MULTI-VIEW ACOUSTIC FEATURE LEARNING USING ARTICULATORY MEASUREMENTS

Sujeeth Bharadwaj[⋆], *Raman Arora*[†], *Karen Livescu*[†], *Mark Hasegawa-Johnson*[⋆]

[⋆] Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL, USA
[†] Toyota Technological Institute at Chicago, Chicago, Illinois, USA

ABSTRACT

We consider the problem of learning a linear transformation of acoustic feature vectors for phonetic frame classification, in a setting where articulatory measurements are available at training time. We use the acoustic and articulatory data together in a multi-view learning approach, in particular using canonical correlation analysis to learn linear transformations of the acoustic features that are maximally correlated with the articulatory data. We also investigate simple approaches for combining information shared across the acoustic and articulatory views with information that is private to the acoustic view. We apply these methods to phonetic frame classification on data drawn from the University of Wisconsin X-ray Microbeam Database. We find a small but consistent advantage to the multi-view approaches combining shared and private information, compared to the baseline acoustic features or unsupervised dimensionality reduction using principal components analysis.

Index Terms— Multi-view learning, canonical correlation analysis, articulatory measurements, dimensionality reduction, acoustic features

1. INTRODUCTION

The question of whether articulatory information can help in automatic speech recognition has been addressed in a number of ways. It is intuitively appealing to think that some form of articulatory information—using either articulatory measurements, such as tracks of flesh points [1, 2], or knowledge about articulatory processes—should help in recognition. Indeed, it has been shown that phonetic recognition can be improved if articulatory measurements are available as observations at test time [3], and that word recognition may be slightly improved if articulatory measurements are included as observed variables in training and as hidden variables at test time [4]. Knowledge-based approaches, in which the articulatory information is never measured but rather inferred from phonetic labels or otherwise used as hidden variables in the recognition model, have also been used with varying degrees of success [5, 6].

In this work we take a new approach to the use of articulatory measurement data that is available at training time but not at test time. We ask whether it is possible to use the measurement data to learn useful transformations of the acoustic feature vector. This is a natural setting, in that corpora of acoustic and articulatory measurements are available and are collected for many purposes. In general, articulatory data is more feasible to collect at training time than at test time.

We rely on ideas from *multi-view learning*, in which multiple "views" of the data (e.g., from multiple measurement modalities) are available for training but possibly not for prediction at test time [7]. We distinguish this term from *multi-modal* approaches, in which the multiple measurement modalities are available at both training and test time.

A typical approach in speech recognition is to generate a high-dimensional acoustic feature vector by appending multiple 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]. In this work we learn transformations in an unsupervised way, 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 very sensitive to scaling of the data, and possibly of supervised approaches, which are more task-specific.

We propose an approach using canonical correlation analysis (CCA), which finds pairs of maximally correlated projections of data in two views [9, 10]. In our 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.

One challenge is that not everything that is uncorrelated is noise: The acoustic view may contain discriminative information that is not correlated with the articulatory view. In this case, we would like to combine the projections learned with CCA ("shared" information) with additional projections that

This research was supported by NSF grant IIS-0905633. The opinions expressed in this work are those of the authors and do not necessarily reflect the views of the funding agency.

are "private" to the acoustic view. We present such combined approaches in the following section.

2. METHODS

We begin with a training data set of N paired vectors $\{(x_i, y_i)\}_{i=1}^N = \{(x_1, y_1), ..., (x_N, y_N)\}$, where $x_i \in \mathbb{R}^{d_1}$, $y_i \in \mathbb{R}^{d_2}$, and d_1 and d_2 are the dimensionalities of the feature vectors in the two views. Let X and Y be the corresponding matrices of training data, i.e. the matrices whose i^{th} columns correspond to x_i and y_i, respectively. In our case, let X be the acoustic training set and Y the articulatory training set. Each pair (x_i, y_i) corresponds to one frame of simultaneously recorded acoustics and articulation.

In this work we consider the task of framewise phonetic classification. We make the assumption that the two views are uncorrelated conditioned on the phonetic class. When this assumption holds, any dimensions that are correlated must relate to the hidden class. To the extent that this assumption holds, then, the learned dimensions will be discriminative for phonetic classification.

2.1. Canonical correlation analysis

Canonical correlation analysis (CCA) [9, 10] finds pairs of directions $v_k, w_k, 1 \leq k \leq \min(d_1, d_2)$ such that the projections of X and Y onto those directions, respectively—the *canonical variables* $v_k^T X$ and $w_k^T Y$—are maximally correlated. The first pair of directions is given by

$$\begin{aligned} \{v_1, w_1\} &= \underset{v,w}{\arg\max}\ corr(v^T X, w^T Y) \\ &= \underset{v,w}{\arg\max}\ \frac{v^T C_{xy} w}{\sqrt{v^T C_{xx} v w^T C_{yy} w}} \end{aligned}$$

where C_{xy} is the cross-covariance matrix between X and Y and C_{xx}, C_{yy} are the auto-covariance matrices. Subsequent direction vectors $\{v_k, w_k\}, k > 1$, maximize the same correlation, subject to the constraint that the resulting projected variables $v_k^T X, w_k^T Y$ are also uncorrelated with all previous ones, $\{v_j^T X, w_j^T Y |\ j < k\}$.

The canonical directions are found as the solution of an eigenvalue problem:

$$\begin{aligned} C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx} v &= \lambda^2 v \\ w &\propto C_{yy}^{-1} C_{yx} v \end{aligned}$$

where the values of λ are the correlations between the projections. We reduce dimensionality by projecting X along the top M eigenvectors.

Unlike PCA, CCA relies on correlation between the projected variables (statistical orthogonality) rather than orthogonality of the direction vectors, and is affine-invariant. This property helps us to avoid the key disadvantage of PCA, which is sensitive to affine transformations of the coordinates. LDA is a special case of CCA where one of the views is the labels represented as a binary matrix of indicator vectors.

CCA is typically regularized by replacing C_{xx} with $C_{xx} + r_x I$ and C_{yy} with $C_{yy} + r_y I$, where I denotes an identity matrix [11]. This ensures that the matrices are invertible and avoids spurious correlations in the data among low-variance input dimensions. The parameters r_x and r_y are tuned on held-out data.

Our assumption of uncorrelatedness given the phone class may not be satisfied. For example, the audio and articulation may be correlated through the speaker identity or emotional state. In this work we restrict ourselves to speaker-dependent experiments—that is, X and Y are data from a single speaker—which partially avoids this problem. This issue, however, requires further study.

Note that CCA provides two projections, one for each view. In our case, we are interested in improving performance on a prediction task that uses acoustic data, so we retain only the projections of the acoustic feature vector. However, the approach can in principle be applied with either or both views available at test time.

2.2. Shared-private representations

CCA finds only those dimensions that are correlated across the views, which we refer to as "shared" information. However, there may be additional discriminative information in the acoustics that is not correlated with the articulatory measurements, which we call "private" information. For example, in our case the articulatory data does not include glottal or velar measurements. Therefore, the acoustic features presumably contain "private" information about voicing and nasality.

In previous work [12], shared and private information were combined by appending the CCA features to baseline MFCC features, which we refer to as MFCCA (for MFCC+CCA). We also explore a different approach that constrains the private dimensions to be non-redundant with the shared ones. The procedure is as follows:

$(V, W) = \text{CCA}(X, Y)$, i.e. use CCA to find the projections $\{v_k\}_{k=1}^M, \{w_k\}_{k=1}^M$ and let V and W be matrices in which the k^{th} column vectors are v_k and w_k, respectively. W is not used from this point on.

$P = \text{PCA}\left((V^\perp)^T X\right)$, i.e. apply PCA to the orthogonal complement of the acoustic subspace defined by V to find projections $\{p_j\}_{j=1}^L$ and let P be a matrix in which the j^{th} column vector is p_j.

$D = [V\ P]$, i.e. form the final feature transformation D by concatenating the CCA and PCA directions.

This is almost identical to the "non-consolidating components analysis" (NCCA) of [13] (up to a difference in regularization) and we refer to it as NCCA henceforth.

After learning a transformation D, all of the acoustic feature vectors (both training and testing) are projected along the vectors in D, forming the new acoustic data $D^T X$. In the case of CCA, $D = V$; in MFCCA, $D = [V\ I]$; and in NCCA, $D = [V\ P]$.

3. RELATED WORK

CCA has rarely been used for speech tasks. In [14], CCA was used to reduce dimensionality of acoustic features for improved clustering into speakers. In [12], it was used to learn transformations of acoustic features for improved speaker recognition in noise. In [15], it was used for speaker normalization, by transforming the acoustics of different speakers so as to be maximally correlated. It has also been used in audio-visual synchronization and speaker recognition [16, 17] where both views are available at test time. In [18], kernel (nonlinear) CCA was used for acoustic-articulatory inversion. We are unaware of any prior work on a speech classification task in which CCA was used to learn an acoustic transform using articulatory training data.

4. EXPERIMENTS

We address two questions in the context of phonetic frame classification: (1) Can we learn useful transformations of the acoustic data using articulatory data for training only? (2) Is it necessary or helpful to combine shared and private dimensions? In both cases, our results provide affirmative answers.

We use a subset of the University of Wisconsin X-ray Microbeam Database (XRMB), which includes simultaneous recordings of acoustic waveforms and articulatory measurements for a number of tasks and speakers [2]. The articulatory data consist of horizontal and vertical displacements of eight pellets on the speaker's lips, tongue, and jaws, relative to reference pellets defining a speaker-specific coordinate system, yielding a 16-dimensional vector at each time point. Our experiments are speaker-dependent, using the two XRMB speakers JW11 (male) and JW30 (female). The coordinate systems can vary drastically between speakers; normalizing for this is a challenge that we defer to future work.

For each utterance, we compute 13 mel-frequency cepstral coefficients (MFCCs) and their first and second derivatives every 10ms with a 25ms window. We downsample the articulatory data to synchronize with the acoustics and discard any frames that have missing measurements. Finally, for each frame we concatenate acoustic features over a window of three frames and articulatory features over a window of seven frames. This results in the data $X \in \mathbb{R}^{117\times N}$, $Y \in \mathbb{R}^{112\times N}$, where X is the acoustic data, Y is the articulatory data, and N is the number of frames. In our case N is about $50,000$ for each speaker.

We consider two types of classifiers: support vector machines (SVMs) using radial basis function kernels with a one-against-one multi-class implementation [19] and k-nearest neighbors (kNN) with a correlation distance $d(x, y) = 1 - corr(x, y)$. We compare the performance of these classifiers on the raw MFCCs (baseline) and on MFCCs transformed with PCA, CCA, NCCA, and MFCCA. The hyperparameters to be tuned are the number of neighbors k in kNN, kernel width and cost in SVMs, PCA dimensionality L, CCA dimensionality M, and CCA regularization parameters r_x and r_y. We use a five-fold cross-validation setup: In each fold, 60% of the utterances are used for training, 20% for tuning (development), and 20% for final testing.

We obtain phone labels for the XRMB corpus using the Penn Phonetics Lab Forced Aligner [20]. The alignments are imperfect, but anecdotally very good. Short pauses and stress are removed, leaving 39 phone classes.

Figure 1 shows the dependence of error rate on the CCA regularization parameters and Figure 2 shows the dependence

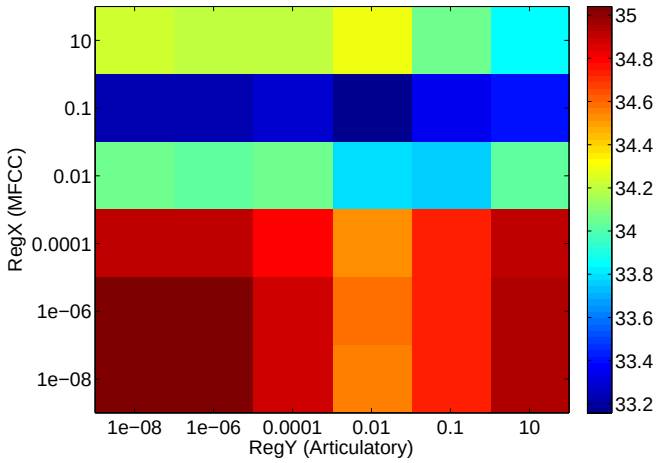

Fig. 1. Dependence of error rate on CCA regularization for speaker $JW30$, using an SVM classifier with CCA-transformed features.

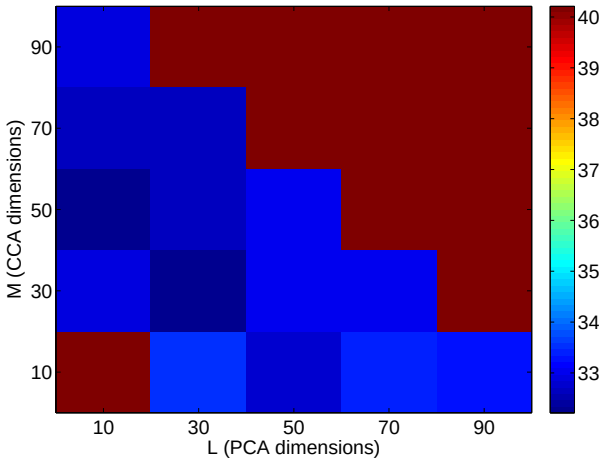

Fig. 2. Dependence of error rate on NCCA dimensionalities for speaker $JW30$, using an SVM classifier with NCCA-transformed features.

Table 1. Best (CCA, PCA) dimensionalities for MFCCA and NCCA

features	MFCCA (M)		NCCA (M, L)	
speaker	JW11	JW30	JW11	JW30
kNN	50	110	$(50, 10)$	$(70, 10)$
SVM	30	30	$(50, 10)$	$(30, 30)$

Table 2. Error rates averaged over five folds for speakers JW11, JW30. Boldface font indicates the best performance in each column. An asterisk indicates a significant improvement over the MFCC baseline (according to a t-test at $p = 0.05$).

speaker	JW11		JW30	
classifer	kNN	SVM	kNN	SVM
MFCC	30.96	26.78	36.63	32.42
PCA	30.76	28.26	35.88	33.49
CCA	30.81	27.99	35.37	33.28
MFCCA	$\mathbf{29.01}^*$	$\mathbf{26.15}^*$	$\mathbf{34.42}^*$	$\mathbf{31.74}^*$
NCCA	29.43^*	26.64	34.85^*	32.26

on the NCCA dimensionalities for speaker JW30 and an SVM classifier. Tuning over all four NCCA hyperparameters as independent variables is computationally intensive; we therefore assume that the best regularization parameters of NCCA are similar to the best regularization parameters of CCA.

Figure 1 shows that performance is very sensitive to the CCA regularization in the acoustic view r_x, but insensitive to the articulatory regularization r_y; this is sensible, since the acoustic view is the noisier one [11]. Figure 2 shows that performance tends to depend more on the sum of the CCA and PCA dimensionalities than on each alone. However, the dependence on hyperparameters is speaker- and classifier-dependent and, to a lesser extent, fold-dependent. The best values of k tend to be in $[8, 12]$, and of the final dimensionality in $[30, 70]$ (with varying divisions between PCA and CCA dimensionalities). Table 1 shows the best CCA (M) and PCA (L) dimensionalities that achieve the lowest misclassification rate on the development set for MFCCA and NCCA.

Table 2 shows the test set error rates averaged over the folds for each experiment. In all cases, one or both of NCCA and MFCCA significantly improve over the baseline and the other techniques. This suggests that CCA helps to clean up those parts of the acoustic signal that have articulatory correlates. However, CCA alone is not sufficient; the articulatory view is missing crucial information (such as voicing and nasality) that is important for phonetic classification. Figure 3 gives a more detailed view of the NCCA and MFCCA results, showing the spread over the five folds.

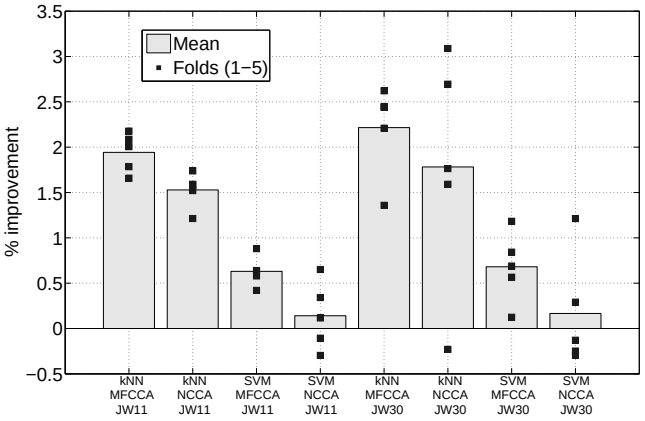

Fig. 3. Improvement (in %) of NCCA/MFCCA over baseline

5. CONCLUSION

We have shown the potential benefit of multi-view learning of acoustic feature transformations using CCA when articulatory measurement data is available at training time. In our experimental setting, the CCA features alone are not sufficient, but in combination with additional "private" feature dimensions—either the baseline features or a subspace orthogonal to the CCA features—they improve over the baseline.

These experiments have been limited to linear transformations and unsupervised learning. Future work includes nonlinear extensions [10, 13], supervised and semi-supervised extensions, application to noise robustness (as in [14, 12]) and domain-independence. In the supervised case, the labels could be considered to be an additional view, or they can be incorporated via additional terms in the objective function to be optimized. A potentially more interesting setting is the semi-supervised case, where labels are available for only a subset of the data, or where some labels are more reliable than others (as in our case, where the ground truth comes from an automatic alignment). In the long run, the practicality of such multi-view techniques will be much greater if they can be shown to extend beyond specific domains for which the views are available. Multi-view methods should be less dependent than supervised methods on a specific task or data set; for example, finding acoustic dimensions that are predictive of articulatory dimensions could be equally useful for phonetic classification, word recognition, or speaker and language identification. An interesting area for future work, therefore, is the study of the domain- and task-independence of features learned with multi-view techniques.

6. REFERENCES

[1] A. Wrench, "A new resource for production modeling in speech technology," in *Workshop on Innovations in Speech Processing*, 2001.

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

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

[4] 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.

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

[6] 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.

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

[10] 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.

[11] T. De Bie and B. De Moor, "On the regularization of canonical correlation analysis," in *ICA*, 2003.

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

[13] C. H. Ek *et al.*, "Ambiguity modelling in latent spaces," in *MLMI*, 2008.

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

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

[16] M. E. Sargin *et al.*, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE. Trans. Multimedia*, vol. 9, no. 7, pp. 1396–1403, 2007.

[17] M. Liu *et al.*, "Audio-visual fusion framework with joint dimensionality reduction," in *ICASSP*, 2008.

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

[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, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

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