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 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 velum motion in the midsagittal plane, and the first and second principal components of TB are used for the determination of velocity and velar closure, 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 \( x \in \mathbb{R}^m \) and \( y \in \mathbb{R}^n \), CCA finds a pair of directions \( \omega_x \in \mathbb{R}^m \) and \( \omega_y \in \mathbb{R}^n \) such that the correlation \( \rho(x, y) \) is maximized between the two projections \( \omega^T_x x \) and \( \omega^T_y y \). Given joint observations \( X = [x_1 x_2 \ldots x_N]^T \) and \( Y = [y_1 y_2 \ldots y_N]^T \), where \( x_i \) co-occurs with \( y_i \), CCA is equivalent...
to finding projection vectors $\omega_i$ and $\omega_j$ that maximize

$$
\rho(X, Y; \omega_i, \omega_j) = \frac{\omega_i^T X Y^T \omega_j}{\sqrt{\omega_i^T X X^T \omega_i} \sqrt{\omega_j^T Y Y^T \omega_j}}.
$$

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 $\Phi$ of the data obtains a higher-dimensional feature space (e.g., $\tilde{X} = \Phi(X)$). 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 $\Phi$, however, since positive definite kernel functions $k(x, y)$ 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 $\frac{1}{2} \|X\omega_i - Y\omega_j\|^2$ such that $\frac{1}{2} \parallel X\omega_i \parallel + \parallel Y\omega_j \parallel = 1$. This allows us to solve the following generalized eigenvalue problem on the transformed data $\tilde{X} \in \mathbb{R}^{N \times m'}$ and $\tilde{Y} \in \mathbb{R}^{N \times m}$ by the method of Lagrange multipliers:

$$
\frac{1}{2} \begin{bmatrix} X^T \tilde{X} & X^T \tilde{Y} \\ \tilde{Y}^T \tilde{X} & \tilde{Y}^T \tilde{Y} \end{bmatrix} \omega = \beta \begin{bmatrix} X^T 0 \\ 0 Y^T \end{bmatrix} \omega.
$$

where $\omega = [\hat{\omega}_x, \hat{\omega}_y]^T$ 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, $K_x = \tilde{X}_x \tilde{X}_x^T \in \mathbb{R}^{N \times N}$, has elements $K_{x_i}[i, j] = k(x_i, x_j)$ defined by vectors in our original data space ($K_x$ is defined similarly for $\tilde{Y}$), we left-multiply eq. 2 by $[X \ 0 \ Y \ 0]$ giving

$$
\frac{1}{2} \begin{bmatrix} K_x^2 + K_y^2 & K_x \ K_y \\ K_y \ K_x & K_y^2 + K_x^2 \end{bmatrix} \alpha = \beta \begin{bmatrix} K_y^2 + K_x^2 \\ K_y \ K_x \ K_x \ K_y \end{bmatrix} \alpha.
$$

where $\alpha = [\alpha_x, \alpha_y]^T \in \mathbb{R}^{2N}$ such that $\hat{\omega}_x = X^T \alpha_x$ and $\hat{\omega}_y = Y^T \alpha_y$ [7]. This gives a generalized eigenvalue problem in the higher-dimensional space where we can minimize $(K_x \alpha_x + K_y \alpha_y)$ / 2 by adjusting $\alpha_x$ and $\alpha_y$ 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 vectors, $X$, of Mel-frequency cepstral coefficients (MFCC) to such a system and to infer the associated articulation vectors, $A$. 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 $g(X[n])$ and $H^{-1}(A[n])$ using KCCA [9]. Since $H()$ is linear, we can reformulate eq. 3 to

$$
\frac{1}{2} \begin{bmatrix} K_x^2 & K_x \Lambda \Lambda^T K_x \\ \Lambda^T K_x \Lambda & \Lambda^T \Lambda \end{bmatrix} \alpha = \beta \begin{bmatrix} K_x \Lambda + \alpha \Lambda \\ \Lambda \alpha \end{bmatrix} \alpha,
$$

where we add a regularizing constant $c$ to prevent overfitting [9]. Here, $\omega_x$ provides the parameters of the linear part of the system, $H^{-1}()$, and $\alpha_y$ provides the parameters of the nonlinear part.

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

$$
g(). \text{ Given a combined average of the output of these two systems, } r = (r_x + r_y) / 2 = (K_x \alpha_x + K_y \alpha_y) / 2, \text{ the eigenvalue problem decomposes to two coupled least squares problems:}
$$

$$
\beta \alpha_x = (K_x + c \Lambda)^{-1} r
$$

$$
\beta \alpha_y = (\Lambda^T \Lambda)^{-1} \Lambda^T r
$$

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 $\alpha_x$ and $\alpha_y$ 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, $X(n) = [x_0, x_1 - 1, \ldots, x_{n - L}]$ and $A(n) = [A_0, A_1 - 1, \ldots, A_{n - L}]$. Assuming that we have matrix $K_{reg}(n-1)$ for the $(n - 1)^{th}$ window of speech, and $K_{reg}(n-1)$ is the matrix formed by its last $n - 1$ rows and columns, then the regularized matrix for the current window is

$$
K_{reg}(n) = \begin{bmatrix} K_{reg}(n-1) & k_{n-1}(X(n)) \\ k_{n-1}(X(n))^T & k_{mn} + c \end{bmatrix},
$$

where $k_{n-1}(X(n)) = [k(x(n-L), x(n)), \ldots, k(x(n-n-1), x(n))]^T$ and $k_{mn} = k(X(n), X(n))$. The inverse of $K_{reg}(n)$ can also be computed quite quickly, given the inverse of $K_{reg}(n-1)$ [10]. We then iteratively update our parameter estimates for $\omega_y$ and $\alpha_y$ as new data arrives using eq. 5. This entire process is summarized in algorithm 1 and is based on work on Wiener systems by Varenenbergh 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 ($k^{(l)}_{h,poly}$),
begin
  Initialize $K^{(0)}_{reg} = (1 + c)I$
  Initialize $\alpha_n$ and $\omega_n$ with random data
  for $n = 1 \ldots N$ do
    Calculate $K^{(n)}_{reg}$ from $x^{(n)}$ as in eq. 6
    $K^{(n)}_{reg} = \kappa(x^{(n)}, x^{(n)})\alpha^{(n-1)}$
    $\omega^{(n)} = \Lambda^{(n)}\omega^{(n-1)}$
    $r^{(n)} = r^{(n-1)} + r^{(n-1)}$
    Calculate $(K^{(n)}_{reg})^{-1}$
    Update solutions for $\alpha_n$ and $\omega_n$ as in eq. 5
    Normalize solutions with $\beta = \|\omega_n\|$
  end
end

Algorithm 1: The adaptive KCCA algorithm.

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

- $K_{\text{poly}}^{(i)}(x_1, x_2) = (x_1 x_2)^i$
- $K_{\text{nh,poly}}^{(c)}(x_1, x_2) = (x_1 x_2 + 1)^i$
- $K_{\text{rbf}}^{(c)}(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right)$
- $K_{\text{sigmoid}}^{(c)}(x_1, x_2) = \tanh(x_1 x_2 + c)$

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 $(K, \alpha_n - \Lambda \omega_n)$ (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.

<table>
<thead>
<tr>
<th>Homogenous polynomial</th>
<th>Nonhomogenous polynomial</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$</td>
<td>MSE reduction</td>
</tr>
<tr>
<td>1</td>
<td>421.6</td>
</tr>
<tr>
<td>2</td>
<td>403.6</td>
</tr>
<tr>
<td>3</td>
<td>394.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sigmoid</th>
<th>Radial-basis function</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(\kappa, c)$</td>
<td>MSE reduction</td>
</tr>
<tr>
<td>(0.2, 0.1)</td>
<td>313.2</td>
</tr>
<tr>
<td>(0.2, 0.5)</td>
<td>321.5</td>
</tr>
<tr>
<td>(0.5, 0.1)</td>
<td>309.7</td>
</tr>
<tr>
<td>(0.5, 0.5)</td>
<td>314.3</td>
</tr>
</tbody>
</table>

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 tangent function given univariate input [7; 8], reaching MSE between $-30\text{dB}$ and $-40\text{dB}$ within 1000 to 1500 iterations. Surprisingly, most of the error in our experiments is dissipated much earlier, within 200 iterations, with MSE fluctuating between $-76.9\text{dB}$ and $39.5\text{dB}$ thereafter across all kernels and parameterizations. This suggests that adaptive KCCA converges rapidly during training on acoustic-articulatory data.
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 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.

![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.](image)

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.

<table>
<thead>
<tr>
<th>TV</th>
<th>MDN μ(σ²)</th>
<th>KCCA μ(σ²)</th>
<th>TV</th>
<th>MDN μ(σ²)</th>
<th>KCCA μ(σ²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VEL</td>
<td>−0.28 (0.08)</td>
<td>−0.25 (0.01)</td>
<td>TCD</td>
<td>−1.00 (0.11)</td>
<td>−1.00 (0.17)</td>
</tr>
<tr>
<td>LTH</td>
<td>−0.18 (0.12)</td>
<td>−0.18 (0.14)</td>
<td>TTCL</td>
<td>−1.62 (0.17)</td>
<td>−1.57 (0.16)</td>
</tr>
<tr>
<td>LA</td>
<td>−0.32 (0.11)</td>
<td>−0.28 (0.10)</td>
<td>TBCD</td>
<td>−0.79 (0.14)</td>
<td>−0.80 (0.15)</td>
</tr>
<tr>
<td>LP</td>
<td>−0.44 (0.12)</td>
<td>−0.41 (0.13)</td>
<td>TDCL</td>
<td>−0.20 (0.11)</td>
<td>−0.18 (0.09)</td>
</tr>
<tr>
<td>GLO</td>
<td>−1.30 (0.16)</td>
<td>−1.14 (0.15)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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