MULTI-VIEW LEARNING WITH SUPERVISION FOR TRANSFORMED BOTTLENECK FEATURES

Raman Arora and Karen Livescu

TTI-Chicago, Chicago, IL 60637
arora@ttic.edu, klivescu@ttic.edu

ABSTRACT

Previous work has shown that acoustic features can be improved by unsupervised learning of transformations based on canonical correlation analysis (CCA) using articulatory measurements that are available at training time. In this paper, we investigate whether this second view (articulatory data) still helps even when labels are also available at training time. We begin with strong baseline bottleneck features, which can be learned when the training set is phonetically labeled. We then compare several options for learning transformations of the bottleneck features in the presence of both articulatory measurements and phonetic labels for the training data. The methods compared include combinations of LDA and CCA, as well as a three-view extension of CCA that simultaneously uses the labels and articulatory measurements as additional views. Phonetic recognition experiments on data from the University of Wisconsin X-ray microbeam database show that the learned features improve performance over using either just the labels or just the articulatory measurements for learning acoustic transformations.

Index Terms— multi-view learning, canonical correlation analysis, articulatory measurements, bottleneck features, supervised transformation learning

1. INTRODUCTION

One of the main recent improvements to automatic speech recognition has come from improving the acoustic features by learning transformations of basic features. A typical approach is to construct high-dimensional features, by concatenating multiple consecutive frames of basic features (e.g., mel-frequency cepstral coefficients, perceptual linear prediction coefficients) around the current frame, and then to reduce dimensionality using a transformation that is learned from data. Such learned transformations can include:

- Unsupervised linear transformations, such as principal components analysis (PCA)
- Unsupervised nonlinear transformations, such as graph-based methods [1]
- Supervised linear transformations, such as linear discriminant analysis (LDA) and its extensions [2, 3]
- Supervised nonlinear transformations, typically using neural networks with phonetic (or state) labels as outputs, such as tandem [4] and bottleneck [5] processing

A research question that has recently been pursued is whether it is possible to learn better transformations if we have access to another view of the speech data at training time, in particular articulatory measurements. Articulatory measurements clearly help if they are available at test time [6, 7], and may also help in the more realistic scenario where they are only available at training time and not test time [8]. We have found that such an additional view can indeed help learn improved acoustic feature transformations, using a multi-view learning approach based on canonical correlation analysis (CCA) [9, 10, 11]. Encouragingly, the improvements seem to hold for speakers for whom no articulatory measurements are available even at training time [11].

Our previous work on CCA-based acoustic feature transformation learning has been in the unsupervised setting. There are certain virtues to this setting, but frame labels are often available, or relatively high-quality ones can be obtained via forced transcription. In such cases much better acoustic features can be obtained, in particular ones based on neural networks as in the tandem [4] and bottleneck [5] approaches. In this paper, we ask whether in this supervised setting, we can start from strong supervised features (in our case bottleneck features) and still obtain an improved feature transformation using the additional information from articulatory measurements and labels at training time.

We consider several ways of combining a second view with labels, using either combinations of CCA and LDA or three-view extensions of CCA that consider the labels to be a third view. Fig. 1 gives a pictorial overview of the approach. We present phonetic recognition experiments on data from the University of Wisconsin X-ray microbeam database (XRMB) [12], comparing the techniques and showing that indeed the new transformations improve over both unsupervised CCA or supervised transformations alone.

This research was supported by NSF grants IIS-0905633 and IIS-1321015. The opinions expressed in this work are those of the authors and do not necessarily reflect the views of the funding agency.
We will use the following shorthand for Problem (1):
\[
\begin{bmatrix} \hat{U}, \hat{V} \end{bmatrix} = \arg \max_{U,V} \rho(U^TX, V^TY).
\]
(2)

where \( \rho \) represents correlation (in this case it should be thought of as multidimensional correlation as above). The solution of Eq. (1) is straightforward [15]: The matrix \( \hat{U} \) consists of the top \( m \) eigenvectors of \( \hat{C}_{xx} = \frac{1}{N} XX^T \) and \( \hat{C}_{yy} = \frac{1}{N} YY^T \) are the sample autocovariance matrices and \( \hat{C}_{xy} = \frac{1}{N} XY^T \) is the sample cross-covariance matrix. The matrix \( \hat{V} \) is composed of columns \( v_k \propto \hat{C}_{yx}^{-1} \hat{C}_{yy} u_k \); in our case, we discard \( \hat{V} \) since we only have access to the acoustic view at test time. In practice, in order to avoid ill-conditioned auto-covariance matrices and mitigate the effects of noise in the data, we regularize by adding a term \( r_x I \) to the auto-covariance estimate \( \hat{C}_{xx} \) and \( r_y I \) to \( \hat{C}_{yy} \).

The largest possible number of learned projections is \( \min(d_x, d_y); \) in order to reduce dimensionality, we retain only the top \( m < \min(d_x, d_y) \) projection vectors, and these form the desired final transformation matrix \( L = \hat{U} = [u_1 \ldots u_m] \).

Nonlinear extensions of CCA can be derived by replacing the linear projections with nonlinear functions \( f(X), g(Y) \) defined via kernels [17, 15] or neural networks [18, 13].

A common motivation for the use of CCA is that if the two views are largely uncorrelated conditioned on some class of interest (here, the phonetic class), then the dimensions that are correlated should be discriminative. In our case, a second motivation is that we know the second view (articulation) to be useful if available at test time, but this is an unrealistic setting, and predicting articulation from acoustics is very challenging and perhaps not necessary; finding good correlated projections may be sufficient.

**Supervised single-view learning via neural networks**: In the single-view supervised setting, neural networks are often used to learn nonlinear transformations. The transformed features are outputs of either the network’s final layer (as in tandem systems [4]) or a narrow hidden layer (bottleneck features [5]). In this work, we use bottleneck features and consider these to be our input \( X \) for all supervised methods.
2.1. Supervised multi-view learning

Multi-view feature learning is typically used as an unsupervised technique, but when training labels are also available, we can use them to learn features that are more discriminative. Why might the second view (here, articulation) be useful even in the presence of labels? One reason is that this second view may “guide” the learning into a good part of the space when directly finding a discriminative subspace is difficult, for example by identifying and removing noisy dimensions. Another reason is that the labels may not be for the target task, but rather for a proxy task: this is often the case in acoustic feature learning, where phonetic frame labels are used but the ultimate task is continuous phone/word recognition.

We next describe several ways of combining labels with the second articulatory view for feature learning. We assume the labels are represented as vectors in \( \mathbb{R}^C \) using the “one-hot” representation of Eq. (3). There are a number of techniques that have been suggested for this setting, including multi-view extensions of LDA [19, 20], labeled extensions of CCA [21], and generalizations that subsume both of these [22]. In this work we select a few approaches that are straightforward to solve via one or two eigenproblems.

2.1.1. CCA on bottleneck features

The first way we combine two views with labels is to simply apply CCA to the learned bottleneck features. The remaining methods below further use the labels in a second stage.

2.1.2. CCA with concatenated labels and articulatory data

One straightforward approach is to consider the labels to be part of the second view, by appending them to the articulatory data and then using CCA. The transformation matrix is then given by solving the modified CCA problem

\[
\begin{align*}
\{ \hat{U}, \hat{V} \} & = \arg \max_{U,V} \rho \left( U^\top X, V^\top \begin{bmatrix} Y \\ Z \end{bmatrix} \right) \\
\{ \hat{W}, \hat{T} \} & = \arg \max_{W,T} \rho \left( W^\top X, T^\top Z \right).
\end{align*}
\]

and the final acoustic feature transformation is \( L = \hat{U} \).  

2.1.3. Concatenated LDA and CCA features

We can also consider solving two multi-view problems, one with articulatory data as the second view and another with the labels as the second view, and concatenating the results; i.e., concatenating CCA and LDA features. Let \( \hat{U}, \hat{W} \) denote transformations learned with CCA and LDA, respectively:

\[
\begin{align*}
\{ \hat{U}, \hat{V} \} & = \arg \max_{U,V} \rho \left( U^\top X, V^\top Y \right) \\
\{ \hat{W}, \hat{T} \} & = \arg \max_{W,T} \rho \left( W^\top X, T^\top Z \right).
\end{align*}
\]

Our final feature transformation matrix is then \( L = \begin{bmatrix} \hat{U} \\ \hat{W} \end{bmatrix} \).

2.1.4. LDA on CCA features

Another option is to first learn CCA-based features, and then to learn a supervised transformation of those features via LDA. The CCA step could be done on either raw features (unsupervised) or on the supervised bottleneck features as in Sec. 2.1.1; here we only consider the latter. The final transformation is given by

\[
\begin{align*}
\{ \hat{W}, \hat{T} \} & = \arg \max_{W,T} \rho \left( W^\top \hat{U}^\top X, T^\top Z \right),
\end{align*}
\]

where \( \hat{U} \) is given in (5). The final transformation is \( L = \hat{U} \hat{W} \).

2.1.5. Many-view learning with generalized CCA (GCCA)

There are several generalizations of CCA for any number of views \( J \geq 2 \) [23, 24, 25]. We follow the formulation introduced by Carroll [24], which seeks a common latent representation (“group configuration”) \( G \in \mathbb{R}^{m \times N} \) and view-specific transformation matrices \( \hat{U}_j \in \mathbb{R}^{d_j \times m}, 1 \leq j \leq J \), that solve

\[
\begin{align*}
\text{minimize} & \sum_{j=1}^J \| G - U_j X_j \|_F^2, \\
\text{subject to} & \quad GG^\top = I
\end{align*}
\]

where \( X_j \in \mathbb{R}^{d_j \times N} \) is the feature matrix for the \( j \)th view. In our case \( J = 3 \), \( X_1 = X, X_2 = Y, X_3 = Z \), and the final transformation matrix is \( L = \hat{U}_1 \). \( G \) is given as the solution to the eigenvalue problem [25]:

\[
\left( \sum_{j=1}^J X_j^\top (X_j X_j^\top)^{-1} X_j \right) G^\top = G^\top \Lambda,
\]

and the transformation matrices \( \hat{U}_j \) are given as

\[
\hat{U}_j = (X_j X_j^\top)^{-1} X_j G^\top.
\]

As before, we implement a regularized variant by adding \( r_j I \) to each autocovariance term \( X_j X_j^\top \) in Eqs. (8, 9).

3. EXPERIMENTS

We test the proposed features for phonetic recognition on a subset of the University of Wisconsin X-ray Microbeam Database (XRMB) [12] of acoustic and articulatory recordings. The setup is the same as in previous related work [11]. Baseline acoustic features are mean- and variance-normalized 13-dimensional mel-frequency cepstral coefficients (MFCCs) and their 1st and 2nd derivatives. The articulatory data are horizontal and vertical displacements of 8 pellets on the speaker’s lips, tongue, and jaw, yielding a 16-dimensional vector at each sample, downsampled to the MFCC frame rate. We use data from two randomly chosen male speakers (JW11, JW24) and two female speakers (JW13, JW30). The input features to the two bottleneck networks are the acoustic and articulatory features concatenated over a 7-frame window around each frame, giving 273-dimensional acoustic inputs and 112-dimensional articulatory inputs.

We compare baseline and transformed acoustic features in a speaker-dependent setting. The recognizers use 3-state monophone HMM/GMMs with a TIMIT bigram language model (LM); an XRMB LM would be too biased due to XRMB’s limited utterance inventory. We use a five-fold setup
with five separate experiments for each speaker, each using 60% of the data for learning projections and HMM/GMM parameters, 20% for development, and 20% for testing. We report average error rates over the five non-overlapping test sets. For baseline bottleneck experiments, we use a bottleneck network topology of 273–1745–35–1745–39, followed by either a PCA rotation or LDA rotation and optional dimensionality reduction, following the setup of [5]. For multi-view experiments, a bottleneck layer of size 150 is used in order to allow for more freedom in dimensionality reduction experiments. We use rectified linear units and pre-training with de-noising auto-encoders (DAE) and dropout. Following [26], we initialize the weights at the $i$th DAE layer uniformly randomly in the interval $\left[ -\frac{1}{\sqrt{w_{i-1} + w_i}}, \frac{1}{\sqrt{w_{i-1} + w_i}} \right]$, where $w_i$ is the width of the $i$th layer. The learning rate was chosen using the adaptive learning rate method ADADELTA [27]. We used the Deepmat toolkit for learning bottleneck features [28]. We tune hyper-parameters (dimensionalities, regularization parameters $r_x, r_y$, number of Gaussians, LM penalty/scale) independently on each development set. For GCCA, we sub-sample the data to 10,000 frames for improved speed (a possible future improvement is to use incremental matrix factorization with all of the data, as in [11]). As in previous work, performance is better when concatenating the learned features with the baseline MFCCs; all of the given results are therefore with concatenated features. We emphasize that the HMM/GMMs are trained and tested on acoustic data only; the articulatory data is not used after the transformation learning step.

Table 1 shows the phonetic error rates for the four speakers, as well as averages over all speakers. For comparison with earlier unsupervised multi-view work, we include results for CCA applied to raw MFCCs and articulatory features\(^1\). As might be expected, the best supervised baselines improve over unsupervised multi-view techniques. Supervised multi-view techniques yield an additional improvement over supervised single-view techniques. Average improvements over the best supervised baseline (LDA on bottleneck features) range from 0.7% to 2.4% absolute (2-9% relative). While the supervised multi-view methods all perform similarly, the most successful method on average (as well as on most speakers) is many-view GCCA with labels as a third view.

### Table 1

<table>
<thead>
<tr>
<th>METHOD</th>
<th>Speaker</th>
<th>JW11</th>
<th>JW30</th>
<th>JW13</th>
<th>JW24</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>unsupervised baseline</td>
<td>MFCC</td>
<td>32.5</td>
<td>34.5</td>
<td>26.1</td>
<td>30.7</td>
<td>31.0</td>
</tr>
<tr>
<td>unsupervised multi-view</td>
<td>CCA (MFCC, Artic)</td>
<td>31.4</td>
<td>33.0</td>
<td>25.9</td>
<td>29.4</td>
<td>29.9</td>
</tr>
<tr>
<td>supervised baselines</td>
<td>LDA (MFCC)</td>
<td>31.8</td>
<td>31.9</td>
<td>25.6</td>
<td>30.8</td>
<td>30.0</td>
</tr>
<tr>
<td></td>
<td>PCA (BN$_y$)</td>
<td>33.0</td>
<td>29.8</td>
<td>23.4</td>
<td>27.7</td>
<td>28.5</td>
</tr>
<tr>
<td></td>
<td>LDA (BN$_y$)</td>
<td>28.1</td>
<td>29.6</td>
<td>23.8</td>
<td>27.8</td>
<td>27.3</td>
</tr>
<tr>
<td>supervised multi-view</td>
<td>Sec. 2.1.1 CCA (BN$_x$, BN$_y$, Lab)</td>
<td>27.2</td>
<td>29.1</td>
<td>22.5</td>
<td>27.9</td>
<td>26.7</td>
</tr>
<tr>
<td></td>
<td>Sec. 2.1.2 CCA (BN$_x$, BN$_y$ $\oplus$ Lab)</td>
<td>26.9</td>
<td>29.4</td>
<td>22.8</td>
<td>27.4</td>
<td>26.6</td>
</tr>
<tr>
<td></td>
<td>Sec. 2.1.3 CCA (BN$_x$, BN$_y$ $\oplus$ LDA (BN$_x$)</td>
<td>26.5</td>
<td>30.4</td>
<td>23.2</td>
<td>27.9</td>
<td>27.0</td>
</tr>
<tr>
<td></td>
<td>Sec. 2.1.4 LDA (CCA (BN$_x$, BN$_y$))</td>
<td>26.4</td>
<td>29.1</td>
<td>21.9</td>
<td>26.9</td>
<td>26.1</td>
</tr>
<tr>
<td></td>
<td>Sec. 2.1.5 GCCA (BN$_x$, BN$_y$, Lab)</td>
<td>25.7</td>
<td>28.9</td>
<td>22.3</td>
<td>26.9</td>
<td>25.9</td>
</tr>
</tbody>
</table>

1Baseline MFCC and CCA results are improved over our earlier results in [11], due to improved tuning. All systems here are tuned in the same way.

4. CONCLUSION

Our results indicate that acoustic features learned using articulatory measurements and phonetic labels, via supervised multi-view learning, are useful for phonetic recognition. Not surprisingly, the use of labels to learn bottleneck features improves over previous unsupervised results. What is less self-evident is that a second data view (here, articulatory measurements and phonetic labels, via supervised multi-view learning, are useful for phonetic recognition. Not surprisingly, the use of labels to learn bottleneck features improves over previous unsupervised results. What is less self-evident is that a second data view (here, articulatory measurements and phonetic labels, via supervised multi-view learning, are useful for phonetic recognition. Not surprisingly, the use of labels to learn bottleneck features improves over previous unsupervised results. What is less self-evident is that a second data view (here, articulatory measurements and phonetic labels, via supervised multi-view learning, are useful for phonetic recognition. Not surprisingly, the use of labels to learn bottleneck features improves over previous unsupervised results. What is less self-evident is that a second data view (here, articulatory measurements and phonetic labels, via supervised multi-view learning, are useful for phonetic recognition.

Previous work has shown that in the unsupervised multi-view case, nonlinear transformations provide further improvements over linear ones, and the features generalize to new speakers for whom no articulatory data are available even at training time [11]. These are natural extensions for our work here with the supervised case. Other directions for future work include additional supervised multi-view approaches [22, 23, 29], as well as the use of additional views besides articulatory tracks. To our knowledge only video has been used with speech in a similar multi-view setting [30, 31], but any of the other types of measurements being explored for speech (MRI [32, 33], ultrasonic sensors [34, 35], tongue ultrasound [36], muscle EMG [37]) could be good candidates.
5. REFERENCES


