Learning vocal tract variables with multi-task kernels

Hachem Kadri, Emmanuel Duflos, Philippe Preux

LEARNING VOCAL TRACT VARIABLES WITH MULTI-TASK KERNELS

Hachem Kadri[1], *Emmanuel Duflos*[1,2], *Philippe Preux*[1,3]

[1]Team-Project SequeL, INRIA Lille Nord-Europe, Villeneuve d'Ascq, France
[2]LAGIS UMR 8146, Ecole Centrale de Lille, Villeneuve d'Ascq, France
[3]LIFL UMR 8022, Université de Lille, Villeneuve d'Ascq, France
{hachem.kadri, emmanuel.duflos, philippe.preux}@inria.fr

ABSTRACT

The problem of acoustic-to-articulatory speech inversion continues to be a challenging research problem which significantly impacts automatic speech recognition robustness and accuracy. This paper presents a multi-task kernel based method aimed at learning Vocal Tract (VT) variables from the Mel-Frequency Cepstral Coefficients (MFCCs). Unlike usual speech inversion techniques based on individual estimation of each tract variable, the key idea here is to consider all the target variables simultaneously to take advantage of the relationships among them and then improve learning performance. The proposed method is evaluated using synthetic speech dataset and corresponding tract variables created by the TAsk Dynamics Application (TADA) model and compared to the hierarchical ε-SVR speech inversion technique.

Index Terms— Multi-task learning, matrix-valued kernel, vocal tract variables, acoustic-to-articulatory inversion.

1. INTRODUCTION

The problem of speech inversion has received increasing attention in the speech processing community in the recent years (see [1, 2] and references therein). This problem is motivated by several applications in which it is required to estimate articulatory parameters from the acoustic speech signal [3, 4, 5]. For example, in speech recognition, the use of articulatory information has been of interest since speech recognition efficiency can be significantly improved [6]. This is due to the fact that automatic speech recognition (ASR) systems suffer from performance degradation in the presence of noise and spontaneous speech. Moreover, acoustic-to-articulatory speech inversion is also useful in many other interesting applications such as speech analysis and synthesis [7].

Most of current research on acoustic-to-articulatory inversion focuses on learning Electromagnetic Articulography (EMA) trajectories from acoustic parameters and frequently uses the MOCHA `fsew0` dataset as training and test data [3, 4]. In a recent work, Mitra et al. [2] suggest the use of the TAsk Dynamics Application (TADA [8]) model to generate acoustic-articulatory database which contains synthetic speech and the corresponding Vocal Tract (VT) time functions. Their results show that tract variables can be better candidates than EMA trajectories for articulatory feature based ASR systems. In this paper, we build upon the works of Mitra et al. [5, 2] by addressing the issue of finding the mapping between acoustic parameters and vocal tract variables. In this context, we use Mel-Frequency Cepstral Coefficients (MFCCs) as input and consider as output eight different vocal tract constriction variables, lip aperture (LA), lip protrusion (LP), tongue tip constriction degree (TTCD), tongue tip constriction location (TTCL), tongue body constriction degree (TBCD), tongue body constriction location (TBCL), Velum (VEL) and Glottis (GLO).

Various nonlinear acoustic-to-articulatory inversion technique [3, 2], and particularly kernel-based methods [9, 5], have been proposed in the literature, in most cases addressing the articulatory estimation problem within a single-task learning perspective. In the speech processing community, the first work we are aware of concerning this problem from a multi-task learning point of view is that of Richmond [10]. The author put forward the idea that we can benefit by viewing a multilayer perception (MLP) based acoustic-articulatory inversion from a multi-task learning perspective. In multi-task learning several related tasks are considered simultaneously [11]. The core idea is that what the machine learns for one task can help improve learning performance of the other tasks. There is a large literature on this subject [12, 13]. One paper that has come to our attention is that of Evgeniou et al. [14] who showed how Hilbert spaces of vector-valued functions [15] and matrix-valued reproducing kernels [16] can be used as a theoretical framework to develop nonlinear multi-task learning methods.

Motivated by the setting of reproducing kernel Hilbert spaces (RKHS) and their extensions considered in multi-task learning, we propose in this paper a matrix-valued kernel based approach for learning vocal tract variables from the the Mel-Frequency Cepstral Coefficients. This is achieved by using a vector-valued RKHS framework to learn a vector-valued function where each of its components corresponds to

a different tract variable.

The remainder of this paper is organized as follows. In Section 2, we review the definition of vector-valued RKHS and matrix-valued kernels. In section 3, we present the vocal tract variables estimation procedure. Experiments and results are presented in section 4. Section 5 presents some conclusions and future work direction.

2. VECTOR-VALUED REPRODUCING KERNEL HILBERT SPACES

Hilbert spaces of scalar functions with reproducing kernels were introduced and studied in [17]. Due to their crucial role in designing kernel-based learning methods successfully applied in several machine learning applications, these spaces have received considerable attention over the last two decades [18]. More recently, interest has grown in exploring Hilbert spaces of vector random functions for learning vector-valued functions [15]. In [14] Hilbert spaces of vector-valued functions [15] are described and matrix-valued reproducing kernels [16] are constructed to learn many related regression or classification tasks simultaneously. Similarly, there are also some works where approaches based on this framework have been proposed to estimate a two or three-dimensional vector-valued function [19].

In this section, we review the theory of reproducing kernel Hilbert spaces of vector-valued functions as in [15] and we discuss the choice of multi-task kernels. Let $X \subseteq \mathbb{R}^d$ and consider functions having values in some euclidean space $Y \subseteq \mathbb{R}^p$. We denote by $\mathcal{L}(\mathbb{R}^p)$ the space of $p \times p$ dimensional matrices and by $\mathcal{F}$ the linear space of function on X with values on Y.

Definition 1 *A function $K : X \times X \to \mathcal{L}(\mathbb{R}^p)$ is a multi-task (matrix-valued) kernel on X if, for any $x, t \in X$, $K(x,t)^T = K(t,x)$, and it is positive semi-definite, i.e., for every natural number n and all $\{(x_i, y_i)_{i=1,\dots,n}\} \in \mathbb{R}^d \times \mathbb{R}^p$ there holds*

$$\sum_{i,j=1}^{n} \langle K(x_i, x_j)u_i, u_j \rangle \geq 0$$

Definition 2 *A Hilbert space of function from X to Y is called a reproducing kernel Hilbert space if there is a positive semi-definite matrix-valued kernel K on X such that for every $f \in \mathcal{F}$,*

$$\langle f, K(x,.)y \rangle_{\mathcal{F}} = \langle f(x), y \rangle \quad (reproducing\ property)$$

Theorem 1 *A matrix-valued kernel K is the reproducing kernel of some vector-valued reproducing kernel Hilbert space, if and only if it is positive definite.*

Similarly to the scalar case, there is a bijection between matrix-valued kernel and vector valued RKHS. Therefore, it is important to consider the problem of constructing positive matrix-valued kernels. This problem has been studied extensively for scalar-valued kernels [18], however it has not been investigated enough in the matrix-valued case. In the context of multi-task learning, matrix-valued kernels are constructed from real kernels which are carried over to the vector-valued setting by a positive definite matrix [16]. This results in a multi-task kernel which has the following form

$$K(x,t) = \sum_{i=1}^{n} k_i(x,t)M_i, \quad \forall x, t \in X$$

where $k_i : X \times X \to \mathbb{R}$ is a scalar kernel and $M_i \in \mathcal{L}_+(\mathbb{R}^p)$. In this paper, we consider kernels of the this form with $n = 2$, M_1 a multiple of the identity matrix and M_2 a low rank matrix [20]. More specifically, we consider the following kernel

$$(K(x,t))_{i,j} = \alpha\langle x, t \rangle + (1-\alpha)\delta_{ij}\langle x, t \rangle^2 \qquad (1)$$

δ_{ij} is the Kronecker delta where $i, j \in \mathbb{N}_p = \{1, \dots, p\}$ and $\alpha \in [0, 1]$. Our choice of multi-task kernels is mainly motivated by the fact that vocal tract time functions are known to be functionally dependent [5].

3. VOCAL TRACT VARIABLES ESTIMATION

In this section, we detail the multi-task kernel based method used to estimate vocal tract variables from the Mel-Frequency Cepstral Coefficients (MFCCs). The problem of learning VT variables from MFCCs can be seen as a vector-valued function estimation problem where the goal is to find the function $f : X \longrightarrow Y$ from input vectors $(x_i)_{i=1}^n \in X \subseteq \mathbb{R}^d$ and associated vector-valued output values $(y_i)_{i=1}^n \in Y \subseteq \mathbb{R}^p$. x_i is a vector that represent the acoustic speech signal, y_i is the vector containing vocal tract variables values and f is the function that performs the mapping between the acoustic and articulatory domains. A good estimate of f is obtained by minimizing the regularized empirical risk $J_\lambda(f)$

$$J_\lambda(f) = \sum_{i=1}^{n} \|y_i - f(x_i)\|^2 + \lambda\|f\|_{\mathcal{F}}^2 \qquad (2)$$

where $\lambda \in \mathbb{R}^+$ is the regularization parameter. In analogy with standard kernel methods [18], a solution of the minimization problem can be provided using the vector-valued version of the representer theorem.

Theorem 2 *Let $\mathcal{F}$ a vector-valued reproducing kernel Hilbert space. The solution of the optimization problem based in minimizing the functional $J_\lambda(f)$ defined by equation (2) has the following form*

$$f(.) = \sum_{i=1}^{n} K(x_i, .)c_i \qquad (3)$$

where K is the reproducing kernel of $\mathcal{F}$ and $c_i \in Y$.

Using the representer theorem and the reproducing property of $\mathcal{F}$, the minimization problem of J_λ is equivalent to finding $C = (c_i)_{i=1}^n \in Y^n$ that minimize the following expression

$$\sum_{i=1}^{n} \|y_i - \sum_{j=1}^{n} K(x_i, x_j)c_j\|^2 + \lambda \sum_{i,j}^{n} \langle K(x_i, x_j)c_i, c_j \rangle \quad (4)$$

Now we can solve the minimization problem by computing the derivative of (4). Using the directional derivative defined by $D_h J_\lambda(C) = \lim_{\tau \longrightarrow 0} \tau^{-1}[J_\lambda(C + \tau h) - J_\lambda(C)]$ and the fact that $D_h J_\lambda(C) = \langle \nabla J_\lambda(C), h \rangle$, we obtain the following solution of the minimization problem

$$C = 2\ (2\ G^T G + \lambda(G^T + G))^{-1} G^T Y \quad (5)$$

where $G = (K(x_i, x_j))_{i,j=1}^n \in \mathbb{R}^{np \times np}$ is the multi-task kernel matrix and $Y = (y_i)_{i=1}^n \in Y^n$. Since K is a reproducing kernel and then $K(x,t) = K(t,x)^T$, the expression (5) is reduced to a more familiar form

$$C = (G + \lambda I)^{-1} Y \quad (6)$$

Following the theory of vector-valued RKHS presented in section 2 and the vector-valued function estimation procedure described above, we propose a multi-task based algorithm (Algorithm 1) for learning vocal tract variables from the acoustic speech signal.

Algorithm 1: acoustic-to-articulatory algorithm

1. Data preparation.
   - Form the training set $X_i = \{x_{i1}, \ldots, x_{id}\}, i = 1, \ldots, n$ containing the MFCCs of the speech signal and the corresponding vocal tract variables $Y_i = \{Y_{i1}, \ldots, Y_{ip}\}, i = 1, \ldots, n$.
   - Set the parameter α of the multi-task kernel (1) to some predetermined value in $[0, 1]$.
2. Training step
   - Solve the minimization problem (4) to get the set of vectors $c_i \in \mathbb{R}^p\ \forall i = 1, \ldots, n$, see equation (6).
3. Test step
   - For each speech signal
     - compute its feature vector denoted as x;
     - compute the corresponding tract variables using equation (3).

---

4. EXPERIMENTS

In this section, we report on a set of experiments similar to those performed by Mitra et al. [5, 2]. Acoustic-articulatory database is generated by the TAsk Dynamics Application (TADA [8]) model, which is a computational implementation of articulatory phonology. The generated dataset consists of acoustic signals for 416 words chosen from the Wisconsin X-ray microbeam data [21] and corresponding Vocal Tract (VT) trajectories sampled at 5 msec. The speech signal was parameterized into 13 Mel-Frequency Cepstral Coefficients (MFCC). These Cepstral coefcients were acquired at a time of 5 ms (synchronized with the TVs) with window duration of 10 ms. As in [9], input vectors are constructed by considering 8 frames before and after each current frame. The dataset is split into 5:1 for training and test sets.

Table 1. *Average RMSE and Corr for the tract variables using hierarchical ε-SVR [5] and our method based on multi-task kernel.*

VT variables	ε-SVR		Multi-task	
	RMSE	Corr	RMSE	Corr
LA	2.763	0.715	**2.341**	**0.832**
LP	0.532	0.722	**0.512**	**0.794**
TTCD	3.345	0.829	**1.975**	**0.935**
TTCL	7.752	0.843	**5.276**	**0.902**
TBCD	2.155	0.866	**2.094**	**0.875**
TBCL	15.083	0.867	**9.763**	**0.924**
VEL	**0.032**	0.937	0.034	**0.944**
GLO	**0.041**	**0.954**	0.052	0.951

For evaluating the performance of the VT variables estimation task, the root mean-squared error (RMSE) and the correlation measure (Corr) are widely used. The RMSE measure the distance between predicted and true VT trajectories and the correlation score provide information about the shape similarity and the synchrony of the two trajectories. These measures are defined as follows

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(e_i - y_i)^2}$$

$$Corr = \frac{\sum_i (e_i - \bar{e})(y_i - \bar{y})}{\sqrt{\sum_i (e_i - \bar{e})^2 \sum_i (y_i - \bar{y})^2}}$$

where e is the estimated vector of the true N-dimensional VT vector y. $\bar{e}$ and $\bar{e}$ are the mean values of e and y.

A recent tract variables learning technique is proposed by Mitra et al. [5]. The authors developed a hierarchical ε-SVR architecture by associating 8 different SVRs, a SVR for each tract variable. To consider the dependencies between VT time functions, the SVRs corresponding to independent VT variables are first created and then used for constructing the others. Table 1 reports the RMSE and correlation results obtained using the multi-task kernel based approach and the hiearchical ε-SVR algorithm after smoothing the estimated VT trajetories using a Kalman filter as described in [5]. It can be observed in the Table 1 that the multi-task kernel method offers better performance than the hierarchical ε-SVR technique. The improvement is significant for TBCL, TTCL and

TTCD, the RMSE values decrease respectively from 15.083, 7.752 and 3.345 to 9.763 5.276 and 1.975. Results obtained for the other tract variables are more or less similar to that of the ε-SVR.

5. CONCLUSION

In this work, we have tackled the problem of vocal tract variables learning from the acoustic speech signal. Using a vector-valued RKHS framework, we have proposed a multi-task kernel based method that has the advantage of taking into account the functional relationships between the tract variables. Experiments on acoustic-articulatory database generated by TADA are conducted to estimate VT time functions from the MFCCs. We have also compared our method with the hierarchical ε-SVR, yielding improved performance. In future it will be also interesting to consider a larger classes of matrix-valued kernels. We will also explore more experiments on articulatory datasets and compare the multi-task kernel approach with previous related methods for speech inversion, such as those in [2].

6. REFERENCES

[1] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Transactions on Acoustics and Speech Signal Processing*, vol. 2, no. 1, Part II, pp. 133–150, 1994.

[2] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, and L. Goldstein, "Retrieving tract variables from acoustics: A comparison of different machine learning strategies," *IEEE Journal of Selected Topics in Signal Processing*, to appear, 2010.

[3] K. Richmond, *Estimating Articulatory Parameters from the Acoustic Speech Signal*, Ph.D. thesis, The Center for Speech Technology Research, Edinburgh, 2002.

[4] A. Toutios and K. Margaritis, "Learning articulation from cepstral coefficients," in *International Speech and Computer Conference (SPECOM'05)*, Patras, Greece, 2005.

[5] V. Mitra, Y. Ozbek, H. Nam, X. Zhou, and C. Y. Espy-Wilson, "From acoustics to vocal tract time functions," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP '09)*, Washington, DC, USA, 2009, pp. 4497–4500.

[6] K. Kirchoff, *Robust Speech Recognition Using Articulatory Information*, Ph.D. thesis, University of Bielefeld, 1999.

[7] T. Toda, A. W. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis," in *ISCA Speech Synthesis Workshop*, 2004.

[8] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, portable task dynamics model in matlab," *Acoustical Society of America Journal*, vol. 115, pp. 2430, 2004.

[9] A. Toutios and K. Margaritis, "A support vector approach to the acoustic-to-articulatory mapping," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 3221–3224.

[10] K. Richmond, "A multitask learning perspective on acoustic-articulatory inversion," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007.

[11] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[12] S. Ben-david and R. Schuller-Borbely, "A notion of task relatedness yielding provable multiple-task learning guarantees," *Machine Learning*, vol. 73, pp. 273–287, 2008.

[13] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[14] T. Evgeniou, C. A. Micchelli, and M. Pontil, "Learning multiple tasks with kernel methods," *Journal of Machine Learning Research*, vol. 6, pp. 615–637, 2005.

[15] C. A. Micchelli and M. Pontil, "On learning vector-valued functions," *Neural Computation*, vol. 17, pp. 177–204, 2005.

[16] C. A. Micchelli and M. Pontil, "Kernels for multi-task learning," in *Advances in Neural Information Processing Systems 17 (NIPS 2005)*, 2005, pp. 921–928.

[17] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337–404, 1950.

[18] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, USA, 2002.

[19] M. Ha Quang, S. H. Kang, and T. M. Le, "Image and video colorization using vector-valued reproducing kernel Hilbert spaces," *Journal of Mathematical Imaging and Vision*, vol. 37, no. 1, pp. 49–65, 2010.

[20] A. Caponnetto, C. A. Micchelli, M. Pontil, and Y. Ying, "Universal multi-task kernels," *Journal of Machine Learning Research*, vol. 68, pp. 1615–1646, 2008.

[21] J.R. Westbury, G. Turner, and J. Dembovski, "X-ray microbeam speech production database users' handbook," vol. Waisman Center, University of Wisconsin, 1994.