SUPPORT VECTOR REGRESSION FOR TONGUE POSITION INFERENCE

Alexander Sepulveda[1] and G. Castellanos-Dominguez

[1]ALEXANDER SEPULVEDA

Alexander got his undergraduate degree in Electronic Engineering, his M.Eng. degree in Industrial Automation from the Universidad Nacional de Colombia Sede Manizales, in 2002, 2004 respectively. During two and a half years he served as a Collegiate Assistant Professor at Universidad Autónoma de Manizales. He is currently working as assistant professor at Escuela Colombiana de Carreras Industriales, Bogotá-Colombia. He is interested in the relationship between the articulatory and acoustic domains. His PhD work is a study about the importance of the acoustic speech representation in acoustic-to-articulatory inversion systems. His interests include time-frequency analysis, articulatory and acoustic phonetics, vocal tract modeling, acoustic-to-articulatory inversion, statistical learning.

ABSTRACT

The articulatory inversion task consists on recovering the articulators' position or the vocal tract shape from the acoustic speech signal. The availability of large corpora of parallel acoustic and articulatory data has made possible the use of data-driven methods as an alternative for the solution of the speech inversion problem. This paper presents a method for the inference of tongue positions based on support vector regression techniques. The acoustic speech signal is parametrized by using perceptual linear prediction coefficients (PLP); then, a nonlinear transformation function is applied to the regressors. Model assessment is performed by measuring the similarity between the estimated and the reference signals and by measuring the correlation between inputs and residuals. The proposed method shows to be promising.

Keywords: acoustic-to-articulatory mapping, transformation of regressors, support vector regression, perceptual linear prediction coefficients.

INTRODUCTION

An adequate system for recovering the articulators' position from the acoustic speech signal is useful for several tasks: applications based on computer animated talking heads [1] (for example, computer guided second language learning programs and visual aids for articulatory training tasks in hearing for speech impaired children); low-bit rate coding due to the relatively slow movement of articulators [2] and; complementing representation in speech recognition systems to improve their performance because of its ability to represent in a better way co-articulatory related phenomena [3]. Even though acoustic-to-articulatory inversion offers a wide range of potential applications, its development still remains an open challenge [4]. In this line, several data-driven based methods for the acoustic-to-articulatory inversion mapping have been proposed. In [5], an inversion mapping using an acoustic-to-articulatory codebook is discussed. In [6] mapping models based on neural networks are suggested. The use of models based on HMMs (Hidden Markov Models) is proposed in [7], where, not only the acoustic features and articulatory parameters are used, but also the phonetic information. Support vector regression (SVR) techniques have been used in [8, 18].

Gaussian mixture models (GMM) with maximum likelihood estimation (MLE) are applied in [9] to determine the articulatory trajectories. They use static as well as dynamic features in order to reduce the presence of unnatural movements in the estimated trajectory. In the above related works the major part of error was obtained on tongue modelling.

METHOD

1. Database

TIMIT sentences are designed to provide phonetically diverse material in order to maximize the usefulness of the data for speech technology and speech science research purposes. The MOCHA-TIMIT database is composed by two speakers; however, in this work only the female speaker (fsewo) is used. It is composed of 460 short phrases from which 368 files are included in the training set. The rest (recordings that end in 2 and 6) are used for testing.

The MOCHA database includes four data streams recorded concurrently: the acoustic waveform (16 kHz sample rate, with 16 bit precision), laryngograph, electropalatograph and electromagnetic articulograph (EMA) data. Movements of receiver coils attached to the articulators are sampled by the electromagnetic articulograph at 500 Hz. Coils were affixed to the lower incisors (li), upper lip (ul), lower lip(ll), tongue tip (tt), tongue blade (tb), tongue dorsum (td) and velum (v), see Fig. (1). The two coils at the bridge of the nose and upper incisors are used to provide reference points in order to correct errors produced by head movements.

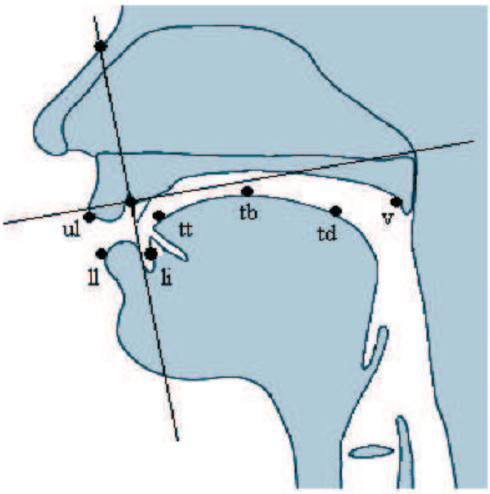

Figure 1. Positions of the EMA contacts.

Label files of MOCHA database were used to discard silent segments at the beginning and end of the utterances. The EMA trajectories are then re-sampled from 500 Hz to 100 Hz. Since the articulators move relatively slowly, crucial information is not lost. Present study is fucoused on the voiced segments of speech. The algorithm for the voiced/unvoiced segmentation is the one explained in [12].

2. REPRESENTATION OF ACOUSTIC SPEECH SIGNAL

Since the EMA data used in this work is sampled at 100 Hz, the speech parametrization is done at a rate of 10 ms. Parameters are calculated on 16 ms width. The articulatory configuration at the current time depends on the context, thus it is desirable to use additional frames such that the regression system takes into account the adjacent information.

For the selection of the size of the context-window the results obtained in [13] are used. Results revealed that along the time axis, individual features of more than 100 ms in the past or in the future are not very relevant for the phoneme category classification at the current time. In addition, the authors of the mentioned work showed that the major part of the information is contained in the range [-80ms; +80ms] around the current time. However, the use of all frames generates a large number of inputs; and, large input sets affect the variance of the output estimation [11]. The input feature set is formed by a log-energy value and 12 PLP coefficients measured for every frame colored gray in Figure 2. PLP coefficients are estimated by using HTK Toolkit.

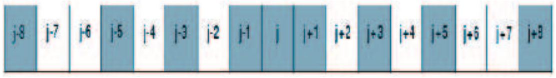

Figure 2. Configuration of the input feature set. Shaded regions indicate the frames in which the log-energy and 12th order PLP features are calculated

3. DATA PREPROCESSING.

A nonlinear transformation is applied to input variables for the sake of alleviating the influence of outliers. In order to obtain a measure of the portion of outliers the kurtosis function may be used. A value of K = 3 is the value for the Gaussian probability density function and values of kurtosis greater than K = 3 implies a greater degree of outliers. Those inputs that have a kurtosis value greater than the threshold K > 6 are transformed by the tangent sigmoidal function,

$$T(x) = a\frac{e^{bx} - e^{-bx}}{e^{bx} + e^{-bx}} \qquad (1)$$

where the values of $a = 1.716$ and $b = 1.5$ are selected such that kurtosis values from *15* and *8* are shifted into the interval *[4.7, 2.56]*. The input transformation is applied in those cases where $K > 6$ and a skewness $S < 0{:}5$.

4. SUPPORT VECTOR REGRESSION

Support vector regression (SVR) is a supervised linear regression method (from the perspective of regression analysis) that can be used to model nonlinear functions. Similar to neural networks, SVR attempts to map the inputs to the outputs of the training data. Estimating the unknown parameters involves the optimization of a convex cost function.

SVR training is carried out by using the SVM-Light software [14]. The radial basis kernel was used throughout. The parameter corresponding to generalization capacity is selected equal to 1. This is a tradeoff between the empirical error and the prediction error. It tends to give the same importance to the training error and the prediction error. In the other hand, the parameter corresponding to the insensitive width is adjusted by using the approximations given in [15]. Finally, in order to adjust the parameter of the RBF kernel, the methodology explained in [16] is used.

5. MODEL ASSESSMENT

The results are organized depending on the kind of behavior of the model we are assessing. First, it is expected the estimated trajectories be similar to reference signals, in this case we use measures of similarity; second, the residual signals should not be correlated with the inputs.

Mean square error and correlation measure are the most common measures used for assessing articulatory inversion performance. The former value is defined as follows,

$$E_{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \qquad (2)$$

where N is the number of input-output vector pairs. Nonparametric correlation measures are more robust than linear correlation measures; and, they are more resistant to unplanned defects in the data [17]. Spearman correlation measure is used in this work.

The ideal situation is when the predicted values $\hat{y}$ are capable of explaining a major part of the actual output. The ratio

$$R_y^2 = 1 - \frac{\sum_{t=1}^{N} e(t)}{\sum_{t=1}^{N} y^2(t)} \qquad (3)$$

known as multiple regression coefficient, measures the proportion of the total variation of y that is explained by the regression model.

It is expected that the residual error be statistically independent from inputs. If the correlation function has high values for different time lags, then it means that he model does not completely represent the behavior of the system [19].

RESULTS

For the six tongue channels the following variables are measured: mean square error in mm, the correlation value between estimated and measured output, the multiple regression coefficient and the maximum value among the set of correlation values between the error and the input regressors. The results are shown in table (1). An example of the predicted trajectories are shown in Figure 3.

Table (1) shows that exists some level of correlation between inputs and the residuals resulting from the regression process. This affirmation is confirmed by using the Brown-Forsythe test, where the variance of the residual error signal varies some input variables. Same table shows that about 1.3 MSE (Emm) is obtained when estimating articulators position; which, is less than the values reported in recent works like [6] and [9].

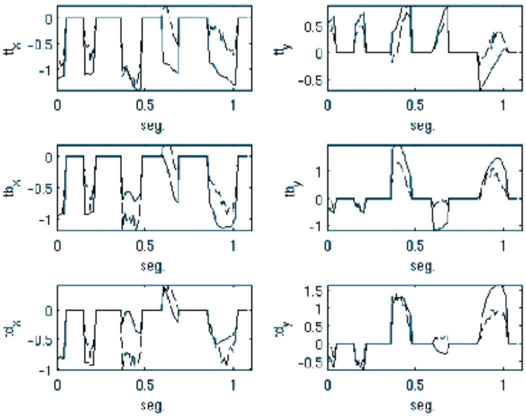

Figure 3. EMA positions for the chanels ttx, ttx, tbx, tby, tdx, tdy. The estimated one correspond to the dashed lines. The acoustic signal correspond to the south british utterance of the phrase *Is this seesaw safe?*, the first phrase in the testing set.

CONCLUSIONS

Present work shows a promising method for the inference of articulators position, which is based on the combination of support vector regression and the adequate selection of regressors. However, the residual error signals are correlated to the inputs; thus, there is a part of the input-output relation that has not been explained yet. That is, further improvements could be carried out.

channel	Emm	R2y	corr.	max. Corr.
ttx	1.32	0.61	0.79	0.15
tty	1.21	0.83	0.83	0.19
tbx	1.26	0.81	0.81	0.17
tby	1.04	0.86	0.86	0.18
tdx	1.28	0.80	0.80	0.16
tdy	1.20	0.79	0.79	0.09

Table 1. Results for the PLP feature set.

REFERENCES

[1] O. Engwall, \Vocal tract modelling in 3D," TMH-QPSR, vol. 40, no. 1-2, pp. 31-38, 1999.

[2] Advances in Speech Signal Processing, ch. Speech coding based on physiological models of speech production. Marcel Decker, 1992.

[3] J. Frankel and S. King, Speech recognition using linear dynamic models," IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 1, pp. 246-256, 2007.

[4] B. Potard, Y. Laprie, and S. Ouni, Incorporation of phonetic constraints in acoustic-to-articulatory inversion," Journal of Acoustical Society of America, vol. 123, no. 4, pp. 2310-2323, 2008.

[5] J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin, and E. Saltzman, Accurate recovery of articulator positions from acoustics: new conclusions based on human data," Journal of Acoustical Society of America, 1996.

[6] K. Richmond, S. King, and P. Taylor, Modelling the uncertainty in recovering articulation from acoustics," Computer, Speech & Language, vol. 17, pp. 153-172, 2003.

[7] L. Zhang and S. Renals, Acoustic-articulatory modeling with the trajectory HMM," IEEE Signal Processing Letters, vol. 15, pp. 245{248, 2008.

[8] A. Toutios and K. Margaritis, Contribution to statistical acoustic-to-EMA mapping," in 16th European Signal Processing Conference (EUSIPCO-2008), 2008.

[9] T. Toda, A. Black, and K. Tokuda, Statistical mapping between articulatory movements and acoustic spectrum using gaussian mixture models," Speech Communication, 2008.

[10] C. Qin, and M. Carreira-Perpiñan, A comparison of acoustic features for articulatory inversion," in InterSpeech.

[11] Moder Regression Methods. John Wiley & Sons, 1997.

[12] Speech Processing and Synthesis Toolboxes. John Wiley & Sons, 2000.

[13] H. Yang, S. Vuuren, S. Sharma, and H. Hermansky, \Relevance of time-frequency features for phonetic and speaker channel classification," Speech Communication, vol. 31, pp. 35-50, 2000.

[14] Adcances in Kernel Methods – Support Vector Learning, ch. Making large-Scale SVM Learning Practical. MIT Press, 1999.

[15] V. Cherkassky and Y. Ma, Practical selection of SVM parameters and noise estimation for SVM regression," Neural Networks, 2004.

[16] e. a. Wanga, Wenjian, Determination of the spread parameter in the Gaussian kernel for classification and regression," Neurocomputing, 2003.

[17] Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, 2001.

[18] F. A. Sep_ulveda, G. Castellanos ,J. I. Godino-Llorente, Acoustic Analysis of the Stop Consonants for Detecting Hypernasal Speech," in 4th International Symposium on Image/ Video Communications, ISIVC2008, Bilbao-Spain.

[19] System Identification: Theory for the user. Prentice Hall PTR, 1999.