KERNEL CCA FOR MULTI-VIEW LEARNING OF ACOUSTIC FEATURES USING ARTICULATORY MEASUREMENTS

Raman Arora and Karen Livescu

Toyota Technological Institute at Chicago (TTIC), Chicago, IL 60637

arora@ttic.edu, klivescu@ttic.edu

ABSTRACT

We consider the problem of learning transformations of acoustic feature vectors for phonetic frame classification, in a multi-view setting where articulatory measurements are available at training time but not at test time. Canonical correlation analysis (CCA) has previously been used to learn linear transformations of the acoustic features that are maximally correlated with articulatory measurements. Here, we learn nonlinear transformations of the acoustics using kernel canonical correlation analysis (KCCA). We present an incremental SVD approach that makes the KCCA computations feasible for typical speech data set sizes. In phonetic frame classification experiments on data drawn from the University of Wisconsin X-ray Microbeam Database, we find that KCCA provides consistent improvements over linear CCA, as well as over single-view unsupervised dimensionality reduction.

Index Terms: multi-view learning, kernel canonical correlation analysis, XRMB, articulatory measurements

1. INTRODUCTION

Articulatory information has been used in automatic speech recognition in a number of ways [1]. Phonetic recognition can be improved if articulatory measurements are available as observations at test time [2], and word recognition may be slightly improved if articulatory measurements are observed in training and hidden at test time [3]. Knowledge-based approaches, in which articulatory information is never measured but rather used to constrain the hidden state structure, have also been proposed [4, 5].

In this work, we ask whether it is possible to use articulatory measurement data that is available only at training time to help learn which aspects of the acoustic feature vector are useful. This is a natural setting, in that articulatory data is much more feasible to collect at training time than at test time. Simultaneous acoustic and articulatory recordings are often collected for various purposes, and a number of public databases are available (e.g., [6]). We apply ideas from *multi-view learning*, in which multiple "views" of the data are available for training but possibly not for prediction (testing) [7].

A typical approach to acoustic feature vector generation in speech recognition is to first construct a high-dimensional acoustic feature vector by concatenating multiple consecutive frames of raw features, and then to reduce dimensionality using either an unsupervised transformation such as principal components analysis (PCA), a linear supervised transformation such as linear discriminant analysis (LDA) and its extensions, or a nonlinear supervised transformation [8]. Our approach here is unsupervised transformation learning, but using the second view (the articulatory measurements) as a form of "soft supervision". This avoids some of the disadvantages of unsupervised approaches, such as PCA, which are sensitive to noise and data scaling, and possibly of supervised approaches, which are more task-specific.

Recent related work [9] has taken this approach using canonical correlation analysis (CCA), which finds pairs of maximally correlated linear projections of data in two views [10, 11]. In this case, the two views are the acoustic and articulatory data, and only the acoustic projections are used at test time. The intuition is that articulatory measurements provide information about the linguistic content, and that the noise in the two views is largely uncorrelated and therefore filtered out by such a technique. CCA has also been used with audio and video for speaker clustering [12] and recognition [13]; for speaker normalization [14], where the views are the speakers; and to study critical articulators [15].

In this paper we extend this approach to non-linear acoustic transformations, using kernel canonical correlation analysis (KCCA) [11, 16], which allows us to learn richer acoustic features. KCCA has found limited use in speech research; the only work we are familiar with is [17] which uses KCCA for acoustic-articulatory inversion. One of the challenges of KCCA is factorizing a kernel matrix (through eigendecomposition or singular value decomposition) of the size of the training set, which is computationally infeasible for most typical speech corpus sizes. In this work we apply an incremental singular value decomposition (SVD) algorithm [18] to KCCA, making the approach feasible for speech tasks.

2. METHODS

We denote the measurable spaces of acoustic and articulatory signals by $\mathcal{X}$ and $\mathcal{Y}$ respectively. The Reproducing Kernel Hilbert Spaces (RKHS) of functions on $\mathcal{X}, \mathcal{Y}$ are denoted as $\mathcal{H}_{\mathcal{X}}$, $\mathcal{H}_{\mathcal{Y}}$ and the associated positive definite ker-

nels by k_x, k_y respectively. We consider a random vector $(X, Y) : \Omega \rightarrow \mathcal{X} \times \mathcal{Y}$ with an unknown joint distribution P_{XY}; the marginal distributions are denoted by P_X, P_Y . We have access to the joint distribution on acoustics and articulation only through n observations, $\{x_i, y_i\}_{i=1}^n$, which comprise our training data. Each pair (x_i, y_i) represents features computed over one frame of simultaneously recorded acoustics and articulation. We make the assumption that the two views are largely uncorrelated conditioned on the phonetic class which implies that the dimensions that are correlated between the two views will be discriminative for phonetic classification. This assumption, however, may not be satisfied in general. For instance, the audio and articulation may be correlated through the speaker identity or emotional state. In this work we restrict ourselves to speaker-dependent experiments which partially avoids this problem.

2.1. Kernel Canonical Correlation Analysis

Canonical correlation analysis (CCA) finds a pair of linear maps ($v \in \mathbb{R}^{d_x}$ and $w \in \mathbb{R}^{d_y}$, where d_x, d_y are the dimensionalities of the two views), that maximize the correlation between $v^T X$ and $w^T Y$ [10, 11]. The nonlinear extension of CCA is defined as the problem of finding functions $f_1 \in \mathcal{H}_x, g_1 \in \mathcal{H}_y$ that solve the optimization problem:

$$\{f_1, g_1\} = \underset{f \in \mathcal{H}_x, g \in \mathcal{H}_y}{\arg\max} \frac{\text{cov}\left(f(X), g(Y)\right)}{\sqrt{\text{var}\left(f(X)\right) \cdot \text{var}\left(g(Y)\right)}}, \quad (1)$$

i.e., that maximize correlation between random variables $f(X)$ and $g(Y)$. Subsequent KCCA directions, $\{f_j, g_j\}$, for $j > 1$, are found iteratively by solving (1) subject to the constraints that the random vector $(f_j(X), g_j(Y))^T$ is uncorrelated with $(f_i(X), g_i(Y))^T$ for all $i < j$.

Since the nonlinear maps $f \in \mathcal{H}_x$, $g \in \mathcal{H}_y$ are in RKHS, we can express them as linear combination of the kernel map evaluated at the data: $f(x) = \sum_{i=1}^n \alpha_i k_x(x, x_i)$, and similarly for $g(y)$. KCCA can then be written as finding directions $\alpha_1, \beta_1 \in \mathbb{R}^n$ that solve the optimization problem [16]

$$\{\alpha_1, \beta_1\} = \underset{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^n}{\arg\max} \frac{\alpha^T K_x K_y \beta}{\sqrt{(\alpha^T K_x^2 \alpha)\ (\beta^T K_y^2 \beta)}}, \quad (2)$$

where $K_x \in \mathbb{R}^{n \times n}$ is the centered Gram matrix:

$$\begin{aligned}[K_x]_{ij} &= k_x(x_i, x_j) - \frac{1}{n}\sum_{i=1}^n k_x(x_i, x_j) \\ &\quad - \frac{1}{n}\sum_{j=1}^n k_x(x_i, x_j) + \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k_x(x_i, x_j),\end{aligned}$$

and K_y is defined similarly. The subsequent weight vectors $\{\alpha_j, \beta_j\}$ are found iteratively by solving (2) with the constraints that resulting maps $\{f_j, g_j\}$ yield random variables uncorrelated with the previous ones.

Due to the high dimensionality of the feature space, with KCCA we run an "elevated" risk of over-fitting when trying to maximize (2). In particular, if K_x or K_y is invertible, then we can find directions $(\hat{\alpha}, \hat{\beta})$ that give perfect correlations [11]. To remedy this, Hardoon et al. [11] propose maximizing the following regularized objective function,

$$\frac{\alpha^T K_x K_y \beta}{\sqrt{(\alpha^T K_x^2 \alpha + r_x \alpha^T K_x \alpha)\ (\beta^T K_y^2 \beta + r_y \beta^T K_y \beta)}}, \quad (3)$$

which is maximized by the top eigenvectors (corresponding to the largest eigenvalue) of the matrix

$$(K_x + r_x I)^{-1}\ K_y\ (K_y + r_y I)^{-1}\ K_x. \quad (4)$$

It can be checked that the best m KCCA directions are given by the eigenvectors of matrix in (4) corresponding to the m largest eigenvalues. In other words, the m-dimensional function space, in $\mathcal{H}_\mathcal{X} \times \mathcal{H}_\mathcal{Y}$, that captures maximal correlation between X and Y is determined by the eigenvectors of matrix in (4) associated with the m largest eigenvalues; the total correlation is given as the sum of those eigenvalues. We reduce dimensionality by projecting K_x onto the subspace spanned by the top m eigenvectors of (4). The parameters r_x and r_y are tuned on held-out data.

2.2. Scalable KCCA

One of the statistical challenges in solving the regularized KCCA problem is possible degeneracy of inverting large kernel matrices when we have a lot of training data. This problem can be alleviated by decomposing the kernel matrix as a gram-product of two lower dimensional matrices, i.e.

$$K_x \approx \hat{F}^T \hat{F}, \quad K_y \approx \hat{G}^T \hat{G}, \quad (5)$$

where $\hat{F}, \hat{G} \in \mathbb{R}^{m \times n}$, for some intermediate dimensionality $m \ll n$. These form a low-dimensional representation of maps f, g evaluated at the training data $\{x_i, y_i\}_{i=1}^n$. Define

$$\begin{aligned}\hat{C}_{ff} &= \hat{F}\hat{F}^T, \quad \hat{C}_{gg} = \hat{G}\hat{G}^T, \\ \hat{C}_{fg} &= \hat{F}\hat{G}^T, \quad \hat{C}_{gf} = \hat{G}\hat{F}^T. \end{aligned} \quad (6)$$

The KCCA directions $(\hat{\alpha}, \hat{\beta})$ in the reduced dimensionality are related to the true KCCA directions (α, β) via $\hat{\alpha} = \hat{F}\alpha$ and $\hat{\beta} = \hat{G}\beta$. As in linear CCA, the reduced dimensionality KCCA directions are solutions to the following generalized eigenvalue problem:

$$\begin{aligned}\hat{C}_{ff}^{-1}\hat{C}_{fg}\hat{C}_{gg}^{-1}\hat{C}_{gf}\hat{\alpha} &= \lambda^2\hat{\alpha} \\ \hat{\beta} &\propto \hat{C}_{gg}^{-1}\hat{C}_{gf}\hat{\alpha}.\end{aligned}$$

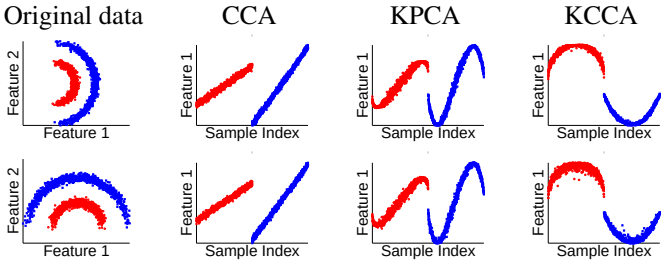

Fig. 1. Simulated data and various projections. The two rows correspond to data from two views.

Input: $\{x_i, y_i\}_{i=1}^n$, m (dimensionality), b (block size)
Output: Basis for KCCA subspace: $\alpha \in \mathbb{R}^{k \times n}$

Set $S_x \leftarrow 0, U_x \leftarrow \mathbf{0}, V_x \leftarrow \mathbf{0}$
Set $S_y \leftarrow 0, U_y \leftarrow \mathbf{0}, V_y \leftarrow \mathbf{0}$
for $i \leftarrow 1$ **to** p **do**
  for $j \leftarrow 1$ **to** n *in steps of* b **do**
    Compute $C_x \leftarrow$ columns $j, \ldots, j+b-1$ of K_x
    $(U_x, S_x, V_x) \leftarrow incrSVD(C_x, U_x, S_x, V_x)$
    Compute $C_y \leftarrow$ columns $j, \ldots, j+b-1$ of K_y
    $(U_y, S_y, V_y) \leftarrow incrSVD(C_y, U_y, S_y, V_y)$
  end
  $\hat{F} \leftarrow S_x^{\frac{1}{2}} U_x^T, \hat{G} \leftarrow S_y^{\frac{1}{2}} U_y^T$
  Compute $\hat{C}_{ff}, \hat{C}_{gg}, \hat{C}_{fg}, \hat{C}_{gf}$ as given in (6)
  $\hat{\alpha} \leftarrow$ eigenvectors of $\hat{C}_{ff}^{-1} \hat{C}_{fg} \hat{C}_{gg}^{-1} \hat{C}_{gf}$
  $\alpha \leftarrow \left(\hat{F}^T \hat{F}\right)^{-1} \hat{F}^T \hat{\alpha}$
end

Algorithm 1: Pseudocode for KCCA.

The KCCA algorithm given in Algorithm 1 returns a basis, α, for the maximally correlated subspace. The projections of the training and test acoustic features are given by $\hat{X} = \alpha^T K_x$ and $\hat{X}^{(test)} = \alpha^T K_x^{(test)}$, where $[K_x^{(test)}]_{ij} = k_x(x_i, x_j^{(test)})$ is the kernel evaluated at the i^{th} training example and j^{th} test example. The most computationally expensive step in KCCA is finding the decomposition in eqn. (5). A standard approach would be to find a truncated rank-m singular value decomposition (SVD) of the matrix $K_x \approx U_x S_x U_x^T$. However, a batch approach to finding SVD would be infeasible for large training sets. We use a block incremental approach [18], described in Algorithm 2, for finding the SVD of K_x, K_y. The computational complexity of the incremental algorithm is $O(m^2 n)$ and the space complexity is $O(mn)$ (compared to $O(n^3)$ computational complexity and $O(n^3)$ space complexity for the batch approach).

A qualitative comparison of CCA, KCCA and Kernel PCA is shown in Fig. 1 for a toy example of two-view, two-dimensional data drawn from two classes. The two classes can be separated by thresholding the top KCCA projection, whereas the top CCA and KPCA projection fail to separate the two classes; whereas CCA fails because the noise is non-linear, KPCA fails because it learns only the noise. KCCA takes advantage of the fact that the non-linear noise in the two views is uncorrelated.

3. EXPERIMENTS

We compare KCCA to CCA and to single-view unsupervised dimensionality reduction (PCA, KPCA), on phonetic frame classification with k-nearest neighbor (kNN) classifiers using the correlation distance and support vector machines (SVM) using a radial basis function (RBF) and a one-vs.-one multi-class implementation [19]. We use RBF kernels for both views for KCCA with $k_x(x_i, x_j) = e^{-\|x_i - x_j\|^2 / 2\sigma_x^2}$ and k_y defined similarly. We use a five-fold cross-validation setup to tune over the hyper-parameters (dimensionality k, regularization parameters r_x, r_y, kernel bandwidths σ_x, σ_y, neighborhood size for kNN, and kernel width and cost for SVMs). Each fold consists of 60% of the utterances for training, 20% for tuning (development), and 20% for final testing.

Input: Columns $C = [c_1, \ldots, c_b]$ of kernel matrix K, rank-r estimate $(U_{n\times r}, S_{r\times r}, V_{t\times r})$ of the SVD of the first t columns of matrix K
Output: updated rank-r SVD $\left(U_{n\times r}, S_{r\times r}, V_{(t+b)\times r}\right)$
Compute projection $L_{r\times b} = U^T C$
Compute residual $H_{n\times b} = C - UL$
Compute QR decomposition $H = J_{n\times b} W_{b\times b}$
Compute $Q_{(r+b)\times(r+b)} = \begin{bmatrix} S_{r\times r} & L_{r\times b} \\ \mathbf{0}_{b\times r} & W_{b\times b} \end{bmatrix}$
Compute SVD, $Q = \tilde{U}\tilde{S}\tilde{V}^T$
Update

$$\begin{aligned} U_{n\times(r+b)} &= [U\ J]\,\tilde{U}_{(r+b)\times(r+b)} \\ S_{(r+b)\times(r+b)} &= \tilde{S}_{(r+b)\times(r+b)} \\ V_{(t+b)\times(r+b)} &= \begin{bmatrix} V_{t\times r} & \mathbf{0}_{t\times b} \\ \mathbf{0}_{b\times r} & W_{b\times b} \end{bmatrix} \tilde{V}_{(r+b)\times(r+b)} \end{aligned}$$

Sort the singular values in S and corresponding singular vectors (in U, V) in a descending order
Truncate: $S = S_{1:r,1:r}, U = U_{:,1:r}, V = V_{:,1:r}$

Algorithm 2: Pseudocode for incremental SVD.

We use a subset of the University of Wisconsin X-ray Microbeam Database (XRMB) [6] of simultaneous acoustic and articulatory recordings. The articulatory data consist of horizontal and vertical displacements of eight pellets on the speaker's lips, tongue, and jaws, yielding a 16-dimensional vector at each time point. Our experiments are speaker-dependent, using the two speakers JW11 (male) and JW30 (female). Our baseline acoustic features consist of 13-dimensional mel-frequency cepstral coefficients (MFCCs) and their first and second derivatives computed every 10ms over a 25ms window. The articulatory data is downsampled to match the MFCC frame rate. For each frame we concatenate acoustic features over a 3-frame window and articulatory features over a 7-frame window, giving acoustic data $X \in \mathbb{R}^{117\times N}$ and articulatory data $Y \in \mathbb{R}^{112\times N}$, where N is the number of frames. We discard any frame with missing measurements, resulting in $N \approx 50,000$ frames for each speaker. For classification experiments, phone labels are obtained using the Penn Phonetics Lab Forced Aligner [20]. Short pauses and stress are removed, leaving 39 phone classes.

In addition to CCA and KCCA, we also consider appending the learned projections to the MFCCs, and refer to these representations as MFCCA (MFCC+CCA) and KMFCCA

Table 1. 5-fold average phonetic frame error rates. Boldface font indicates the best performance in each column. An asterisk indicates a significant improvement over the 39-dimensional MFCC baseline (as per a t-test with p = 0.05).

speaker	JW11		JW30	
classifier	kNN	SVM	kNN	SVM
MFCC (39)	31.98	30.58	37.59	36.15
PCA	30.85*	26.83*	36.07	32.48*
KPCA	30.40*	28.80*	35.20*	34.73
CCA	28.95*	27.90*	34.29*	32.89*
MFCCA	27.89*	25.93*	33.46*	31.05*
KCCA	27.79*	26.61*	32.70*	31.95*
KMFCCA	**26.87***	**24.99***	**32.02***	**30.01***

(MFCC+KCCA). This is based on the observation, made previously by us and others [9], that CCA and KCCA capture only those dimensions that are correlated across views, but there may be additional useful information in the acoustics that is not correlated with the articulatory measurements. For example, the XRMB data that we use does not include glottal or velar measurements, so the CCA and KCCA projections capture little information about voicing or nasality.

Regularization parameters r_x, r_y for CCA and KCCA were tuned over the range $[10^{-8}, 10]$ and PCA/CCA dimensionalities were tuned over $[10, 110]$. Kernel bandwidths were fixed at $\sigma_x = 4 \times 10^6, \sigma_y = 2 \times 10^4$. Table 1 shows the test set error rates averaged over five folds.[1] KCCA consistently improves over CCA by $1-1.5\%$ absolute. KMFCCA consistently improves over MFCCA by $\sim 1\%$ and over the baseline by roughly $5-6\%$ absolute. Improvements are seen across all phone classes, with the largest improvements being on consonants, especially labials. The single largest improvement is on the phone [s], which has a very high prior probability.

Figure 2 shows the top two CCA and KCCA projections for frames corresponding to monophthongs, showing that the KCCA projections form better-separated clusters, with the clusters forming roughly the standard vowel quadrilateral.

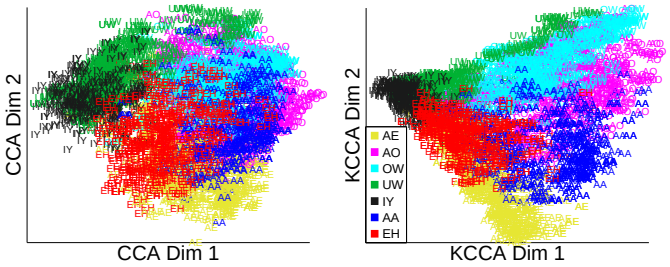

Fig. 2. Top two CCA (left) and KCCA (right) projections for frames corresponding to monophthongs.

[1]Experiments reported here are identical to those in [9], except for improved data preparation (in [9], frames adjacent to mis-tracked frames were concatenated). We also use 39-dimensional MFCCs as the baseline, rather than higher-dimensional windowed features.

4. CONCLUSION

Our results show the potential benefit of using kernel CCA in multi-view learning of acoustic feature transformations when articulatory measurements are available during training. We found consistent gains with KCCA features combined with the baseline MFCC features. Future work will address speaker- and domain-independence to make the approach more applicable to new speakers and domains where articulatory data is not necessarily available at all, as well as using the learned transformations in word recognition. Another possible extension is dynamic context-sensitive CCA to take into account changes in critical articulator.

5. REFERENCES

[1] S. King *et al.*, "Speech production knowledge in automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[2] J. Frankel and S. King, "ASR - articulatory speech recognition," in *Eurospeech*, 2001.

[3] K. Markov *et al.*, "Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework," *Speech Communication*, vol. 48, pp. 161–175, 2006.

[4] L. Deng *et al.*, "Production models as a structural basis for automatic speech recognition," *Speech Comm.*, vol. 33, pp. 93–111, 1997.

[5] K. Livescu *et al.*, "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop," in *ICASSP*, 2007.

[6] J. R. Westbury, *X-ray microbeam speech production database user's handbook*, Waisman Center on Mental Retardation & Human Development, U. Wisconsin, Madison, WI, version 1.0 edition, June 1994.

[7] S. M. Kakade and D. P. Foster, "Multi-view regression via canonical correlation analysis," in *COLT*, 2007.

[8] H. Hermansky *et al.*, "Tandem connectionist feature extraction for conventional HMM systems," in *ICASSP*, 2000.

[9] S. Bharadwaj *et al.*, "Multiview acoustic feature learning using articulatory measurements," in *Intl. Workshop on Stat. Machine Learning for Speech Recognition*, 2012.

[10] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[11] D. R. Hardoon *et al.*, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[12] K. Chaudhuri *et al.*, "Multi-view clustering via canonical correlation analysis," in *ICML*, 2009.

[13] K. Livescu and M. Stoehr, "Multi-view learning of acoustic features for speaker recognition," in *ASRU*, 2009.

[14] K. Choukri and G. Chollet, "Adaptation of automatic speech recognizers to new speakers using canonical correlation analysis techniques," *Speech Comm.*, vol. 1, pp. 95–107, 1986.

[15] T. Kato, S. Lee, and S. Narayanan, "An analysis of articulatory-acoustic data based on articulatory strokes," in *ICASSP*, 2009.

[16] K. Fukumizu *et al.*, "Statistical consistency of Kernel Canonical Correlation Analysis," *Journal of Machine Learning Research*, vol. 8, pp. 361–383, 2007.

[17] F. Rudzicz, "Adaptive kernel canonical correlation analysis for estimation of task dynamics from acoustics," in *ICASSP*, 2010.

[18] M. Brand, "Incremental singular value decomposition of uncertain data with missing values," in *Eur. Conf. on Comp. Vision*, 2002.

[19] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[20] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," in *Acoustics*, 2008.