Multi-View Learning of Acoustic Features for Speaker Recognition

Karen Livescu [1], Mark Stoehr [2]

[1] *TTI-Chicago,* [2] *University of Chicago*
Chicago, IL 60637, USA
[1] klivescu@uchicago.edu, [2] stoehr@uchicago.edu

***Abstract*—We consider learning acoustic feature transformations using an additional view of the data, in this case video of the speaker's face. Specifically, we consider a scenario in which clean audio and video is available at training time, while at test time only noisy audio is available. We use canonical correlation analysis (CCA) to learn linear projections of the acoustic observations that have maximum correlation with the video frames. We provide an initial demonstration of the approach on a speaker recognition task using data from the VidTIMIT corpus. The projected features, in combination with baseline MFCCs, outperform the baseline recognizer in noisy conditions. The techniques we present are quite general, although here we apply them to the case of a specific speaker recognition task. This is the first work of which we are aware in which multiple views are used to learn an acoustic feature projection at training time, while using only the acoustics at test time.**

I. Introduction

The extraction of acoustic features useful for a given task – automatic speech recognition, speaker recognition, and so on – has received a great deal of attention in speech technology research. Techniques such as principal components analysis (PCA) and linear discriminant analysis (LDA) [1], and their variants, are popular and effective in many settings. However, they have drawbacks: For example, PCA is highly sensitive to the scaling of the data, making it unable to distinguish between signal and noise. LDA and other discriminative transforms, on the other hand, are much more effective for finding the important dimensions for the task at hand, but they rely on labeled data for estimating the transform.

In this paper, we consider an unsupervised approach to learning an acoustic feature transform. Rather than labels, we assume that we instead have access to a second "view" of the data at training time (but not necessarily at test time). This is often a natural assumption, as we may be able to collect a great deal of multi-view (e.g., audio-visual) data, while not necessarily having access to all of their labels nor having both views at test time. We differentiate this approach from *multimodal* approaches, in which multiple views are available at both training and test time. In particular, we focus in this paper on the problem of speaker recognition, with audio and video available at training time and only audio available at test time.

Why might a second view help in estimating a discriminative transform? This is a question that has been addressed thoroughly in the area of multi-view learning. Multi-view learning assumes that we have multiple (usually two) "views" of the data, and the goal is to use the relationship between these views to alleviate the difficulty of a learning problem of interest [2], [3], [4]. The definition of "views" may be quite natural, such as audio and video recordings of speech, or images and associated captions; or they may be quite abstract, such as random divisions of a feature vector [5]. In this work, we consider how having two views contributes to the speaker classification problem. Specifically, we consider the problem of learning a linear projection of the acoustic data. We explore the use of canonical correlation analysis (CCA) [6], [7] as a dimensionality reduction technique.

In many multi-view scenarios, we can assume that sources of noise in each modality do not affect the other modality. For example, in speaker classification, the visual noise may include lighting and pose variation; the corresponding audio is likely to be unaffected by these, but will be affected by independent sources such as the background acoustic noise. When this assumption holds, the information that appears in both views is likely to be related to the semantic content in the data (e.g. the speaker identity) and not to the noise. This provides some intuition for the multi-view approach. Figure 1 shows a graphical model that represents this assumption.

CCA looks for information that appears in both views by finding those linear projections of each view that are most correlated with the corresponding projections of the other view. Using CCA for dimensionality reduction, we only retain the correlated information between the two views, which hopefully captures the information about the class identity while reducing the noise. Some multi-view learning approaches make stronger assumptions than those of Figure 1; for example, co-training [2] makes the additional assumption that each view is "sufficient" for classification, a strong assumption that may not hold in practice. In addition, co-training simultaneously learns two *classifiers*, one for each view. Here we learn only a feature transform, and are free to use any classifier on the resulting features.

II. Learning feature transforms with canonical correlation analysis

Given a data set of paired vectors $\{(x_1, y_1), ..., (x_n, y_n)\}, X = \{x_i\}, Y = \{y_i\}$, CCA [6], [7] finds pairs of directions $v_k, w_k, 1 \leq k \leq M$ such that the

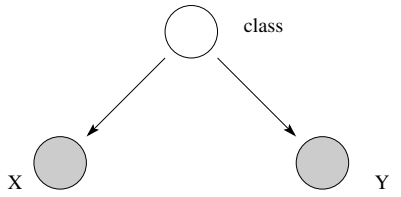

Fig. 1. A graphical model representing a two-view setting in which the two (observed) views X and Y are independent given the (hidden) class of interest.

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

$$\{v_1, w_1\} = \arg\max_{v,w} \; corr(v^T X, w^T Y) \quad (1)$$

$$= \arg\max_{v,w} \; \frac{v^T C_{xy} w}{\sqrt{v^T C_{xx} v w^T C_{yy} w}} \quad (2)$$

where C_{xy} is the cross-covariance matrix between X and Y (i.e., the (i,j) entry of C_{xy} is $cov(x_i, y_j)$) and C_{xx}, C_{yy} are the auto-covariance matrices of X and Y. 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, $\{v_j^T X, w_j^T Y | j < k\}$.

It is straightforward to show [7] that the canonical directions can be found as the solution of an eigenvalue problem. In particular, the v_k are eigenvectors of $C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx}$ and the w_k are eigenvectors of $C_{yy}^{-1} C_{yx} C_{xx}^{-1} C_{xy}$. Only one of the eigenvector problems needs be solved: Given v_k, $w_k = C_{yy}^{-1} C_{yx} v_k$. Therefore, the problem we solve is

$$C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx} v = \lambda^2 v \quad (3)$$

$$w \propto C_{yy}^{-1} C_{yx} v \quad (4)$$

where the top eigenvectors v_1, w_1 corresponding to the largest λ are the most highly correlated ones across the views, and the values of λ are the correlations between the projections. To reduce dimensionality, we keep the top eigenvectors corresponding to the most correlated projections.

Because of its reliance on correlation, rather than orthogonality of the direction vectors, CCA is affine-invariant (unlike, for example, principal components analysis). It can be shown that under the multi-view assumption, we are able to (approximately) find the low-dimensional subspace spanned by the means of the classes in each view [8]. This subspace is important, because, when the data is projected onto this subspace, the means of the classes are well-separated, yet the typical distance between points from the same distribution is smaller than in the original space.

In practice, we also add a regularizing term of $\gamma_x I$ to C_{xx} and $\gamma_y I$ to C_{yy} (where γ_x, γ_y are tuned on held-out data), as done in prior work [3]. The regularization ensures that the matrices are invertible, as well as smoothing out some of the spurious correlations in the data (i.e., directions that appear correlated in the data due to chance variation in the sample rather than due to the class identity).

In addition, it is clear that there may be multiple hidden variables, other than the class of interest, that may account for correlations between the two views. In our case, the views are audio and corresponding face video of speakers, and the hidden variables may include the (desired) speaker identity as well as the (undesired) phonetic state, emotional state, and so on. In our experiments, we alleviate this problem by randomizing the vectors in one of the views for each speaker, so that the only consistent connection between the views is (hopefully) the speaker identity. This issue, however, requires further study.

We think of each view as providing a sample of the same class, plus (high-dimensional) additive noise in each view. We retain only the top M directions, thus using CCA as a dimensionality reduction. It is easy to show that in the resulting subspace found by CCA, the noise covariance is reduced relative to the signal covariance. As mentioned previously, we assume that the two views are independent given the hidden class variables; if the noise is also independent in the two views, then the correlated dimensions must correspond to some aspect of the hidden class.[1]

Figures 2 and 3 motivate the usefulness of projections learned using CCA, using a (very simplistic) simulated example. In each view, there is clearly a single "good" dimension along which classification should be done. It would be difficult to find this direction given one (unlabeled) view alone. PCA would of course find the direction orthogonal to the desired one.[2] If we were to train a typical speaker recognition system using diagonal Gaussians, this would also be a poor fit to the data. However, the two views are correlated given the class, in such a way that the dimension that is correlated across views is also the correct dimension for classification in each of the views. Figure 3 shows the result of performing CCA on the simulated data and projecting to the first dimension. The projected data is now easy to classify using a single one-dimensional Gaussian in either view.

Note that CCA, like many other multi-view learning methods, provides two projections, one for each view, and is agnostic as to which view is used at test time. In our case, we are interested in improving the performance of a classifier using acoustic data. However, we could just as well use this approach to improve classification in the other (visual) view.

CCA has been used in previous work on audio-visual synchronization and speaker recognition [9], [10], but to our knowledge, only in the context of multi-modal tasks where both views are available at test time. CCA has also been

[1] In fact, we are slightly abusing the term "independent" as it is intuitive to think about the dependence or independence of the views; however, we only assume that the views are uncorrelated given the class.

[2] Clearly we could "fix" this example to improve the behavior of PCA, but it is easy to extend this to more challenging cases.

applied to speaker clustering using both audio and video for projection learning and only one view for clustering [8]; here we base our experimental setup on this clustering work.

III. Experiments

We use 41 speakers from the VidTIMIT database [11], speaking 10 sentences (about 20 seconds) each, recorded at 25 frames per second in a studio environment with no significant lighting or pose variation. The sentences are drawn from the TIMIT database [12]. The task is speaker identification, i.e. a 41-way classification task. We use a standard mixture-of-Gaussians approach [13]: We train a mixture of diagonal Gaussians for each speaker, and at test time we hypothesize the speaker whose model has the highest likelihood on the current utterance, where the utterance likelihood is taken to be the product of the frame likelihoods.

The baseline audio features are 12-dimensional mel frequency cepstral coefficients (MFCCs) and their derivatives. We also extract a larger feature vector, which we then project using CCA. This larger vector consists of MFCCs and their derivatives and double derivatives, computed every 10ms over a 20ms window, and finally concatenated over a window of 440ms centered on the current frame (i.e. corresponding to a total of 11 video frames), for a total of 1584 dimensions. Note that it may seem that the CCA-based approach is given a unfair advantage, as it uses a larger number of raw features. However, the baseline performance is not improved by simply adding more of these raw features without the CCA projection step. The video features are pixels of the face region extracted from each image (2394 dimensions).

We use a 5-fold cross-validation scheme. For each speaker, 6 sentences are used for training, 2 for tuning, and 2 for final testing, for a total of 82 utterances for development and 82 for testing in each fold. The five folds use disjoint development and test sets. For each fold, we find the parameters that produce the best performance on the development set. In these experiments, the tuning parameters are the number of Gaussians in each mixture, the dimensionality of CCA projection, and the two CCA regularization parameters γ_x, γ_y. For final testing, we re-train on the combined training and development sets for each fold, using the best parameters found above, and use the resulting models for final testing.

For each fold, we learn a CCA projection of the training data. We randomize the vectors of one view for each speaker, to reduce correlations between the views due to other latent variables such as the current phoneme. We find that the CCA features alone do not outperform the baseline MFCC-based approach (see the Discussion section below). Instead, we append the CCA features to the baseline MFCCs and use the combined vectors for speaker recognition.

We learn the CCA projections using *clean* audio data, while the speaker recognition is done using noisy data, with white noise added at 0dB or -10dB. This is intended to simulate a natural scenario in which cooperative speakers provide training data in a controlled environment, whereas the system may be deployed in much noisier environments. This setup is still not entirely natural, of course; see Section IV for discussion of more natural extensions.

Figure 4 shows the results of our experiments for clean speech and noisy speech at 0dB and -10dB. For clean speech, the performance of the baseline and CCA-based features is the same (the difference between them is statistically insignificant according to a t-test). For the 0dB and -10dB cases, there is a modest but statistically significant improvement (according to a t-test; p-value $= .04$ for the 0dB case and .005 for the -10dB case). The best parameter values differ somewhat across folds; the chosen number of Gaussians per speaker is typically between 3 and 9, the CCA dimensionality is usually 5 or 10 (for a total dimensionality of 29 or 34), and the CCA regularization parameters are usually 10 (note that this may depend on the variance in the data). Again, note that the CCA-based approach does not have an advantage due to the higher dimensionality; the baseline does not improve when its dimensionality is increased (e.g., by adding double derivatives of the MFCCs). For completeness, we note that the performance of the visual classifier is extremely good, with typically less than 5% error rate, and does not gain from the use of CCA-based features learned using the audio. This is to be expected, as the visual data is very clean.

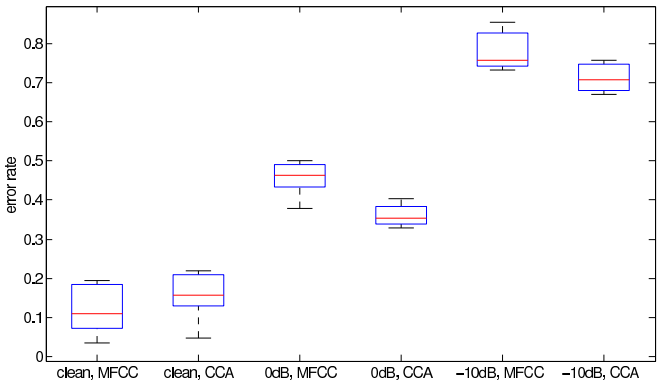

Fig. 4. Box plots of speaker recognition error rates over the five folds for clean and noisy speech using baseline MFCC features or CCA-based features appended to MFCCs.

IV. Discussion

Our experiments show that a multi-view learning approach using CCA to extract features from speech, with video as the other view, can improve the performance of a speaker recognition system. In particular, under the assumption that clean audio and video data is available to learn the CCA projections, we find modest but statistically significant improvements for speaker recognition on VidTIMIT data in additive noise at 0dB and -10dB. This is the first work of which we are aware in which an unsupervised feature transformation is learned

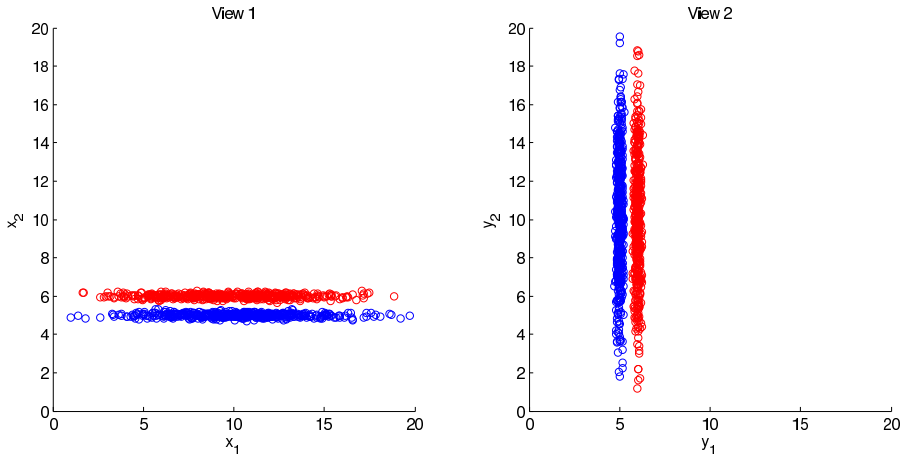

Fig. 2. Scatter plots of simulated two-view data, with two dimensions in each view. Red and blue points correspond to different classes, e.g. speakers.

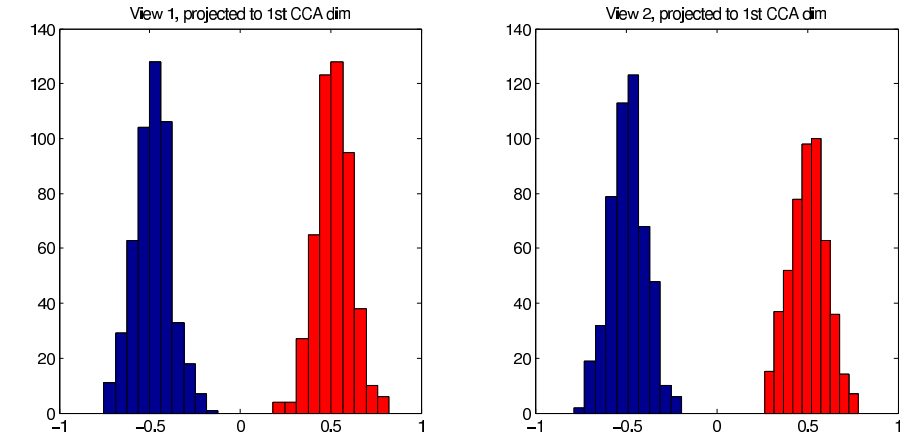

Fig. 3. Histogram of each view in Figure 2 projected onto the first CCA dimension.

from multiple views of speech data, while only the audio is available at test time (unlike multi-modal approaches, in which it is assumed that the multiple views are available at test time).

There are some clear extensions to this work. First, the improvements we have seen are not large, and it is somewhat unsatisfying that the CCA-based features alone do not improve over MFCCs. One reason may be that there are aspects of the audio that are relevant to speaker recognition but uncorrelated with the video, for example information about the rear of the vocal tract. CCA is only effective to the extent that the correlated directions in the two views are informative about the task. A natural extension, besides simply appending the raw and CCA features, would be a more principled approach to look for just that acoustic information that is not included in the CCA features.

Another limitation of our setup is the assumption of a linear relationship between the audio and video. In this work, we are essentially estimating each view from the other using a linear mapping. This is unlikely to be a good assumption, and we are currently exploring non-linear extensions such as kernel CCA [7].

The setup in these experiments is, of course, somewhat contrived and more experiments are needed for a fuller comparison against approaches for noise robustness. We have made the natural assumption that clean data is available at training time but not at test time. However, we are using an unsupervised learning approach, but using the same (labeled) data for both the projection learning and the model training. A more natural scenario is one where the projection may be learned on arbitrary data, collected not necessarily from the same speakers we will eventually test on, and the labeled data for model training may be a separate (perhaps smaller) set. The key to this approach is indeed that it is unsupervised: While for our data we may have been able to learn a better transform using a supervised approach, the multi-view approach allows us to use a potentially larger, unlabeled data set. We therefore would like to extend our work to larger data sets that allow such experiments. For this initial work, we have chosen the

small VidTIMIT set because of its clean video data.

Another natural setting is one in which some labeled data is used in addition to the unlabeled data; this suggests extending the approach to one combining unsupervised transforms learned with CCA with discriminative transforms using labels (e.g., as in [14]).

REFERENCES

[1] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Int. Conf. on Acoustics, Speech, and Signal Processing*, 1992.

[2] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Conf. on Learning Theory*, 1998.

[3] S. M. Kakade and D. P. Foster, "Multi-view regression via canonical correlation analysis." in *Conf. on Learning Theory*, 2007.

[4] R. K. Ando and T. Zhang, "Two-view feature generation model for semi-supervised learning," in *Int. Conf. on Machine Learning*, 2007, pp. 25–32.

[5] K. Nigam and R. Ghani, "Analyzing the effectiveness and applicability of co-training," in *Conf. on Information and Knowledge Management*, 2000.

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

[7] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[8] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Int. Conf. on Machine Learning*, 2009.

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

[10] M. Liu, Y. Fu, and T. S. Huang, "Audio-visual fusion framework with joint dimensionality reduction," in *Int. Conf. on Acoustics, Speech, and Signal Processing*, 2008.

[11] C. Sanderson, *Biometric Person Recognition: Face, Speech and Fusion*. VDM-Verlag, 2008.

[12] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "TIMIT acoustic-phonetic continuous speech corpus," 1993, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1.

[13] D. A. Reynolds and R. C. Rose, "Text-independent speaker identification using Gaussian mixture models," *IEEE Trans. Speech and Audio Proc.*, vol. 3, no. 1, pp. 72–83, 1995.

[14] T.-K. Kim, J. Kittler, and R. Cipolla, "Learning discriminative canonical correlations for object recognition with image sets," in *Eur. Conf. on Comp. Vision*, 2006.