FROM ACOUSTICS TO VOCAL TRACT TIME FUNCTIONS

Vikramjit Mitra[1], *İ Yücel Özbek*[2], *Hosung Nam*[3], *Xinhui Zhou*[1], *Carol Y. Espy-Wilson*[1]

[1]**Department of Electrical and Computer Engineering**, University of Maryland, College Park, MD
[2]**Department of Electrical and Computer Engineering**, Middle East Technical University, Turkey
[3]**Haskins Laboratories**, New Haven, CT
[1]{vmitra@umd.edu, zxinhui@umd.edu, espy@umd.edu}, [2]{iozbek@illinois.edu},
[3]{nam@haskins.yale.edu}

ABSTRACT

In this paper we present a technique for obtaining Vocal Tract (VT) time functions from the acoustic speech signal. Knowledge-based Acoustic Parameters (APs) are extracted from the speech signal and a pertinent subset is used to obtain the mapping between them and the VT time functions. Eight different vocal tract constriction variables consisting of five constriction degree variables, lip aperture (LA), tongue body (TBCD), tongue tip (TTCD), velum (VEL), and glottis (GLO); and three constriction location variables, lip protrusion (LP), tongue tip (TTCL), tongue body (TBCL) were considered in this study. The TAsk Dynamics Application model (TADA [1]) is used to create a synthetic speech dataset along with its corresponding VT time functions. We explore Support Vector Regression (SVR) followed by Kalman smoothing to achieve mapping between the APs and the VT time functions.

Index Terms— *Speech inversion, Support Vector Regression, vocal tract time functions, Acoustic-to-articulatory inversion.*

1. INTRODUCTION

Acoustic-to-articulatory inversion of speech has received a great deal of attention from researchers for the past 35 years. Kirchhoff [2] has demonstrated that articulatory features can significantly improve the performance of an automatic speech recognition (ASR) system when the speech is noisy. In fact she has shown that this effectiveness increases with a decrease in the Signal-to-Noise ratio (SNR). Articulatory information is also useful for speech synthesis, speech therapy, language acquisition, speech visualization and extraction of information about vowel lengthening [3] and prosodic stress [4].

Most of the current work on acoustic-to-articulatory inversion is based on the data acquired from Electromagnetic Mid-sagittal Articulography (EMMA) or Electromagnetic Articulography (EMA) [5]. A huge collection of data is available from the MOCHA [6] and the Microbeam [7] databases. Most of the research on acoustic-to-articulatory inversion [8, 9] has used these corpora. Although these databases contain natural speech and have various effects like speaker and gender variability, they are often contaminated with measurement noise and are not suitable for studying gestural and prosodic variability. The TAsk Dynamics Application model (TADA [1]), on the other hand, is completely free from measurement noise and is designed such that it generates VT time functions similar to that obtained from EMA or EMMA; moreover it has a greater degree of flexibility in adding gestures, prosodic stress etc. such that their effects on the VT time functions can be observed.

Speech recognition models have suffered from poor performance in casual speech because of the significant increase in acoustic variations relative to that observed in clearly articulated speech. This problem can be attributed to the intrinsic limitation of the phone unit used in many systems. While phone units are distinctive in the cognitive domain, they are not invariant in the physical domain. Further, phone-based ASR systems do not adequately model the temporal overlap that occurs in more casual speech. In contrast to segment-based phonology and phone-based recognition models, articulatory phonology proposed the articulatory constriction *gesture* as an invariant action unit and argues that human speech can be decomposed into a constellation of *articulatory gestures* [10, 11] allowing for temporal overlap between neighboring gestures. Thus, in this framework, acoustic variations can be accounted for by gestural coarticulation and reduction. Recently, some speech recognition models [12] using articulatory gestures as units have been proposed as an alternative to traditional phone-based models. Also, Zhuang et. al. [13] proposed an instantaneous gestural pattern vector and a statistical method to predicting these gestural pattern vectors from VT time functions. The VT time functions are time-varying physical realizations of gestural constellations at the distinct vocal tract sites for a given utterance. This study aims to predict the VT functions from acoustic signals as a component model in a complete gesture-based speech recognition system. The prediction of the VT time function from the acoustic speech signal is performed by Support Vector Regression (SVR). The SVR output is often noisy; hence a Kalman-filter based post processor is used to smooth the reconstructed VT time function.

The organization of the paper is as follows: Section 2 briefly describes VT time functions and how they are obtained in this study; Section 3 describes the proposed Support Vector Regression (SVR) based mapping model; Section 4 presents the results obtained followed by the conclusion and future work in Section 5.

2. VOCAL TRACT (VT) TIME FUNCTIONS

Gestures are primitive units of a produced word and represent constricting motions at distinct constricting devices/organs along the vocal tract, which are lips, tongue tip, tongue body, velum, and glottis. The constriction is the task goal of each gesture and can be described by its location and degree. Since the constriction in the glottis and velum are not varied in location, it is defined by degree only. Gestures can be defined in eight VT constriction variables as shown in Table 1. When a gesture is active in each VT variable, it is distinctively specified by such dynamic parameters as constriction target, stiffness, and damping. The gestures are allowed to temporally overlap with one another within and across tract variables. Note that even when a tract variable does not have an active gesture, the resulting tract variable time function can be varied passively by another tract variable sharing the same articulator. For example, TTCD with no active gesture can also change when there is an active gesture in LA because LA involves jaw articulator movement and at the same time it passively changes TTCD since they share the jaw articulator. *A priori* knowledge about these functional dependencies along with data driven correlation information can be used to effectively design the mapping process from acoustics to VT time functions. The *task-dynamic model* of speech production [14] employs a constellation of gestures with dynamically specified parameters, i.e. *gestural scores*, as a model input for an utterance. The model computes task-dynamic speech coordination among the articulators, which are structurally coordinated with the gestures along with the time function of the physical trajectories for each VT-variable. The time function of model articulators is input to the vocal tract model [15] and then the model computes the area function and the corresponding formants. Given English text or ARPABET, TADA [1] (Haskins laboratories articulatory speech production model that includes the task dynamic model and vocal tract model) generates input in the form of formants and VT time functions for HLsyn™ (a parametric quasi-articulator synthesizer, Sensimetrics Inc.). The TADA output files are then manually fed to HLsyn™ to generate acoustic waveform. The dataset generated for this study consists of VT trajectories (sampled at 5 msec) and corresponding acoustic signals for 363 words, which were chosen from the Wisconsin X-ray microbeam data [7] and identical to that used in [13].

3. THE PROPOSED MAPPING ARCHITECTURE

The acoustic speech signal is converted to acoustic parameters (APs) [16,17,18] (e.g. formant information, mean Hilbert envelope, energy onsets and offsets, periodic and aperiodic energy in subbands [19] etc.). The APs are measured at a frame interval of 5 msec (hence synchronized properly with the VT time functions). The APs are then normalized to have zero mean and unity standard deviation. Altogether 53 APs were considered for the proposed task. A subset of these APs was selected for each of the VT time functions based upon their relevance. Relevance is decided based on: (1) Knowledge about the attributes of speech that is well reflected by a particular AP and (2) manual observation of the variation of the APs with respect to each of the VTs, supported by their correlation information. Some APs may be uncorrelated with certain VT time functions. In addition, there may be strong cross-correlation among a certain number of APs which may render them as redundant for a specific VT time function. In this case, the AP with the strongest correlation with the respective VT time function was selected and the others were discarded. Moreover certain VT time functions (TTCL, TBCL, TTCD and TBCD) are known to be functionally dependent upon other VT time functions and can be represented by equation 1, where as the remaining four VTs (GLO, VEL, LA and LP) are relatively independent and can be obtained directly from the APs.

Table 1. *Constriction organ, vocal tract variables & involved model articulators*

Constriction organ	VT variables	Articulators
Lip	Lip Aperture (LA)	Upper lip, lower lip, jaw
	Lip Protrusion (LP)	
Tongue Tip	Tongue tip constriction degree (TTCD)	Tongue body, tip, jaw
	Tongue tip constriction location (TTCL)	
Tongue Body	Tongue body constriction degree (TBCD)	Tongue body, jaw
	Tongue body constriction location (TBCL)	
Velum	Velum (VEL)	Velum
Glottis	Glottis (GLO)	Glottis

$$\begin{aligned} &f_{TTCL} : TTCL \leftarrow (AP, LA) \\ &f_{TBCL} : TBCL \leftarrow (AP, LA) \\ &f_{TTCD} : TTCD \leftarrow (AP, TTCL, TBCL, LA) \\ &f_{TBCD} : TBCD \leftarrow (AP, TBCL, LA) \end{aligned} \qquad (1)$$

where *AP* denotes the set of pertinent APs for that specific VT time function. The *ε*-SVR [20] (which is a generalization of the Support Vector Classification algorithm) works for only single output. *ε*-SVR uses the parameter *ε* (the unsusceptible coefficient) to control the number of support vectors. The main advantage of SVR is that it projects the input data into a high dimensional space via non-linear mapping and then performs linear regression in that space. For the 8 VT time functions, 8 different *ε*-SVRs were created and equation 1 suggests that some *ε*-SVRs need to be created before the other. For example, LA needs to be created first followed by TTCL, TBCL and finally followed by TTCD and TBCD. Based upon the knowledge-based information regarding the VT time functions GLO, VEL and LP can be considered relatively independent of the others. Each of the VT time functions are centered at zero and scaled by 4 times the standard deviation so that most of them fall in the interval (-1,1) (this processing is similar to [8] and is pertinent for LibSVM *ε*-SVR implementation). Table 2 shows the number of pertinent APs for each VT, their optimal context and the input dimension of their corresponding *ε*-SVRs. For each of the VT time function 5 different *ε*-SVRs were created for 5 different contextual windows: 5, 6, 7, 8 and 9.

Table 2. *Pertinent APs for each VT*

VT time function	Number of APs	Optimal Context	Input Dimension (d)
GLO	15	6	195
VEL	20	7	300
LP	15	6	195
LA	23	8	391
TTCL	22	7	345
TTCD	22	5	275
TBCL	18	5	209
TBCD	18	6	260

The optimal contextual window is obtained for the case where the least mean square error (MSE) is obtained from the *ε*-SVR. For a context-window of length *N*, *N* frames are selected before and after the current frame with a frame shift of 2 (time shift of 10 msec) between the frames giving rise to a vector of size (2*N*+1)*d*, where *d* is the dimension of the input feature space. It should be noted that *d* is different for TTCL, TBCL, TTCD and TBCD. For example in the case of TTCD, *d* is the sum of the number of pertinent APs (=22) and the number of VTs (=3) upon which TTCL is dependent, (refer to equation 1) which is 25. Prior research [8] has shown that the Radial basis function (RBF) kernel with $\gamma = 1/d$, and C = 1 [21] is near optimal for the proposed task. However, given that the optimal context window is known, C is varied between 0.5, 1 and 1.5, to select the best configuration based upon the MSE from ε-SVR. The final ε-SVR configuration is evaluated against three separate training-test sets to obtain cross-validation performances and error bounds for the proposed system. The dataset is split into 5:1 for training and test sets. The overall hierarchical system is shown in Fig. 1, where independent VT time functions are obtained first and the dependent ones are obtained later. The output from the ε-SVR are noisy due to estimation error. An averaging filter using a window of 7 samples was initially used to smooth the reconstructed VT time functions.

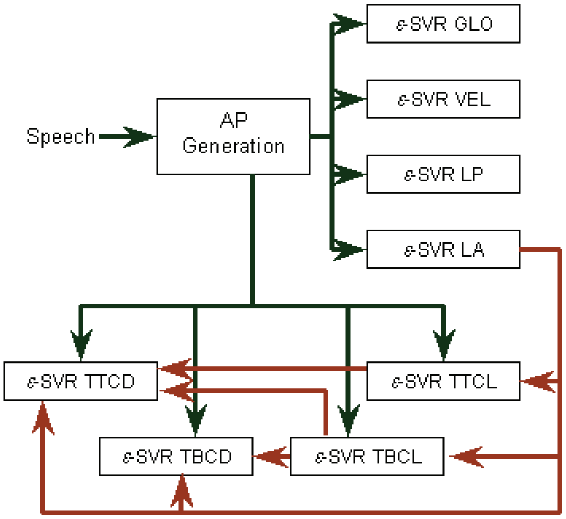

Fig. 1. ε-SVR architecture for generating the VT time functions

It was observed that smoothing the estimated VT time functions improved estimation quality and reduces root mean square error (RMSE). This led to the use of a Kalman Smoother as the post processor for the reconstructed VT time functions from ε-SVR. Since articulatory trajectories are physical quantities, they can be approximately modeled as the output of a dynamic system. For the proposed architecture, we selected the following state-space representation

$$\begin{aligned} x_k &= Fx_{k-1} + w_{k-1} \\ y_k &= Hx_k + v_k \end{aligned} \qquad (2)$$

with the following model parameters

$$F = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \text{ and } H = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

$$\begin{aligned} x_0 &\sim \mathrm{N}(x_0, \bar{x}_0, \Sigma_0) \\ w_k &\sim \mathrm{N}(w_k, 0, Q) \\ v_k &\sim \mathrm{N}(v_k, 0, R) \end{aligned} \qquad (3)$$

T is the time difference (in ms) between two consecutive measurements, $x_k=[x_k^p \quad x_k^v]^T$ is the state vector and contains the position and velocity of the VT time function at time instant k. y_k is the output of the ε-SVR estimator which is considered as noisy observation of the first element of the state x_k. The variables w_k and v_k are process and measurement noise, which have zero mean, known covariance Q and R, and they are considered to be Gaussian. The goal is to find the smoothed estimate of the state $x_{k|N}$ given the observation sequence $Y = \{y_1,\ldots,y_n\}$, i.e, $x_{k|N} = E[x_k \mid y_1,\ldots, y_N]$. Although, F and H are known parameters of the state space representation, the unknown parameter set $\Theta = \{Q, R, \bar{x}_0, \Sigma_0\}$ should be learnt from the training dataset. After learning the unknown parameter set $\Theta = \{Q, R, \bar{x}_0, \Sigma_0\}$ the smoothed state $x_{k|N}$ is estimated by the Kalman Smoother in optimal sense. It is observed that smoothing reduces the RMSE of the reconstructed VT time functions.

4. RESULTS

The parameters of the ε-SVR and the optimal context window were obtained using a single test-train set and then the remaining 2 test-train sets were used with the same configuration to obtain the error bounds. The results obtained from ε-SVR, after the averaging filter and Kalman smoothing is shown in Table 3. It should be noted that for GLO and VEL, the VT time function values are in terms of abstract numbers; hence the RMSE doesn't have a unit. The values of LP, LA, TBCD and TTCD are in terms of mm, hence the RMSE is in terms of mm, and finally TTCL and TBCL are in degrees, hence the RMSE is in terms of degree. In Table 3, the entries correspond to the average RMSE across the three test-train sets and (+N / -M) entries in the lower row depict the maximum and the minimum deviation from the average RMSE. Table 3 shows that Kalman smoothing offered better RMSE than average smoothing. The Kalman smoothing is also found to offer a tighter bound in most of the cases and, on average offers a 9.44% reduction in the RMSE over the unprocessed ε-SVR output. This RMSE reduction is significantly better than the 3.94% offered by the averaging filter. Table 4 presents the correlation coefficient of the ε-SVR reconstructed VT time functions, which indicates the similarity in shape and trajectory between the actual and the reconstructed VT time functions. Fig. 2 shows the plot of the actual and reconstructed (Kalman smoothed) VT time function. RMSE of GLO and VEL are found to be very low, to analyze their result, the fraction of cases where open/close is missed or falsely detected for GLO and VEL was obtained and it was found to be 4.6% for GLO and 2.9% for VEL.

Table 3. *Average RMSE for the different VTs*

VT time function	RMSE		
	ε-SVR	after averaging filter	after Kalman smoothing
GLO	0.039 (+0.004/-0.002)	0.040 (+0.003/-0.002)	0.036 (+0.004/-0.003)
VEL	0.025 (+0.002/-0.003)	0.025 (+0.002/0.003)	0.023 (+0.002/-0.003)
LP	0.565 (+0.007/-0.007)	0.536 (+0.011/-0.012)	0.508 (+0.018/-0.016)
LA	2.361 (+0.091/-0.063)	2.227 (+0.107/-0.084)	2.178 (+0.115/-0.091)
TTCD	3.537 (+0.075/-0.118)	3.345 (+0.089/-0.067)	3.253 (+0.073/-0.079)
TBCD	1.876 (+0.129/0.139)	1.749 (+0.138/-0.158)	1.681 (+0.141/-0.162)
TTCL	8.372 (+0.263/-0.0.257)	8.037 (+0.285/-0.329)	7.495 (+0.221/-0.266)
TBCL	14.292 (+1.319/-1.895)	13.243 (+1.465/-1.921)	12.751 (+1.313/-1.829)

Table 4. *Correlation coefficient for each VT obtained from ε*-SVR

GLO	VEL	LP	LA	TTCD	TBCD	TTCL	TBCL
0.951	0.944	0.754	0.745	0.889	0.857	0.849	0.849

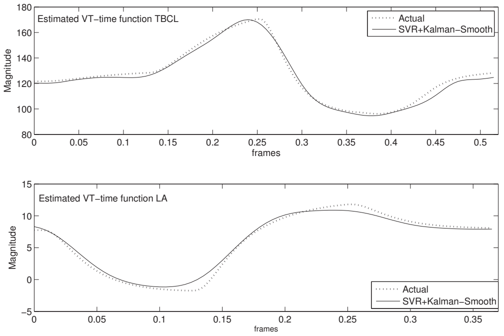

Fig. 2. *Overlaying plot of the actual VT along with the ε*-SVR output followed by Kalman smoothing *for TBCL and LA*

5. CONLUSION

This paper demonstrated the use of ε-SVR algorithm to obtain VT time functions from the acoustic signal. The ε-SVR parameters are optimized for each VT time functions. It is observed from Tables 3 and 4 that the ε-SVR corresponding to the independent VT time functions GLO and VEL offered the least RMSE and best correlation coefficient, indicating best estimation. RMSE of TTCL and TBCL may seem to be high compared to the others; however they represent the RMSE in degrees. LP, LA, TTCD and TBCD are measured in millimeters; hence their RMSE is in mms. On average, the Kalman smoothing reduced the RMSE of the reconstructed data by 9.44%.

Future work should consider improving the performance of the SVR by using a data driven Kernel. Currently, TADA outputs are fed manually to HLsyn to obtain the synthetic speech. Future research should automate the process so that more data can be generated to appropriately estimate the robustness of the proposed architecture. The mapping should also be evaluated in noisy scenarios, where noise at different signal-to-noise ratios is added to the speech signal and the effect of the noise on the reconstructed VT time functions is observed and evaluated in terms of RMSE. Spectral parameters like MFCCs have been used for a similar task in [22], however SVRs were not used in such a setup. Future research should compare such features with APs using the same SVR framework to see the difference in performance.

6. ACKNOWLEDGEMENT

This research was supported by NSF grant IIS0703859 and NIH grant DC-02717. We also would like to acknowledge the helpful advice from Elliot Saltzman (BU), Mark Hasegawa-Johnson (UIUC) and Louis Goldstein (USC).

7. REFERENCES

[1] H. Nam, L. Goldstein, E. Saltzman and D. Byrd, “Tada: An enhanced, portable task dynamics model in matlab”, *Journal of the Acoustical Society of America*, Vol. 115, Iss. 5, pp. 2430, 2004.

[2] K. Kirchhoff, “Robust Speech Recognition Using Articulatory Information”, *PhD Thesis*, University of Bielefeld, 1999.

[3] D. Byrd, “Articulatory vowel lengthening and coordination at phrasal junctures”, *Phonetica*, 57 (1), pp. 3-16, 2000.

[4] T. Cho, “Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /A, i/ in English”, *Journal of the Acoustical Society of America*, 117 (6), pp. 3867-3878, 2005.

[5] J. Ryalls and S. J. Behrens, *Introduction to Speech Science: From Basic Theories to Clinical Applications*, Allyn & Bacon, 2000.

[6] A.A. Wrench and H.J. William, “A multichannel articulatory database and its application for automatic speech recognition”, In 5th Seminar on *Speech Production: Models and Data*, pp. 305–308, Bavaria, 2000.

[7] Westbury “X-ray microbeam speech production database user’s handbook”, Univ. of Wisconsin, 1994

[8] A. Toutios and K. Margaritis, “A Support Vector Approach to the Acoustic-to-Articulatory Mapping”, In Proceedings of *Interspeech*, Eurospeech-2005, pp. 3221-3224, Portugal, 2005.

[9] K. Richmond, “Estimating Articulatory Parameters from the Speech Signal”, *PhD thesis*, The Center for Speech Technology Research, Edinburgh, 2002.

[10] C. Browman and L. Goldstein, “Articulatory Gestures as Phonological Units”, *Phonology*, 6: 201-251, 1989

[11] C. Browman and L. Goldstein, “Articulatory Phonology: An Overview”, *Phonetica*, 49: 155-180, 1992

[12] K. Livescu, O. Cetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezman, S. Dawson-Haggerty, B. Woods, J. Frankel, M. Magimai-Doss and K. Saenko, “Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop,” in Proc. *ICASSP*, Hawaii, U.S.A., 2007.

[13] X. Zhuang, H. Nam, M. Hasegawa-Johnson, L. Goldstein and E. Saltzman, “The Entropy of Articulatory Phonological Code: Recognizing Gestures from Tract Variables”, In Proceedings of *Interspeech* 2008, pp. 1489-1492, 2008.

[14] L. Saltzman and K. Munhall, “A Dynamical Approach to Gestural Patterning in Speech Production”, *Ecological Psychology* 1(4): 332-382, 1989.

[15] K. Iskarous, L. Goldstein, D. Whalen, M. Tiede and P. Rubin, “CASY: the Haskins configurable articulatory synthesizer”, 15th *International Congress of Phonetic Sciences*, Universitat Autònoma de Barcelona, Barcelona, Spain, 2003.

[16] A. Juneja, “Speech recognition based on phonetic features and acoustic landmarks”, *PhD thesis*, University of Maryland College Park, December 2004.

[17] K. Stevens, S. Manuel and M. Matthies, “Revisiting place of articulation measures for stop consonants: Implications for models of consonant production”. Proceedings of *International Congress of Phonetic Science*, Vol-2, pp. 1117-1120, 1999.

[18] S. Chen and A. Alwan, “Place of articulation cues for voiced and voiceless plosives and fricatives in syllable-initial position”, Proceedings of *ICSLP*, vol.4, 113-116, 2000.

[19] O. Deshmukh, C. Espy-Wilson, A. Salomon and J. Singh, “Use of Temporal Information: Detection of the Periodicity and Aperiodicity Profile of Speech”, *IEEE Trans. on Speech and Audio Process*ing, Vol. 13(5), pp. 776-786, 2005.

[20] C. Chang and C. Lin, “LIBSVM: a library for support vector machines, 2001”, http://www.csie.ntu.edu.tw/~cjlin/libsvm

[21] J. Weston, A. Gretton and A. Elisseeff, “SVM practical session – How to get good results without cheating”, *Machine Learning Summer School*, Tuebingen, Germany, 2003.

[22] A. Lammert, D.P. Ellis and P. Divenyi, “Data-driven articulatory inversion incorporating articulator priors”, *Statistical and Perceptual Audition*, Brisbane, AU, pp.29-34, 2008.