., "Mapping between acoustic and articulatory gestures," Speech Communication, vol. 53, no. 4, pp. 567–589, 2011.

[15] P. K. Ghosh and S. S. Narayanan, "A computational framework for exploring the role of speech production in speech processing from a communication system perspective," University of Southern California, Ph.D. Thesis, 2011.

[16] A. Toutios and K. Margaritis, "A rough guide to the acoustic-to-articulatory inversion of speech," in 6th Hellenic European Conference of Computer Mathematics and its Applications, HERCMA, pp. 1–4, 2003.

[17] Z. Wu, et al. "Acoustic to articulatory mapping with deep neural network," Multimedia Tools and Applications, pp.1-19, 2014.

[18] A. Wrench and W. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition," in Proc. *5th* Seminar of Speech Production (SSP), KlosterSeeon, 2000, pp. 305–308.

[19] S. Al Moubayed and G. Ananthakrishnan, "Acoustic-to-articulatory inversion based on local regression," in Proc. INTERSPEECH, Makuhari, Chiba, Japan, September 2010, pp. 937– 940.

[20] A. Toth and A. Black, "Using articulatory position data in voice transformation," ISCA Speech Synthesis Workshop (SSW6), pp. 182-187, 2007

[21] D. Felps, C. Geng, and R. Gutierrez-Osuna, "Foreign accent conversion through concatenative synthesis in the articulatory domain," IEEE Trans. on Audio, Speech, and Language Processing, vol. 20, no. 8, pp. 2301-2312, 2012.

[22] Z. H. Ling, K. Richmond, J. Yamagishi, and R. H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," IEEE Trans. on Audio, Speech, and Language Processing, vol. 17, no. 6, pp. 1171-1185, 2009.

[23] P. Ghosh and S. Narayanan, "Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion," *J. Acoust. Soc. Am.*, vol. 130, no. 4, pp. EL251-EL257, 2011.

[24] S. Aryal, and R. Gutierrez-Osuna. "Articulatory inversion and synthesis: towards articulatory-based modification of speech," in Proc. Int'l Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, Canada, pp. 7952-7956.

[25] A. W. Black, et al. "Articulatory features for expressive speech synthesis," in Proc. Int'l Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, Japan, pp. 4005-4008.