Using articulatory measurements to learn better acoustic features

Galen Andrew[1], Raman Arora[2], Sujeeth Bharadwaj[3],
Jeff Bilmes[1], Mark Hasegawa-Johnson[3], Karen Livescu[2]

[1]U. Washington, Seattle, USA [2]TTI-Chicago, USA [3]U. Illinois at Urbana-Champaign, USA

galen@cs.washington.edu, arora@ttic.edu, sbhara3@uiuc.edu,
bilmes@ee.washington.edu, jhasegaw@uiuc.edu, klivescu@ttic.edu

Abstract

We summarize recent work on learning improved acoustic features, using articulatory measurements that are available for training but not at test time. The goal is to improve recognition using articulatory information, but without explicitly solving the difficult acoustics-to-articulation inversion problem. We formulate the problem as learning a (linear or nonlinear) transformation of standard acoustic features, such that the transformed vectors are maximally correlated with some (linear or nonlinear) transformation of articulatory measurements. This formulation leads to the standard statistical technique of canonical correlation analysis (CCA) and its nonlinear extension kernel CCA. Along the way, we have developed a scalable variant of kernel CCA and a new type of nonlinear CCA via deep neural networks (deep CCA). The learned features can improve phonetic classification and recognition and generalize across speakers, and deep CCA shows promise over kernel CCA.[1]

1. Introduction

Articulatory measurements have been used in speech recognition in a number of ways. Recognition can be improved if measurements of articulatory motions are available at test time (e.g. [5]), but this is an impractical setting. In the absence of articulatory data at test time, we may consider predicting the articulation from acoustics and using the predicted values as additional observations. However, acoustic-to-articulatory inversion is a complex task, and to date it has been difficult to improve recognition performance in this way [5]. The work described here takes a different approach, based on the intuition that while predicting articulation may be difficult, learning acoustic features that are somehow *informed* by articulation may be easier. We have previously reported on this line of work [1, 2, 3, 4] and summarize the methods and main results here.

A common approach to acoustic feature learning is to first construct a high-dimensional acoustic feature vector by concatenating multiple consecutive frames of raw features such as MFCCs or PLPs, and then to reduce dimensionality in a supervised or unsupervised way (using PCA, LDA, neural networks, etc.). In this work we consider unsupervised transformation learning, but in a setting where parallel acoustic and articulatory measurements are available for some training data (but not at test time). We ask whether we can use the articulatory information to learn which linear or nonlinear directions in the acoustic space are most useful. Such an approach avoids some of the disadvantages of unsupervised approaches, such as PCA, which are sensitive to noise and data scaling, and possibly also of supervised approaches, which are task-specific.

Our approach is based on canonical correlation analysis (CCA), a standard statistical technique that finds pairs of maximally correlated linear projections of data in two views [6], and its nonlinear extensions. The most common nonlinear extension is kernel CCA [7]; we have also developed a deep network-based extension, deep CCA [4]. The two views here 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 much of the non-discriminative information in the two views is largely uncorrelated and therefore filtered out.

2. Methods

Let $X \in \mathcal{X}, Y \in \mathcal{Y}$ denote random vectors in two views, and let $\mathcal{H}_X, \mathcal{H}_Y$ denote spaces of real-valued functions on $\mathcal{X}$ and $\mathcal{Y}$, respectively. We are given N paired training examples, $\{x_i, y_i\}_{i=1}^N$, drawn from the (unknown) joint distribution of X, Y. In our case, each pair (x_i, y_i) represents feature vectors of acoustics (x_i) and articulation (y_i). All of the methods we use are based on finding pairs of functions, $f(X) \in \mathcal{H}_\mathcal{X}$ and $g(Y) \in \mathcal{H}_\mathcal{Y}$, that are highly correlated. We can think of these functions as (linear or nonlinear) projections of X and Y. The first such pair of functions is the one with maximal correlation:

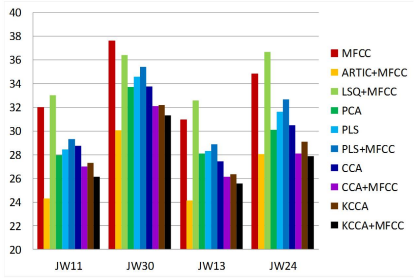
$$\{f_1, g_1\} = \underset{f \in \mathcal{H}_x, g \in \mathcal{H}_y}{\arg\max} \frac{\text{cov}\left(f(X), g(Y)\right)}{\sqrt{\text{var}\left(f(X)\right) \cdot \text{var}\left(g(Y)\right)}}, \qquad (1)$$

Subsequent pairs $\{f_j, g_j\}$ for $j > 1$ are solutions to (1) subject to the constraints that $f_j(X)$ is uncorrelated with $f_i(X)$, $g_j(Y)$ is uncorrelated with $g_i(Y)$ and $f_j(X)$ is uncorrelated with $g_i(Y)$ for all $i \neq j$. In our case X and Y are high-dimensional and we perform dimensionality reduction of X by keeping the k first (most correlated) projections $f_1(X), \ldots, f_k(X)$. This k-dimensional vector is our learned feature vector.

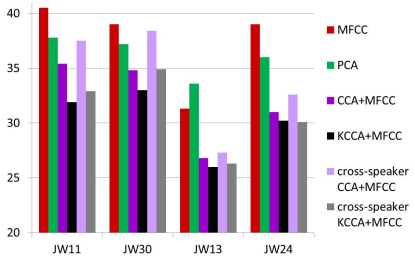
CCA: CCA solves Problem (1) for the case where the projection functions are linear, $f(X) = v^T X, g(Y) = w^T Y$. In this case the solution is straightforward: The vectors v that maximize the objective are the top eigenvectors of $C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx}$, and w are given as $w \propto C_{yy}^{-1} C_{yx} v$, where C_{xx}, C_{yy} are the autocovariance matrices in each view and C_{xy} is the cross-covariance matrix between X and Y. To alleviate over-fitting, one typically regularizes the problem by replacing the autocovariance matrices with $C_{xx} + r_x I$ and $C_{yy} + r_y I$, where r_x, r_y are regularization parameters.

KCCA: Kernel CCA (KCCA) is a nonlinear extension to CCA, and is useful when the relationship between the two views is believed to be nonlinear (as in our case). In KCCA, the

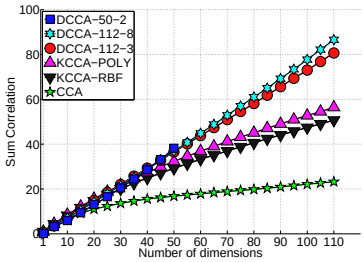
[1]Many details and citations are left out for brevity; please see [1, 2, 3, 4] for a more complete description. This research was supported by NSF grant IIS-0905633 and by the Intel/UW ISTC. The opinions expressed in this work are those of the authors and do not necessarily reflect the views of the funders.



(a) Speaker-dependent phonetic frame error rates (%).



(b) Phonetic recognition error rates (%), using a 3-state monophone HMM-GMM recognizer.



(c) Total correlation (sum over all dimensions) between acoustic and articulatory projections.

function spaces $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ are Reproducing Kernel Hilbert Spaces (RKHS) with associated positive definite kernels k_x, k_y. Using the "kernel trick" one can express f and g as linear combinations of the kernel evaluated at the data: $f(x) = \sum_{i=1}^{N} \alpha_i k_x(x, x_i)$, and similarly for $g(y)$. Problem (1) then becomes one of finding directions $\alpha_1, \beta_1 \in \mathbb{R}^N$ that satisfy

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

where $K_x \in \mathbb{R}^{N \times N}$ is the centered version of the Gram matrix $K_{ij} = k_x(x_i, x_j)$, and similarly for K_y. Subsequent vectors $\{\alpha_j, \beta_j\}$ are solutions of (2) with the constraints that the resulting $\{f_j(X), g_j(Y)\}$ are uncorrelated with the previous ones.

The optimal vectors, after a similar regularization, are the top eigenvectors of $(K_x + r_x I)^{-1}\ K_y\ (K_y + r_y I)^{-1}\ K_x$, where r_x and r_y are regularization parameters. In practice, the kernel matrices may be too large for this computation. We have addressed this issue with a variant of KCCA in which, first, the kernel matrices are reduced to a lower rank via an efficient incremental SVD; and then, a linear CCA is done in the resulting intermediate-dimensionality space. See [2, 3] for details.

DCCA: Deep CCA (DCCA) is a different nonlinear extension of CCA [4]. In DCCA, rather than constraining the nonlinear projections to be in RKHS, they are the outputs of neural networks with multiple hidden layers. In each layer, each node computes a nonlinear function of a linear combination of the inputs from the previous layer. In our implementation of DCCA, we have found that the types of nonlinearities typically used in neural network research (e.g., logistic) do not work well for DCCA, and have introduced a nonlinearity related to the cube root which does not saturate (asymptote) for very high positive/negative inputs. Training is performed by backpropagation of gradients of the correlation objective through the layers of the network. For more details, see [4]. The benefits of DCCA over KCCA include that it is not constrained to a particular choice of kernel function, and that as a parametric technique, it does not require accessing the training data at test time (as KCCA does).

3. Experimental results

We have evaluated CCA and KCCA in terms of phonetic frame classification and phonetic recognition, and deep CCA in terms of correlation between the projected acoustics and articulation, using a subset of the University of Wisconsin X-ray Microbeam Database (XRMB) [8]. The input acoustic features are MFCCs and their derivatives concatenated over seven frames (273 dimensions); the input articulatory features are horizontal and vertical displacements of 8 pellets on the speaker's lips, tongue, and jaw, concatenated over seven frames (112 dimensions).

Figs. 1(a) and 1(b) show frame classification and phonetic recognition results on four speakers (JW11, JW30, JW13, JW24) from XRMB, using a k-NN frame classifier. ARTIC+MFCC is the "gold standard" of appending the true articulatory measurements to the MFCCs; LSQ+MFCC refers to concatenation of MFCCs with articulatory predictions using least-squares; PLS refers to partial least squares, an alternative to CCA in which one maximizes the covariance rather than correlation. The recognition results include speaker-dependent and cross-speaker projection learning (in the latter case, there is one test speaker and the projections are learned on the remaining three speakers). The main results are that CCA- and KCCA-based features appended to MFCCs outperform the baseline features and other projections, and KCCA outperforms CCA. The most exciting result is that the performance of KCCA takes a very small (or no) hit in the cross-speaker condition, for which articulatory data has not been seen at all for the target speaker. This result suggests that the learned features generalize beyond the typical speaker-dependent settings in which articulatory measurements have previously been found to be helpful.

Fig. 1(c) shows acoustic-articulatory correlations obtained with CCA, KCCA, and DCCA on unseen test data. KCCA-POLY and KCCA-RBF refers to KCCA with a polynomial/radial basis function kernel. DCCA-o-d refers to o output units and d layers. DCCA convincingly outperforms the other techniques except at low dimensionalities.

These results demonstrate some potential for acoustic features learned with CCA, KCCA, and DCCA. The results transfer well to new speakers for which we have no articulatory data. In cross-domain experiments on data outside of XRMB, the improvements have been more modest [3]. Ongoing work is exploring better transfer to new domains, acoustic conditions, etc.; alternatives to appending the learned projections to MFCCs, by retaining just a "sufficient subspace" of the MFCCs that is uncorrelated with the learned projections [1]; sparse and dynamic extensions; and applying DCCA to speech recognition tasks.

4. References

[1] S. Bharadwaj *et al.*, "Multiview acoustic feature learning using articulatory measurements," *IWSML* 2012.

[2] R. Arora and K. Livescu, "Kernel CCA for multi-view learning of acoustic features using articulatory measurements," *MLSLP* 2012.

[3] R. Arora and K. Livescu, "Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains," *ICASSP* 2013.

[4] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," *ICML* 2013.

[5] A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," *ICSLP* 2000.

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

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

[8] J. R. Westbury, *X-ray microbeam speech production database user's handbook*, Waisman Center, U. Wisconsin, 1994.