ADAPTIVE KERNEL CANONICAL CORRELATION ANALYSIS FOR ESTIMATION OF TASK DYNAMICS FROM ACOUSTICS

Frank Rudzicz

University of Toronto, Department of Computer Science

frank@cs.toronto.edu

ABSTRACT

We present a method for acoustic-articulatory inversion whose targets are the abstract tract variables from task dynamic theory. Towards this end we construct a non-linear Hammerstein system whose parameters are updated with adaptive kernel canonical correlation analysis. This approach is notably semi-analytical and applicable to large sets of data. Training behaviour is compared across four kernel functions and prediction of tract variables is shown to be significantly more accurate than state-of-the-art mixture density networks.

Index terms: acoustic-articulatory inversion, kernel canonical correlation analysis, task dynamics.

1. INTRODUCTION

Differences between speakers are the result of purely endogenous phenomena distinguished by their mechanics of articulation. Such distinctions cannot be codified into automatic speech recognition (ASR) systems that are agnostic of speech production, however. It is therefore desirable to find an accurate method of projecting acoustic speech data onto a lower-dimensional space which is more indicative of the linguistic intentions of the speaker, namely, to the space of physical properties of vocal tract motion.

Although acoustic-to-articulatory inversion is a one-to-many relationship [1], such protestation has not limited research in this area. For example, Richmond et al. [2] estimated the 2-dimensional midsagittal positions of 7 articulators given kinematic data using both a multi-layer perceptron and discriminatively trained Gaussian mixture models to within 0.41mm and 2.73mm. Toda et al. [3] achieved almost identical results on the same data by applying expectation-maximization using both minimum mean-squared error and maximum likelihood estimation to a Gaussian mixture mapping function with low-pass filtering. Simpler approaches achieved similar results (errors less than 2mm, typically around 1mm) using simple vector quantization with an appropriate number of vectors [4].

One commonality in existing work is that the target dimensions consist of the absolute physical positions of points in the vocal tract. Typical points include the upper and lower lips (**UL**, **LL**), the upper and lower incisors (**UI**, **LI**), the tongue tip, body, and dorsum (**TT**, **TB**, **TD**), and the velum (**V**). Despite the popularity of this approach, neither its generalizability among speakers nor its representation of linguistic intent has been justified. Why would the physical position of the upper lip be as explicative of intent or of acoustic consequence as a measure of the distance between the lips, for example?

In this paper we estimate features of the vocal tract from acoustics using adaptive kernel canonical correlation analysis (KCCA). We choose features of the vocal tract derived from the theory of task dynamics, as described below.

1.1. Tract variables and task dynamics

Task dynamics constitute a combined model of skilled articulator motion and the planning of vocal tract configurations [5]. This theory introduces the notion that the dynamic patterns that occur in speech are the result of overlapping *gestures*, which are high-level abstractions of goal-oriented reconfigurations of the vocal tract. An instance of a gesture in this theory is any articulatory movement towards the completion of a particular speech-relevant goal, such as bilabial closure, or velar opening. The progenitors of this theory claim that all the implicit spatiotemporal behaviour underlying speech is the result of the interaction between the abstract *intergestural* dimension (between tasks) and the geometric *interarticulator* dimension (between physical actuators) [5].

Each gesture in this theory occurs within one of the following *tract variables* (TVs): lip aperture and protrusion (**LA**, **LP**), tongue tip constriction location and degree (**TTCL**, **TTCD**), tongue body constriction location and degree (**TBCL**, **TBCD**), velar opening (**VEL**), glottal vibration (**GLO**), and lower tooth height (**LTH**). A gesture to close the lips, for example, would occur within the LA variable, and would set that variable close to zero.

Articulatory data consists of spoken utterances and their aligned articulator positions as described in section 3. In order to convert the articulator space to tract variable space, we transform the midsagittal articulatory data using a combination of principal component analysis and sigmoid activation functions. For example, we describe VEL by calculating the first principal component of velum motion in the midsagittal plane, finding the minimum and maximum deviations from the mean in this transformed space, and applying a sigmoid to that unidimensional space to retrieve a real function on $[0..1]$. Similarly, the first and second principal components of the distance between UL and LL are used for the determination of lip aperture and protrusion, respectively, the first and second principal components of TT are used for the determination of TTCL and TTCD, respectively, and the first and second principal components of TB are used for the determination of TBCL and TBCD, respectively. Voicing detection on energy below 150Hz is used to estimate the GLO tract variable.

2. ADAPTIVE KCCA

Canonical correlation analysis (CCA) is a popular technique in statistical analysis used in a variety of contexts, including communication theory and statistical signal processing, to measure linear relationships between sets of variables. Given vector variables $\mathbf{x} \in \mathbb{R}^{\mathbf{m_x}}$ and $\mathbf{y} \in \mathbb{R}^{\mathbf{m_y}}$, CCA finds a pair of directions $\omega_x \in \mathbb{R}^{m_x}$ and $\omega_y \in \mathbb{R}^{m_y}$ such that the correlation $\rho(\mathbf{x}, \mathbf{y})$ is maximized between the two projections $\omega_x^T \mathbf{x}$ and $\omega_y^T \mathbf{y}$. Given joint observations $\mathbf{X} = [\mathbf{x}_1 \mathbf{x}_2 ... \mathbf{x}_N]^T$ and $\mathbf{Y} = [\mathbf{y}_1 \mathbf{y}_2 ... \mathbf{y}_N]^T$, where $\mathbf{x}_i$ co-occurs with $\mathbf{y}_i$, CCA is equivalent to finding projection vectors ω_x and ω_y that maximize

$$\rho(\mathbf{X},\mathbf{Y};\omega_x,\omega_y) = \frac{\omega_x^T \mathbf{X}\mathbf{Y}^T \omega_y}{\sqrt{\omega_x^T \mathbf{X}\mathbf{X}^T \omega_x}\sqrt{\omega_y^T \mathbf{Y}\mathbf{Y}^T \omega_y}}. \qquad (1)$$

Although this method can find good linear relationships between sets of data, it is incapable of capturing nonlinear relationships, which limits its application in many aspects of speech. In order to overcome this limitation, we employ the “kernel trick” in which a nonlinear transformation Φ of the data obtains a higher-dimensional feature space (e.g., $\hat{\mathbf{X}} = \Phi(\mathbf{X})$). The linear solution of CCA within this higher-dimensional space is equivalent to a non-linear solution in the original data space [6]. We can avoid the need to explicitly define Φ, however, since positive definite kernel functions $\kappa(\mathbf{x},\mathbf{y})$ satisfying Mercer’s condition can implicitly map their input to higher-dimensional spaces. We specify a set of such kernels in section 3.

Reformulating eq. 1 within a framework of least-squares regression allows us to minimize $\frac{1}{2}\|\mathbf{X}\omega_x - \mathbf{Y}\omega_y\|^2$ such that $\frac{1}{2}\left(\|\mathbf{X}\omega_x\| + \|\mathbf{Y}\omega_y\|\right) = 1$. This allows us to solve the following generalized eigenvalue problem on the transformed data $\hat{\mathbf{X}} \in \mathbb{R}^{N\times m'_x}$ and $\hat{\mathbf{Y}} \in \mathbb{R}^{N\times m'_y}$ by the method of Lagrange multipliers:

$$\frac{1}{2}\begin{bmatrix} \hat{\mathbf{X}}^T\hat{\mathbf{X}} & \hat{\mathbf{X}}^T\hat{\mathbf{Y}} \\ \hat{\mathbf{Y}}^T\hat{\mathbf{X}} & \hat{\mathbf{Y}}^T\hat{\mathbf{Y}} \end{bmatrix} \hat{\omega} = \beta \begin{bmatrix} \hat{\mathbf{X}}^T\hat{\mathbf{X}} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{Y}}^T\hat{\mathbf{Y}} \end{bmatrix} \hat{\omega}, \qquad (2)$$

where $\hat{\omega} = [\hat{\omega}_x \hat{\omega}_y]^T$ is the concatenation of the transformed direction vectors. We can now avoid explicit data transformation by applying a kernel function. Since the kernel matrix describing our transformed data, $\mathbf{K}_x = \hat{\mathbf{X}}_k\hat{\mathbf{X}}_k^T \in \mathbb{R}^{N\times N}$, has elements $\mathbf{K}_x[i,j] = \kappa(\mathbf{x}_i,\mathbf{x}_j)$ defined by vectors in our original data space ($\mathbf{K}_y$ is defined similarly for $\hat{\mathbf{Y}}$), we left-multiply eq. 2 by $\begin{bmatrix} \hat{\mathbf{X}} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{Y}} \end{bmatrix}$, giving

$$\frac{1}{2}\begin{bmatrix} \mathbf{K}_x^2 & \mathbf{K}_x\mathbf{K}_y \\ \mathbf{K}_y\mathbf{K}_x & \mathbf{K}_y^2 \end{bmatrix} \alpha = \beta \begin{bmatrix} \mathbf{K}_x^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_y^2 \end{bmatrix} \alpha. \qquad (3)$$

Here, $\alpha = [\alpha_x \alpha_y]^T \in \mathbb{R}^{2N}$ such that $\hat{\omega}_x = \hat{\mathbf{X}}^T\alpha_x$ and $\hat{\omega}_y = \hat{\mathbf{Y}}^T\alpha_y$ [7]. This gives a generalized eigenvalue problem in the higher-dimensional space where we can minimize $\left(\mathbf{K}_x\alpha_x + \mathbf{K}_y\alpha_y\right)/2$ by adjusting α_x and α_y according to our original data space [8].

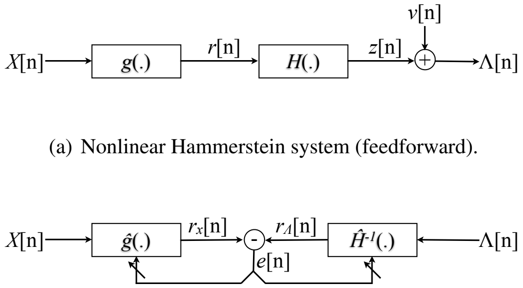
2.1. KCCA and Hammerstein systems

A nonlinear Hammerstein system is a memoryless nonlinear function $g()$ followed by a linear dynamic system $H()$ in series, as shown in Figure 1(a). Our goal is to input acoustic observations, $\mathbf{X}$, of Mel-frequency cepstral coefficients (MFCC) to such a system and to infer the associated articulation vectors, Λ. In order to accomplish this accurately, we must learn the parameters of the two components of the Hammerstein system.

A mechanism for identifying these parameters has recently been proposed that takes advantage of the cascade structure by inverting the linear component, as in Figure 1(b), and minimizing the difference, $e[n]$, between $g(\mathbf{X}[n])$ and $H^{-1}(\Lambda[n])$ using KCCA [9]. Since $H()$ is linear, we can reformulate eq. 3 to

$$\frac{1}{2}\begin{bmatrix} \mathbf{K}_x^2 & \mathbf{K}_x\hat{\Lambda} \\ \hat{\Lambda}^T\mathbf{K_x} & \hat{\Lambda}^T\hat{\Lambda} \end{bmatrix} \begin{bmatrix} \alpha_x \\ \omega_\Lambda \end{bmatrix} = \beta \begin{bmatrix} \mathbf{K}_x(\mathbf{K}_x + c\mathbf{I}) & \mathbf{0} \\ \mathbf{0} & \hat{\Lambda}^T\hat{\Lambda} \end{bmatrix} \begin{bmatrix} \alpha_x \\ \omega_\Lambda \end{bmatrix}, \qquad (4)$$

where we add a regularizing constant c to prevent overfitting [9]. Here, ω_Λ provides the parameters of the linear part of the system, $H()^{-1}$, and α_x provides the parameters of the nonlinear part, $g()$. Given a combined average of the output of these two systems, $r = (r_x + r_\Lambda)/2 = (\mathbf{K}_x\alpha_x + \Lambda\omega_\Lambda)/2$, the eigenvalue problem decomposes to two coupled least squares problems:



(a) Nonlinear Hammerstein system (feedforward).

(b) System for identifying the parameters of the nonlinear Hammerstein system.

Fig. 1. The feedforward Hammerstein system and its associated identification system.

$$\begin{aligned} \beta\alpha_x &= (\mathbf{K}_x + c\mathbf{I})^{-1} r \\ \beta\omega_\Lambda &= (\Lambda^T\Lambda)^{-1}\Lambda^T r \end{aligned} \qquad (5)$$

This representation allows us to minimize a Euclidean error measurement $\|r_x - r_\Lambda\|$ by analytically solving for α_x and ω_Λ. In order to estimate articulation at run time, we compute $r_x = \mathbf{K}_x\alpha_x$, since we can construct the kernel matrix from observed acoustics, and then solve for $\Lambda \approx \mathbf{K}_x\alpha_x\omega_\Lambda^{-1}$, since $\Lambda\omega_\Lambda = r_\Lambda \approx r_x = \mathbf{K}_x\alpha_x$.

2.2. Adaptive algorithm

Unfortunately, for problems involving large amounts of data, as is typical in speech, the sizes of the kernel matrices described above become prohibitively large. An online algorithm that iteratively adjusts the estimates of α_x and ω_Λ based on subsequent segments of data is therefore desirable. We assume that we have a sliding context window covering L aligned frames from each data source, namely, $\mathbf{x}^{(n)} = [\mathbf{x}_n, \mathbf{x}_{n-1}, ..., \mathbf{x}_{n-L+1}]$ and $\Lambda^{(n)} = [\Lambda_n, \Lambda_{n-1}, ..., \Lambda_{n-L+1}]$. Assuming that we have matrix $\mathbf{K}_{reg}^{(n-1)}$ for the $(n-1)^{th}$ window of speech, and $\hat{\mathbf{K}}_{reg}^{(n-1)}$ is the matrix formed by its last $n-1$ rows and columns, then the regularized matrix for the current window is

$$\mathbf{K}_{reg}^{(n)} = \begin{bmatrix} \hat{\mathbf{K}}_{reg}^{(n-1)} & \mathbf{k}_{n-1}(\mathbf{x}^{(n)}) \\ \mathbf{k}_{n-1}(\mathbf{x}^{(n)})^T & k_{nn} + c \end{bmatrix}, \qquad (6)$$

where $\mathbf{k}_{n-1}(\mathbf{x}^{(n)}) = [\kappa(\mathbf{x}^{(n-L+1)}, \mathbf{x}^{(n)}, ..., \kappa(\mathbf{x}^{(n-1)}, \mathbf{x}^{(n)}]^T$ and $k_{nn} = \kappa(\mathbf{x}^{(n)}, \mathbf{x}^{(n)})$. The inverse of $\mathbf{K}_{reg}^{(n)}$ can also be computed quickly, given the inverse of $\mathbf{K}_{reg}^{(n-1)}$ [10]. We then iteratively update our parameter estimates for ω_Λ and α_x as new data arrives using eq. 5. This entire process is summarized in algorithm 1 and is based on work on Wiener systems by Vaerenbergh et al. [7].

3. EXPERIMENTS

Our experiments evaluate the stability of the error-correction method and the estimation of tract variables from acoustics. We apply four kernel functions, namely the homogenous polynomial ($K_{h\_poly}^{(i)}$), the

```
begin
    Initialize $\mathbf{K}_{reg}^{(0)} = (1+c)\mathbf{I}$
    Initialize $\alpha_x$ and $\omega_\Lambda$ with random data
    for $n = 1..N$ do
        Calculate $\mathbf{K}_{reg}^{(n)}$ from $x^{(n)}$ as in eq. 6
        $\mathbf{r}_x^{(n)} = \kappa(\mathbf{x}^{(n)}, \mathbf{x}^{(n-1)})\alpha_x^{(n-1)}$
        $\mathbf{r}_\Lambda^{(n)} = \Lambda^{(n)}\omega_\Lambda^{(n-1)}$
        $\mathbf{r}^{(n)} = \frac{\mathbf{r}_x^{(n-1)} + \mathbf{r}_\Lambda^{(n-1)}}{2}$
        Calculate $\left(\mathbf{K}_{reg}^{(n)}\right)^{-1}$
        Update solutions for $\alpha_x$ and $\omega_\Lambda$ as in eq. 5
        Normalize solutions with $\beta = \|\omega_\Lambda\|$
    end
end
```

Algorithm 1: The adaptive KCCA algorithm.

non-homogenous polynomial ($K_{nh\_poly}^{(i)}$), the radial-basis function ($K_{rbf}^{(\sigma)}$), and the sigmoid ($K_{sigmoid}^{(\kappa,c)}$) kernels:

$$K_{h\_poly}^{(i)}(x_1,x_2) = (x_1 x_2)^i$$

$$K_{nh\_poly}^{(i)}(x_1,x_2) = (x_1 x_2 + 1)^i$$

$$K_{rbf}^{(\sigma)}(x_1,x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right)$$

$$K_{sigmoid}^{(\kappa,c)}(x_1,x_2) = \tanh(x_1 x_2 + c)\,.$$

Training data consists of midsagittal tract variables (10-dimensional vectors) and aligned acoustics (42-dimensional MFCCs) selected from approximately 460 sentences uttered by a male speaker from Edinburgh's MOCHA database [11]. The positions and velocities of the jaw, lips, and tongue, as exemplified in Figure 2, are recorded with electromagnetic articulography using alternating electromagnetic fields generated by a cube that surrounds the speaker's head. These data are then converted to the tract variable space as described in section 1.1. Results reported below are averages of 10-fold cross validation. Until otherwise indicated, the window length $L = 150$.

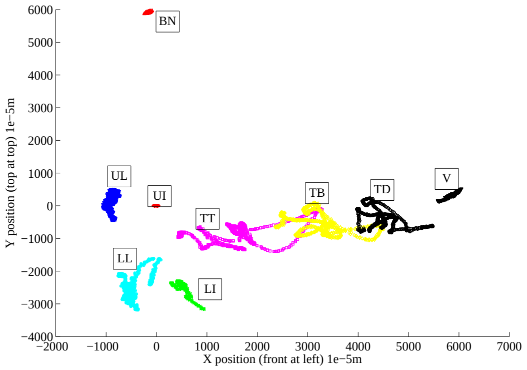


Fig. 2. The midsagittal motion of the articulators during the phrase "*This was easy for us*".

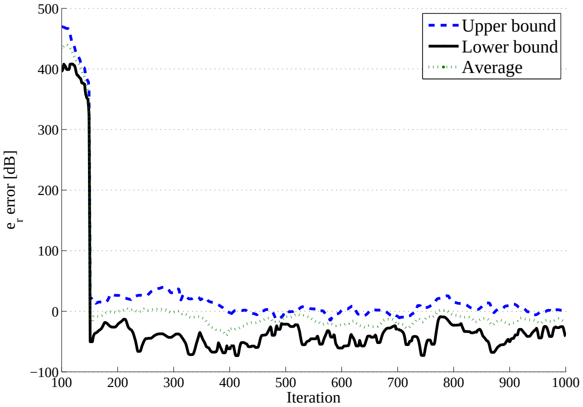


Fig. 3. Normalization error, $e[n]$, for the first-order homogenous polynomial kernel at window size $L = 150$.

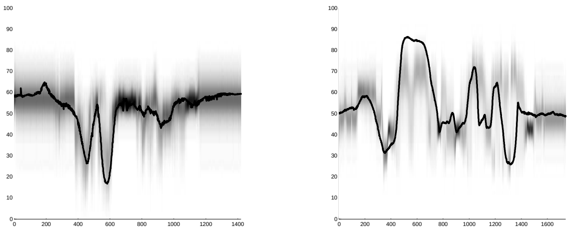
3.1. Stability and convergence during training

The goal of auto-correction is for the Euclidean error $(\mathbf{K}_x\alpha_x - \Lambda\omega_\Lambda)$ (i.e., $e[n]$ in Figure 1(b)) to approach zero during training. Figure 3 shows the best, average, and worst mean squared errors in decibels during training given the homogenous polynomial kernel and 10 random initial parameterizations. This example is indicative of all other kernels whereby a period of fluctuation tends to follow a rapid decrease in error. Table 1 shows the total decrease in mean squared error (dB) between the first 20 and last 20 windows of the adaptive KCCA training process. As one increases the order of both the homogenous and non-homogenous kernels, the MSE also increases. In both the tan-sigmoid and radial-basis function kernels, however, our choice of parameters seems to have little discernible effect.

Homogenous polynomial		Nonhomogenous polynomial	
i	MSE reduction	i	MSE reduction
1	421.6	1	441.9
2	403.6	2	413.1
3	394.5	3	382.9
Sigmoid		**Radial-basis function**	
$(\kappa$, c)	MSE reduction	σ	MSE reduction
$(0.2, 0.1)$	313.2	0.1	406.5
$(0.2, 0.5)$	321.5	0.5	410.4
$(0.5, 0.1)$	309.7	1.0	406.7
$(0.5, 0.5)$	314.3		

Table 1. Total reduction in MSE (dB) between Hammerstein components during training across kernels and parameterizations.

Vaerenbergh et al. apply a nearly identical approach to learning Wiener systems on the comparatively simple problem of estimating a hyperbolic tangeant function given univariate input [7; 8], reaching MSE between −30dB and −40dB within 1000 to 1500 iterations. Surprisingly, most of the error in our experiments is dispelled much earlier, within 200 iterations, with MSE fluctuating between −76.9dB and 39.5dB thereafter across all kernels and parameterizations. This suggests that adaptive KCCA converges rapidly during training on acoustic-articulatory data.



TTCD TTCL

Fig. 4. Example intensity maps of Gaussian mixtures produced by mixture density networks (MDNs) that estimate tongue tip constriction degree and location. Darker sections represent higher probability. The true trajectories are superimposed as black curves.

TV	**MDN** $\mu(\sigma^2)$	**KCCA** $\mu(\sigma^2)$	**TV**	**MDN** $\mu(\sigma^2)$	**KCCA** $\mu(\sigma^2)$
VEL	−0.28 (0.08)	−0.23 (0.07)	TTCD	−1.60 (0.17)	−1.60 (0.17)
LTH	−0.18 (0.12)	−0.18 (0.14)	TTCL	−1.62 (0.17)	−1.57 (0.16)
LA	−0.32 (0.11)	−0.28 (0.10)	TBCD	−0.79 (0.14)	−0.80 (0.15)
LP	−0.44 (0.12)	−0.41 (0.13)	TDCL	−0.20 (0.11)	−0.18 (0.09)
GLO	−1.30 (0.16)	−1.14 (0.15)			

Table 2. Average log likelihoods of true tract variable positions in test data, under distributions produced by mixture density networks (MDNs) and the KCCA method, with variances.

3.2. KCCA versus mixture density networks

In order to judge the accuracy of the articulatory estimates produced by adaptive KCCA against the state-of-the-art, we consider mixture density neural networks (MDNs) that output parameters of Gaussian mixture probability distributions, as described by Richmond [2]. We train MDNs to estimate the likelihood of tract variable positions given MFCC input and 2 frames of surrounding acoustic context. Figure 4 shows the estimated likelihood of tract variable positions over time produced by trained MDNs as intensity maps superimposed with the true trajectories. MDNs are trained on the same data as KCCA. Articulatory estimates for KCCA are smoothed with third-order median filters.

We assess the accuracy of the MDN and KCCA methods by comparing their estimates of the log likelihood of the true articulatory trajectories. A more accurate method will assign a higher probability to the actual trajectory. The likelihood of a frame of articulation is easily computed by MDNs whose output is a probability distribution over tract variable positions. We approximate the likelihood of a frame of articulation in the KCCA approach with the radial-basis kernel by fitting a Gaussian to the estimates of 10 trials having different initial parameterizations. Test data in each trial consists of approximately 60 utterances from our male speaker.

The mean and variance of the log likelihoods of true articulatory positions across all test frames is summarized in Table 2 for both methods. According to the t test with $9.6E^4 < n_1 = n_2 < 9.9E^4$ frames and one degree of freedom, KCCA is significantly more accurate than the MDN method at the 95% confidence level for **VEL**, **LA**, **LP**, **TTCL**, and **TDCL** and at the 99% confidence level for **GLO**, and statistically indistinguishable at these levels for the remaining tract variables.

4. CONCLUDING REMARKS

Some high-level questions remain. For example, if the eventual aim is to use estimated articulatory trajectories to constrain hypotheses in speech recognition, then it is possible that a quantized representation may be more amenable to training in such systems. A similar (though non-adaptive) kernel-based system has recently been proposed that inverts acoustic to articulatory data according to discrete categories [12]. Likewise, a k-means clustering of the tract variable motion estimated by our adaptive KCCA process might be applicable as conditioning variables in dynamic Bayes networks for speech classification [13].

Our analysis has demonstrated that adaptive KCCA can effectively learn non-linear relationships between co-occurring variables in speech, and perform more accurate acoustic-to-articulatory inversion than the state-of-the-art. This approach combines a semi-analytical (non-statistical) kernel-based approach with an iterative, adaptive learning process.

5. ACKNOWLEDGMENTS

This work is funded by the Natural Sciences and Engineering Research Council of Canada and the University of Toronto.

6. REFERENCES

[1] Sam T. Roweis, *Data Driven Production Models for Speech Processing*, Ph.D. thesis, California Institute of Technology, Pasadena, California, 1999.

[2] Korin Richmond, Simon King, and Paul Taylor, "Modelling the uncertainty in recovering articulation from acoustics," *Computer Speech and Language*, vol. 17, pp. 153–172, 2003.

[3] Tomoki Toda, Alan W. Black, and Keiichi Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, March 2008.

[4] John Hogden, Philip Rubin, Erik McDermott, Shigeru Katagiri, and Louis Goldstein, "Inverting mappings from smooth paths through r^n to paths through r^m: A technique applied to recovering articulation from acoustics," *Speech Communication*, vol. 49, no. 5, pp. 361–383, 2007.

[5] Elliot L. Saltzman and Kevin G. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological Psychology*, vol. 1, no. 4, pp. 333–382, 1989.

[6] Pei Ling Lai and Colin Fyfe, "Kernel and nonlinear canonical correlation analysis," *International Journal of Neural Systems*, vol. 10, no. 5, pp. 365–377, 2000.

[7] Steven Van Vaerenbergh, Javier Via, and Ignatio Santamaria, "Online kernel canonical correlation analysis for supervised equalization of Wiener systems," in *Proceedings of the 2006 International Joint Conference on Neural Networks*, Vancouver, Canada, July 2006, pp. 1198–1204.

[8] Steven Van Vaerenbergh, Javier Via, and Ignacio Santamaria, "Adaptive kernel canonical correlation analysis algorithms for nonparametric identification of Wiener and Hammerstein systems," *EURASIP Journal on Advances in Signal Processing*, vol. 8, no. 2, pp. 1–13, January 2008.

[9] Ernst Aschbacher and Markus Rupp, "Robustness analysis of a gradient identification method for a nonlinear Wiener system," in *Proceedings of the 13th Statistical Signal Processing Workshop (SSP)*, Bordeaux, France, July 2005.

[10] Steven Van Vaerenbergh, Javier Via, and Ignatio Santamaria, "A sliding-window kernel RLS algorithm and its application to nonlinear channel identification," in *Proceedings of the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, May 2006.

[11] Alan Wrench, "The MOCHA-TIMIT articulatory database," November 1999.

[12] Wenming Zheng, Xiaoyan Zhou, Cairong Zou, and Li Zhao, "Facial expression recognition using kernel canonical correlation analysis (KCCA)," *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 233–238, 2006.

[13] Frank Rudzicz, "Applying discretized articulatory knowledge to dysarthric speech," in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP09)*, Taipei, Taiwan, April 2009.