MULTIVIEW FEATURE LEARNING FOR SPEECH RECOGNITION

BY

SUJEETH SUBRAMANYA BHARADWAJ

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Adviser:

Associate Professor Mark A. Hasegawa-Johnson

ABSTRACT

In this thesis, we study the problem of learning a linear transformation of acoustic feature vectors for speech recognition, in a framework where apart from the acoustics, additional views are available at training time. We consider a multiview learning approach based on canonical correlation analysis to learn linear transformations of the acoustic features that are maximally correlated with the data. We propose simple approaches for combining information shared across the views with information that is private to the acoustic view. We apply these methods to a specific scenario in which articulatory data is available at training time. Results of phonetic frame classification on data drawn from the University of Wisconsin X-ray Microbeam Database indicate a small but consistent advantage to the multiview approaches that combine shared and private information, compared to the baseline acoustic features or unsupervised dimensionality reduction using principal component analysis. We then discuss limitations of canonical correlation analysis and possible extensions.

To Thatha, Late. Sri. H. Ramabhat, for the early introduction to mathematics

To Ajji, Smt. Kanthamma, for instilling in me faith, devotion, and the drive to succeed

To Appa, Amma, and Shruthi, for the infinite love, support, and guidance

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my adviser, Professor Mark Hasegawa-Johnson, without whom none of this would have been possible. He has always been supportive of my endeavors, complimenting me for the successes and consoling me for the failures. Thank you so much.

I am equally indebted to Professor Karen Livescu of Toyota Technological Institute at Chicago (TTIC), where much of the work presented in this thesis was completed. I am thankful for her cheerful nature, because of which I could truly learn from my mistakes; I made many last summer.

Several other TTIC members have been extremely helpful. I would like to thank Raman Arora for restructuring my code; thanks to him, I can now call myself a “healthy” coder. I am grateful to Professor Raquel Urtasun and Mathieu Salzmann for their guidance and suggestions throughout the semester, without which my study of factorized latent spaces would have been incomplete. Thanks also to Professor Nati Srebro and Joseph Keshet for useful tips and tricks essential for effective stochastic gradient descent implementation.

I am thankful to the National Science Foundation (NSF) and TTIC for the financial support; this research was funded by NSF grant IIS-0905633.

I am grateful to Prof. Livescu for thoroughly editing and rewriting parts of my ICASSP submission, the contents of which form a large portion of this thesis. I am also thankful to the ECE Publications Office and Prof. Hasegawa-Johnson for their help in editing this document.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The speech processing community has witnessed several advances over the last few decades. Sophisticated statistical models such as the hidden Markov model (HMM) and deep belief networks (DBNs) have been proposed to create an accurate internal representation of the non-stationary process that is so fundamental to humans; however, the problem of feature selection has received little attention – few researchers have looked past mel-frequency cepstral coefficients (MFCC) and have attempted to construct optimal feature spaces for classification. One could argue that this is only natural, for why try to improve upon a system that already has recognition accuracies close to 95%? These elevated figures are achieved by systems restricted to the machine learning analog of STP conditions – low noise and trained to a particular speaker. Although intriguing and significant in its own right, it forms a very small subset of a much larger group of unresolved classification problems. A particularly interesting extension is the integration of multiple views (for example, acoustical and articulatory features) of the same semantic object; the most naive approach would be to simply concatenate the feature vectors, but such a procedure requires that we have access to all of the views at both training and test time. In this thesis, we present the general multiview framework, statistical methods such as canonical correlation analysis (CCA), and their applications to feature selection when articulatory information is available at training time, but not at test time.

1.1 Multiview Learning

Commercially available speech recognizers depend solely on recorded speech; however, it is intuitively obvious that humans benefit from additional views of the data. Introspection is arguably the greatest avenue for accurate psychological analysis. What exactly are we doing when we talk to somebody? We hear the voice, see the face, and keep account of significant emotions and gestures. If we are later asked to recognize the speaker in a video, we first try to identify the face. But suppose the image is corrupted; we would next try to identify the speaker based on his/her voice. But let us assume that the audio is also noisy; we are left with no choice but to rely on the emotional gestures that are unique to a particular person. One can envision many other situations where multiple modalities are of significant value; for example, we tend to track lip movement while conversing with a foreign national, whose accent typically differs from our own.

Levinson's group [1] has taken the first step by constructing an associative memory, wherein a separate HMM is trained for each part of the sensory system – audio, visual, tactile, etc. The maximum likelihood state sequences of these sensory HMMs are fed into another HMM to learn higher level dependencies; their approach appears to be working remarkably well as demonstrated by their robots [1]. Similarly, multimodal speech recognition systems have been described for audiovisual [2] and audio-articulatory [3] settings. Levinson and others assume that instances of all of the different views are available at both training and test times; in this work, we provide a framework which relaxes this assumption to a more general scenario in which only a subset of the views are available at test time.

In multimodal settings, such as the one described by Levinson, all views are expected to be available at test time; in some cases, taking measurements from every modality is either too expensive or impossible. Multimodal recognition of simultaneous acoustic and articulatory measurements as recorded in the X-ray Microbeam (XRMB) database would require speakers to have pellets every time they speak; a multimodal approach is therefore useless and we look towards methods that can both harness additional information available at training time and rely solely on the audio at test time. Multiview learning is in fact more general than what we have described above; even if only one view is available, an arbitrary partitioning of the extracted features into two sets might be beneficial. We will explore the paradigm in its most general form for the specific problem of feature learning.

1.2 Feature Learning

A 30 millisecond speech signal sampled at 44 kHz has 1300 samples; it is clearly infeasible to train a classifier on hundreds of samples from such a high dimensional space – similarly, inference is also difficult. Based on a physical model of the auditory system, scientists in the first half of the twentieth century concluded that a 12-dimensional space is sufficient to represent all of the important features of speech, obtained via linear predictive coding (LPC) [1]. Current feature selection methods seek to automate this process under some notion of optimality [4].

Feature selection and dimensionality reduction have been problems of great interest to researchers in computer vision and machine learning. Feature selection is broadly defined as the process of selecting a (small) subset of features sufficient to predict the target class. A more general framework is that of dimensionality reduction, wherein we learn a few functions that map a high dimensional feature space into one that is of much lower dimensionality. Most learning methods perform well (after tuning) if we have an appropriate representation of the data and tend to fail when we have poorer representations [4]. The consensus within the machine learning community is that a good representation; i.e., one that leads to impressive classification results, is one that is both compact and meaningful [4].

Feature selection methods offer several advantages. First and foremost is dimensionality reduction: feature vector dimension is generally proportional to the computation time necessary in both the training and testing stages of a model. Feature selection can also increase the classification accuracy by enhancing the signal-to-noise ratio (SNR) in certain situations [4, 5, 6]. Another benefit lies in realms where the physical properties of the system are not very well known. Feature selection methods can still find key features that are optimal for classification; moreover, these features provide some insight into the nature of the system itself. Akin to most other processes replicated on a computer, feature selection also occurs in biological systems [4]. Any sensory system is exposed to infinitely many possible features, but chooses only a handful of them for the specific task of interest. Different hierarchies within the visual cortex, for example, select a very small subset of the features presented by the lower levels. Within the olfactory system, neurons detect odor molecules and send significant features to the olfactory bulb, which typically represents one particular set of smells [4]. The human auditory system is yet another example of sophisticated feature selection [1, 4]. Please refer to [1] for a succinct overview of the auditory model.

1.3 Articulatory Information

Articulatory information has been known to boost the performance of automatic speech recognizers in several settings. It is intuitive that some form of articulatory information – using either articulatory measurements, such as tracks of flesh points [7, 8], or knowledge about articulatory processes – should help in recognition. Indeed, it has been shown, for example, that phonetic recognition can be improved if articulatory measurements are available as observations at test time [9], 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 [10]. 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 [11, 12].

In this thesis, we present a new approach to the use of articulatory measurement data that are 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 are more feasible to collect at training time than at test time.

We use ideas from *multiview learning*, in which multiple "views" of the data (e.g., from multiple measurement modalities or an arbitrary partitioning of a single modality into two or more distinct feature sets) are available for training but possibly not for prediction at test time [13]. We distinguish this term from *multimodal* 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 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 [14]. 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), a statistical technique originally proposed by Hotelling in 1936 [15] that has recently been gaining popularity in machine learning research [13, 16]. Given training data corresponding to two views, CCA finds pairs of maximally correlated projections of the data [15, 16]. 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. Intuitively, projections of the acoustic features that are highly correlated with the articulation should be more discriminative than those that are not.

One challenge is that 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 are "private" to the acoustic view. We also present such combined approaches.

CHAPTER 2

MULTIVIEW 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 $x_i \in \Re^{d_1}$, $y_i \in \Re^{d_2}$, 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.

Figure 2.1 depicts the overall multiview architecture. Given two views, X and Y, we first learn a mapping from view X to a lower dimensional space; we then apply the learned transformation to new instances of data from view X. It is this intermediate feature transformation phase that characterizes all of multiview learning; in this thesis, we focus on linear mappings for their theoretical properties and ease of computation.

After the feature transformation is learned, we cosider the task of frame-wise 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. In particular, noise that is uncorrelated across views will be removed by CCA; to the extent that this assumption holds, then, the learned dimensions will be discriminative for phonetic classification. Figure 2.2 is a graphical model that illustrates this assumption.

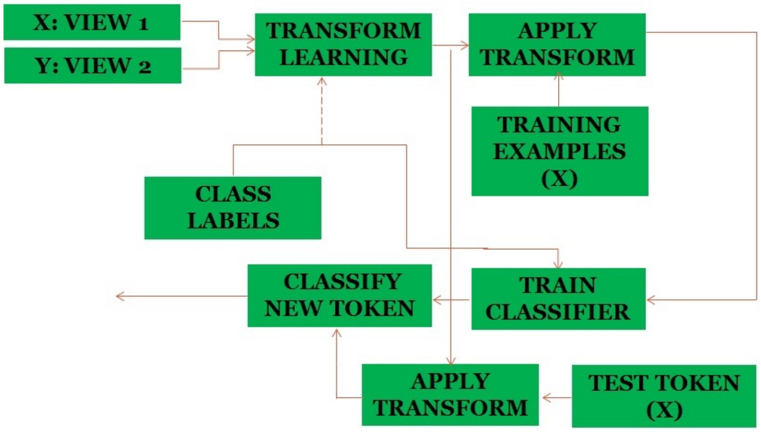

Figure 2.1: The multiview architecture

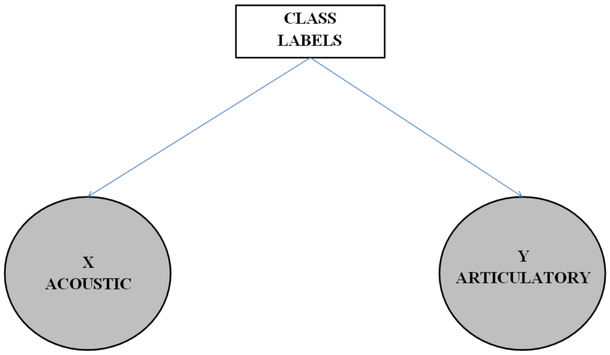

Figure 2.2: Graphical model: Representation of the uncorrelatedness assumption

2.1 Canonical Correlation Analysis

Canonical correlation analysis (CCA) [15, 16] 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

$$\{v_1, w_1\} = \arg\max_{v,w} \; corr(v^T X, w^T Y) \tag{2.1}$$

$$\{v_1, w_1\} = \arg\max_{v,w} \; \frac{v^T C_{xy} w}{\sqrt{v^T C_{xx} v w^T C_{yy} w}} \tag{2.2}$$

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 beauty of CCA lies in a very simple but useful observation: v and w are invariant to scaling, even when they are scaled independently; that is, suppose we replace v with αv and w with βw. As per Eq. (2.1), we have:

$$\{v_1, w_1\} = \arg\max_{v,w} \ \frac{\alpha v^T X Y^T \beta w}{\sqrt{(\alpha v^T X X^T \alpha v)(\beta w^T Y Y^T \beta w)}}$$

which equals

$$\{v_1, w_1\} = \arg\max_{v,w} \ \frac{\alpha \beta v^T X Y^T w}{\sqrt{(\alpha^2 \beta^2 v^T X X^T v)(w^T Y Y^T w)}}$$

and is identical to Eq. (2.2). We can exploit this relation by imposing the following two constraints:

$$\begin{aligned} v^T C_{xx} v &= 1 \\ w^T C_{yy} w &= 1 \end{aligned} \tag{2.3}$$

Consider the constrained optimization problem

$$\{v_1, w_1\} = \arg\max_{v,w} \ v^T C_{xy} w \tag{2.4}$$

$$\text{s.t.}$$

$$\text{Eq. (2.3) Holds}$$

The solution to Eq. (2.4) is the same as the solution to Eq. (2.2). The Lagrangian of Eq. (2.4) is given by

$$L(\lambda, v, w) = v^T C_{xy} w - \frac{\lambda_x}{2}(v^T C_{xx} v - 1) - \frac{\lambda_y}{2}(w^T C_{yy} w - 1) \tag{2.5}$$

Taking partial derivatives of Eq. (2.5) with respect to v and w yields:

$$\frac{\partial L}{\partial v} = C_{xy} w - \lambda_x C_{xx} v \tag{2.6}$$

$$\frac{\partial L}{\partial w} = C_{yx} v - \lambda_y C_{yy} w \tag{2.7}$$

Setting Eq. (2.6) and Eq. (2.7) to 0, we obtain $\lambda_x = \lambda_y$; let us denote them by λ. We obtain the following generalized eigenvalue problem:

$$C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx} v = \lambda^2 v$$
$$C_{yy}^{-1} C_{yx} C_{xx}^{-1} C_{xy} w = \lambda^2 w$$

There are several methods for solving a generalized eigenvalue problem, one of which is reducing it to a standard eigenvalue problem via a Cholesky decomposition.

It is straightforward to show [16] that the canonical directions are found as the solution of the following 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, corresponding to the M most correlated projections.

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 its sensitivity 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 [17]. 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, like many other multiview learning methods, 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.

2.1.1 Related Work

CCA has rarely been used for speech tasks. In [18], CCA was used to reduce dimensionality of acoustic features for improved clustering into speakers. In [6], it was used to learn linear transformations of acoustic features for improved speaker recognition in noise; in that work, the problem of shared vs. private dimensions was addressed by appending the CCA features to the baseline acoustic features (MFCCs). In [19], 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 [20, 21] where both views are 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, and we call this "private" information. For example, in our case the articulatory data do not include glottal or velar measurements. Therefore, the acoustic features are expected to contain "private" information about voicing and nasality.

In previous work [6], 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 recovers both shared and private dimensions, while making sure that they are not redundant. In this work we take a simple approach to find a set of correlated projections and a set of private projections, which are then concatenated to form our final acoustic feature vector. The procedure is as follows:

$(V, W) = \mathrm{CCA}(X, Y)$, 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.

$P = \mathrm{PCA}\left((V^{\perp})^T X\right)$, 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 [22] (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, $D = [V \; P]$.

CHAPTER 3

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 (XRMB) database, which includes simultaneous recordings of acoustic waveforms and articulatory measurements for a number of tasks and speakers [8]. 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 10 ms with a 25 ms window. We downsample the articulatory data to synchronize with the acoustics and discard any frames that have missing measurements (usually occurring due to mistracked pellets). Finally, for each frame we concatenate features (acoustic or articulatory) over a window of seven frames. This results in the data $X \in \Re^{273XN}$, $Y \in \Re^{112XN}$, where the columns of X are the acoustic data, the columns of Y are 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) with radial basis function kernels and k-nearest neighbors (kNN) using 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. The effects of data scaling on different classifiers tend to be different, so we also compare two forms of scaling – z-scaling (mean- and variance-normalization) and scaling to the range $[0, 1]$ – and treat this choice as a tuning parameter. 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. For SVMs, tuning is done on the first fold, and the resulting tuning parameters are used for testing in the remaining folds.[1]

We obtain phone labels for the XRMB corpus using the Penn Phonetics Lab Forced Aligner [23]. The alignments are imperfect, but anecdotally very good. Short pauses and stress are removed, leaving 39 phone classes.

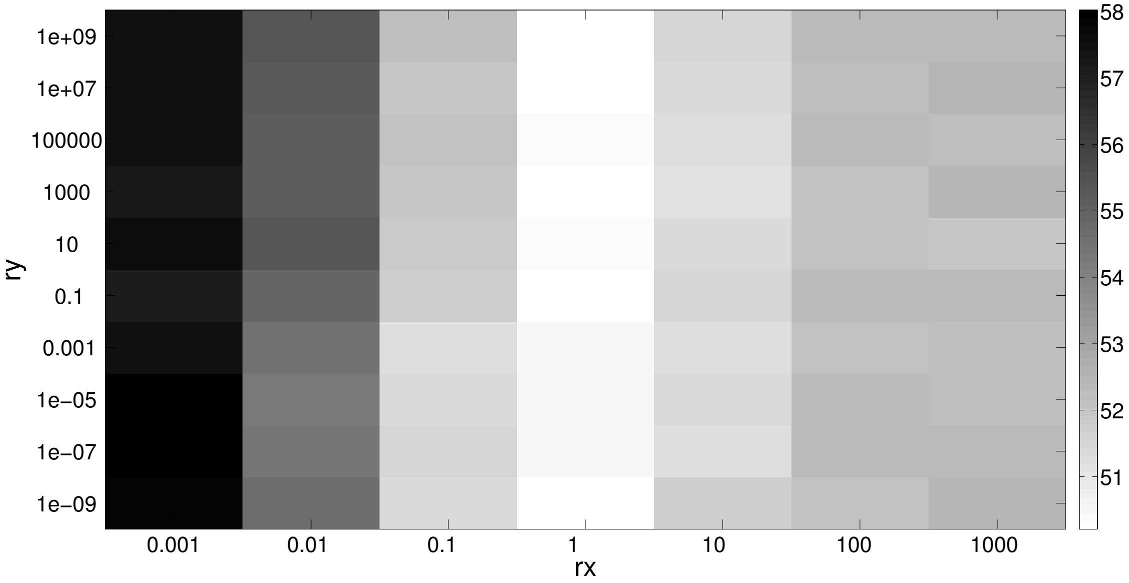

Figure 3.1: Dependence of error rate on CCA regularization

Figures 3.1 and 3.2 show the error rate as a function of the NCCA hyperparameters for speaker JW11 and kNN classifiers. 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 [17]. Figure 3.2 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,

[1] This may give a slight advantage to the results in the fifth fold, where the test set is the same as the development (tuning) set of Fold 1.

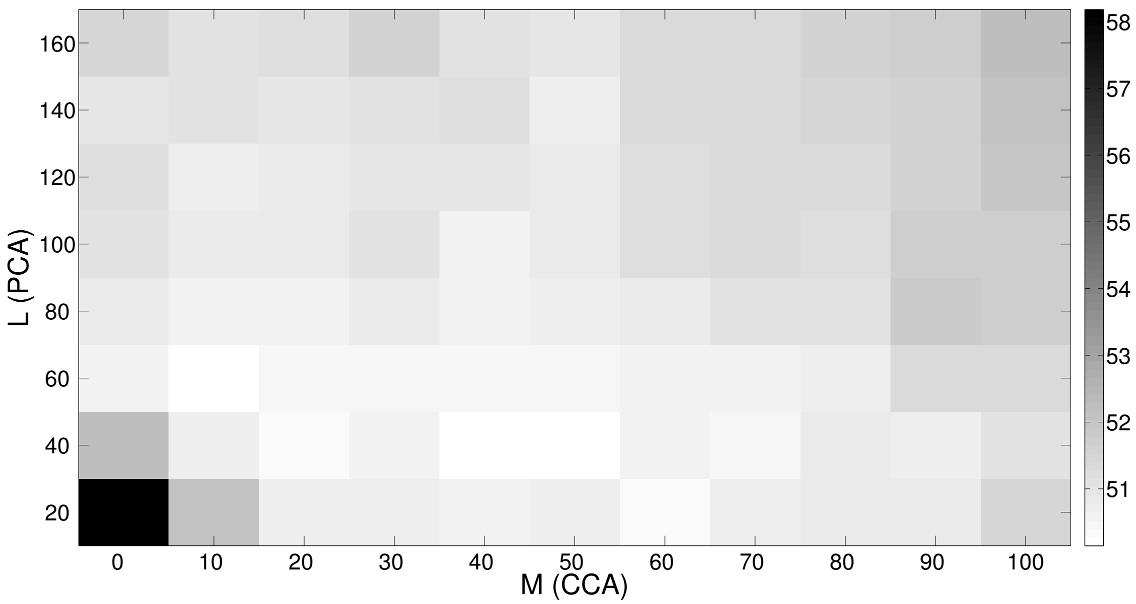

Figure 3.2: Dependence of error rate on NCCA dimensionalities

to a lesser extent, fold-dependent. The best values of k tend to be in $[8, 16]$, and of the final dimensionality in $[40, 120]$ (with varying divisions between PCA and CCA dimensionalities).

One of the assumptions made in many multiview learning problems (but not in ours, because it is not true) is the following: given the two views X and Y, and the class labels, Z, $I(X;Z) \approx I(Y;Z)$. To emulate a setting close to this assumption, we test on another dataset. We keep the articulatory features unmodified; for the acoustic data, rather than using the first and second derivatives in all seven frames, we use them for the end frames; the middle frames are the 13-dimensional MFCC vectors. This gives us a feature space with dimensionality $13 * 5 + 39 * 2 = 143$. We only test on one fold, under the assumption that results across different folds are approximately consistent.

CHAPTER 4

RESULTS AND DISCUSSION

In this chapter, we interpret our results, compare them with previous methods, and propose a fundamentally different approach that can alleviate many of the problems posed by CCA, but at the cost of computation.

4.1 Results

Table 4.1 shows the test set error rates averaged over the folds for each experiment. In all but one case (speaker JW30, SVM classifier), one or both of NCCA and MFCCA significantly improve over the baseline and the other techniques. CCA alone, however, does not usually improve over the baseline. This is in line with our intuition that the articulatory view is missing crucial information (such as voicing and nasality) that is important for phonetic classification. Figures 4.1, 4.2, 4.3, and 4.4 give a more detailed view of the NCCA and MFCCA results, showing the spread over the five folds.

Variance of classification error across the different folds stems from the fact that the XRMB corpus consists of drastically different tasks; random sampling on a frame-by-frame level should reduce differences in performance across the folds. The figures also indicate that the multiview approaches are not as useful when we classify with SVMs. This is expected because SVMs are insensitive to high dimensional data; dimensionality reduction is therefore not very useful, and sometimes even harmful.

We note that when experimental conditions are closer to the mutual information assumption, information that is private to the acoustics is not as essential; performance of CCA alone is significantly better than the baseline MFCCs and PCA, as demonstrated by Table 4.2. However, extensive experimentation is necessary to draw more concrete conclusions in this low-dimensional setting.

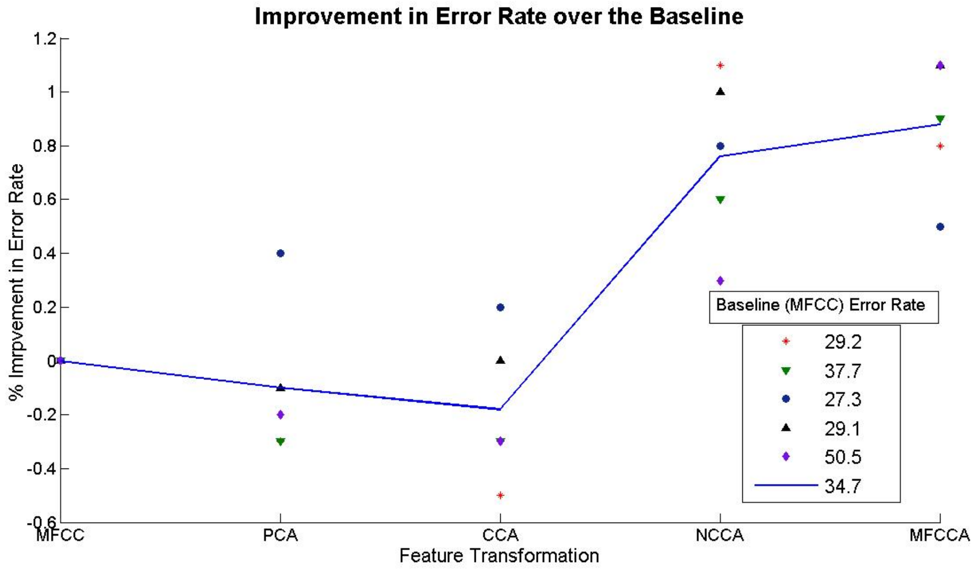

Figure 4.1: Improvement over baseline – Speaker JW11 using kNN classifier

In both cases, it is surprising that CCA performs no worse than MFCCs, despite the fact that CCA necessarily discards information about nasality and voicing. Likewise, NCCA performs significantly better, even though it contains no more information than the initial MFCCs. As discussed earlier, NCCA learns an intelligent coding of the MFCCs that concentrates the phonetically relevant information into a lower-dimensional subspace, by selecting the dimensions that are highly correlated with articulation and the top few PCA dimensions of the residual. Voicing and nasality typically are high variance entities; it is therefore not surprising that we select only the first few (high energy) dimensions of PCA and discard the rest, which are presumed to be noise.

Table 4.1: Error rates (in %) averaged over five folds

	JW11		JW30	
	kNN	SVM	kNN	SVM
MFCC	34.76	33.61	37.12	33.28
PCA	34.86	*35.93*	36.63	*36.80*
CCA	34.94	33.39	*38.22*	*35.10*
NCCA	**34.00**	33.39	**35.89**	*35.83*
MFCCA	**33.88**	**33.14**	36.83	33.17

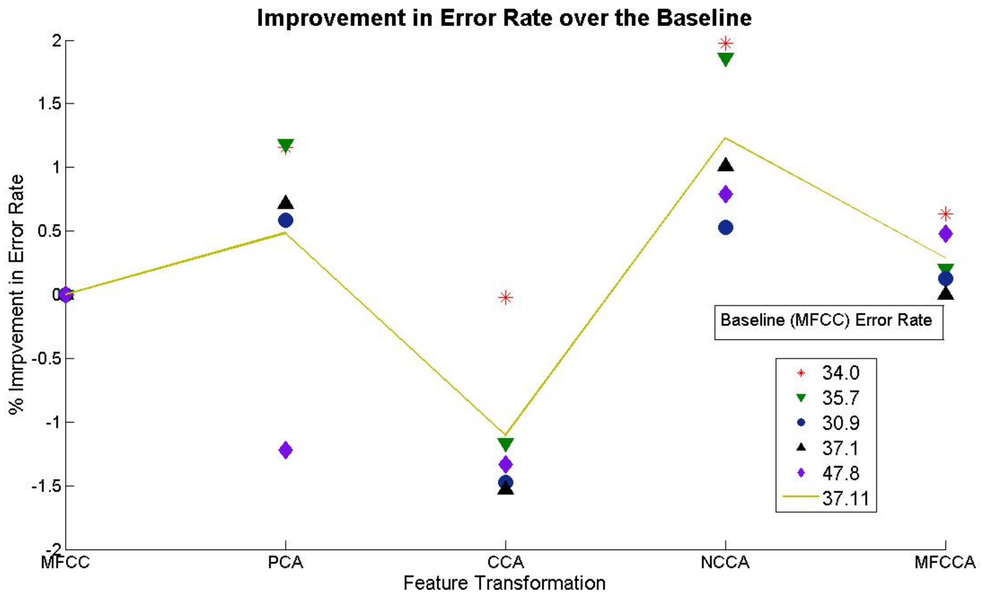

Figure 4.2: Improvemement over baseline – Speaker JW30 using kNN classifier

Table 4.2: Classification error for speaker JW11 using kNN (143 MFCC dimensions)

MFCC	PCA	CCA	NCCA
35.4	35.4	**30.0**	**30.0**

4.2 Comparison with Previous Approaches

Previous works in the use of articulatory information to design acoustic transformations for speech recognition have mostly taken distinctive feature based approaches, in which it is assumed that articulatory cues relevant to the phone label can be extracted from the acoustics [24, 25]. Statistical classifiers are generally trained to extract discriminative articulatory features, which are then used to train another classifier (usually HMMs). Borys in [24] used binary SVMs to detect the existence of specific cues relevant to the phone class. The work can generally be described as "landmark-based" learning, in which abrupt changes in distinctive features are detected. The SVMs were trained on manually labeled data. The key difference between the approach taken in [24] and ours is that we explicitly use recorded articulatory data, and learn a transformation by using the additional measurements as "soft supervision." In [24, 25], human experts decided on a set of acoustic

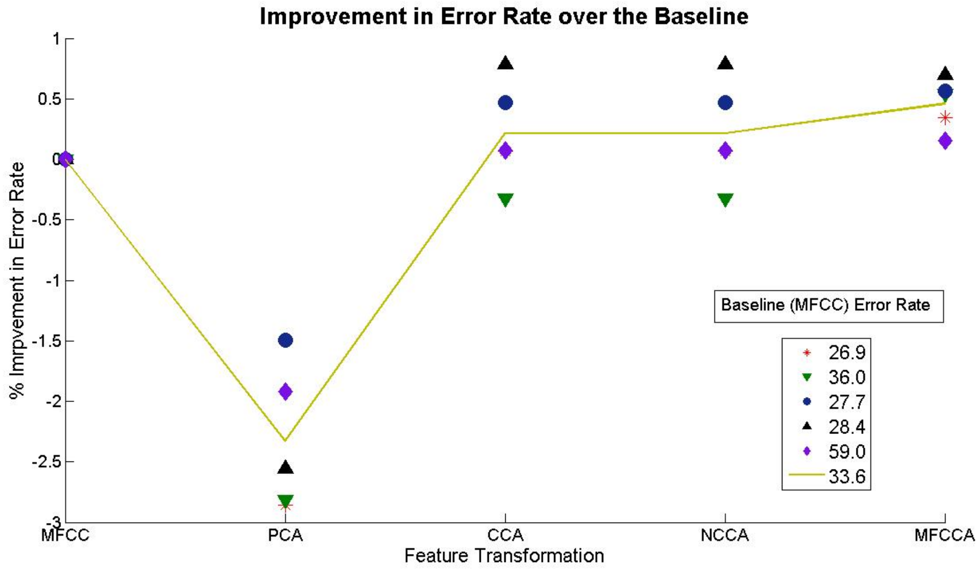

Figure 4.3: Improvement over baseline – Speaker JW11 using SVM classifier

and articulatory features as most relevant for phonetic classification, resulting in a manually constructed shared/private space factorization of acoustics and articulation. We, however, learn the optimal linear transformation such that correlation between the transformed spaces is maximized.

Markov in [10] considered articulatory measurements at training time, but not at test time. In [10], probabilistic dependancies between acoustics and articulation were learned by a hybrid HMM/BN model. The BN variables were partitioned into two sets and each set corresponded to a specific view. Variables representing articulation were assumed to be observable at training time, but not at test time. This makes Markov's approach closest to ours in spirit; however, the key difference is that we learn an explicit mapping from the acoustics to a feature space that depends on articulatory information available at training. Our approach is similar to some of the articulatory inversion methods explored by Vikramjit, where a mapping from acoustics to articulation is sought [26, 27]. While articulatory gestures can be inferred from the acoustics, it is not clear how the two can be integrated. A shared-private space factorization based approach such as ours automatically integrates articulatory information with acoustics based on some optimality criterion. In the case of CCA, the criterion is maximum correlation between the transformed spaces.

Our experiments have been limited to linear transformations and unsu-

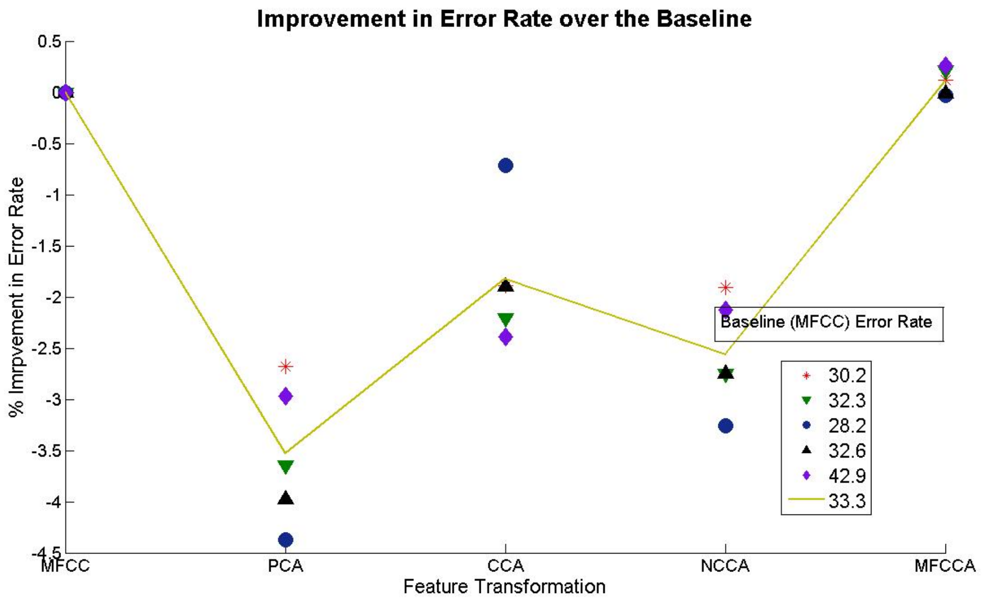

Figure 4.4: Improvement over baseline – Speaker JW30 using SVM classifier

pervised learning. Future work includes non-linear extensions [16, 22], supervised and semi-supervised extensions, application to noise-robustness (as in [6, 18]) 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 availale 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 multiview techniques will be much greater if they can be shown to extend beyond specific domains for which the views are available. Multiview 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 multiview techniques. In the next section, we discuss a radically different approach that has most of the characteristics we desire.

4.3 Factorized Latent Spaces

Suppose we have N observations obtained from V views:
$X = \{X^{(1)}, X^{(2)}, ..., X^{(V)}\}$, where $X^{(i)} \epsilon R^{P_i XN}$; we seek to find a latent representation $\alpha \epsilon R^{dXN}$ of dimension d and view specific projections

$$D = \{D^{(1)}, D^{(2)}, ..., D^{(V)}\}$$

such that the views are factored as follows: $X^{(1)} = D^{(1)}\alpha, ..., X^{(V)} = D^{(V)}\alpha$. For most applications of interest, such an α may not exist; it is most natural to consider instead the following optimization problem:

$$(\alpha^*, D^*) = argmin_{D,\alpha} \sum_{i=1}^{V} ||X^{(i)} - D^{(i)}\alpha||_F^2 \tag{4.1}$$

where $||.||_F$ is the Frobenius norm; of course, one could consider other matrix norms. The factorization obtained by solving Eq. (4.1) gives us one shared space α and we face the same problem as with CCA: How best can we reconstruct the spaces that are private to each view? Further, we may be interested in spaces that are shared only by some subset of the views and not all of them. Structured sparsity addresses both issues.

Consider instead the following optimization problem:

$$(\alpha^*, D^*) = argmin_{D,\alpha} \sum_{i=1}^{V} ||X^{(i)} - D^{(i)}\alpha||_F^2 + \Psi(\alpha^T) + \sum_{i=1}^{V} \Psi(D^{(i)}) \tag{4.2}$$

where $\Psi(.)$ is some convex relaxation that enforces sparsity, typically the l_1/l_2 norm or the l_1/l_∞ norm. The sparsity constraint encourages each projection, $D^{(i)}$, to use only a subspace of the latent space, α. The sparsity constraint on α automatically chooses the dictionary size as the smallest one that can still reconstruct X well [28]. Figure 4.5 is a graphical illustration of the sparse matrix factorization method described above. Extension to more than two views is also evident from Figure 4.5. We can incorporate class labels as another view, or in the form of an additional term in the objective function. In the next section, we describe some of the possible approaches.

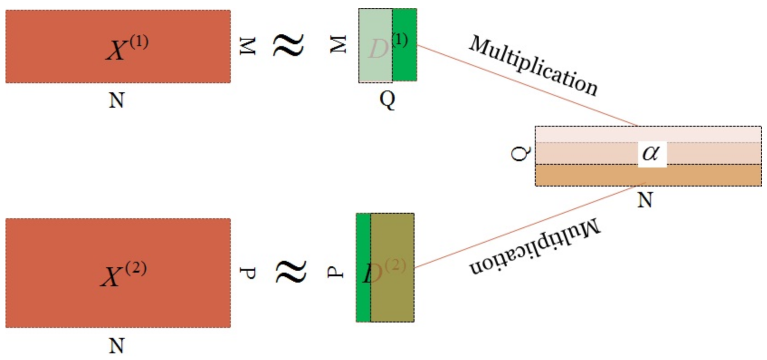

Figure 4.5: FLS as sparse matrix factorization

4.3.1 Supervised FLS

The primal formulation of the kernel support vector machine (SVM) problem is given by:

$$argmin_{\beta,b}\frac{1}{2}\sum_{i,j}\beta_i\beta_j k(x_i,x_j)+\frac{1}{n}\sum_{i=1}^{n}(L(y_i,b+\sum_{j=1}^{n}\beta_j k(x_i,x_j))) \tag{4.3}$$

where y_i is the class label corresponding to the sample x_i, $L(y_i,.)$ is some loss function, typically the hinge loss, and $k(.,.)$ is the kernel that implicity maps the data into a high dimensional space. Further, note that this is an unconstrained optimization problem which can easily be solved with gradient descent. The minimization problem can further be convex in β if we pick an appropriate $L(y_i,.)$. We can therefore readily embed this into the latent space formulation of Eq. (4.2) by including the SVM objective of Eq. (4.3) along with the original objective function.

$$(\beta,b,\alpha,D)=argmin_{\beta,\alpha,D}\{(4.2)+(4.3)\} \tag{4.4}$$

In Eq. (4.4), the arguments passed to the kernel function are vectors from the union of the acoustic/articulatory joint space and the space that is private to the acoustic view. This is under the assumption that while there are certainly aspects of the articulatory data that are irrelevant to the phonetic label, all of the acoustic features are relevant. The graphical model in Figure 4.6 better illustrates this.

The above problem is not jointly convex, but can be convex in each of its three arguments; as is typically done, we can alternate between minimizing

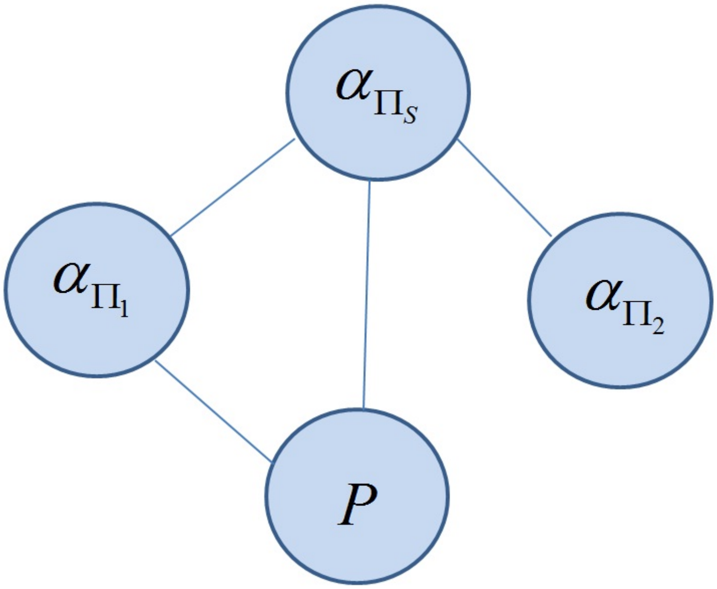

Figure 4.6: Graphical model depicting the dependencies among the various views

the three respective convex problems [28]. We would need a convex kernel, $k(.,.)$. Further, the training tokens in the SVM problem of Eq. (4.3) depend on $D^{(1)}$ and α in a non-convex fashion. The term x_i is the i^{th} column of alpha with only the rows that correspond to the non-zero columns of $D^{(1)}$. One possible relaxation is to pick a linear kernel $k(.,.)$ and relax $< w, x_i >$ to $||d_i^{(1)}|| < w, s_i >$, where $< .,. >$ denotes the dot product, and $d_i^{(1)}$ is the i^{th} column of $D^{(1)}$. The relaxation is appropriate for the linear case; whenever there is a zero column in $D^{(1)}$, the norm is zero and will have no effect on the inner product. It is continuous, with its effect proportional to its norm. We must be careful while extending this notion to non-linear kernels such as the radial basis fucntion (RBF). A nicer formulation is to consider the following: $||D||_{1,2} = \sum_j \sqrt{\sum_i D_{ij}^2}$, which can be rewritten as $min_{||\mu||_1 \leq 1} \sum_j \sum_i (\frac{D_{ij}}{\mu_j})^2$. This equivalence allows us to elegantly reformulate the optimization problem in terms of μ and write the above relaxation as $\mu. * \alpha$, where $.*$ denotes element-wise product. Yet another option is to prespecify dimensions of α that form the joint/private space of the acoustic view. This approach drastically simplifies the minimization problem; we would no longer need the sparsity regularizer on $\{D^{(1)}, D^{(2)}, ..., D^{(V)}\}$ and optimization over D reduces to a least squares problem.

CHAPTER 5

CONCLUSION

In this thesis, we have made an early attempt to develop a multiview feature learning framework based on CCA. We have introduced a novel method to integrate articulatory information with the acoustics; CCA alone would not have sufficed. Our implementation of shared-private factorizations such as NCCA and MFCCA suggests that our approach is effective. We are also among a handful of researchers to have incorporated recorded articulatory data effectively into a classifier. The real benefit lies in our multiview approach, which allows us to assume that articulatory information is not available at test time. Although our experiments have focused primarily on articulatory data, the methods described in this thesis are far more general. Temporally aligned data that are recorded from multiple views are abundant; for instance, we can use the same methods for audio-visual speech recognition, where we have access to both audio and video at training, but only audio at test time. A still more general and useful situation is where we have just the audio, but would like to reduce the cost of sampling. We can arbitrarily partition the space into two views: at training time, we can sample at the higher rate and learn a transformation based on CCA; at test time, we can sample at the lower rate, and assume that the "missing" samples correspond to the second view.

CCA, however, has some fundamental drawbacks. Some of these drawbacks can be alleviated by simple extensions such as the proposed notions of shared/private space factorizations, or extension to nonlinear versions such as kernel canonical correlation analysis. A bigger concern is that CCA is fundamentally limited to two views, with heuristic extensions when the number of views is greater than two. Furthermore, it is difficult to incorporate arbitrary task-specific functionals such as classification performance based on labels or a metric for the cluster quality. A more appropriate and general framework is that of FLS, which is inherently suited for several views, with the ability to tackle a much broader class of regression and classification problems. As we have shown, it is simple to introduce additional objectives to the optimization problem. Semi-supervised learning in this setting is also straightforward; whenever a specific view is unavailable, we simply ignore it and minimize over the known views. A key issue is the implementation of factorized latent spaces for large datasets. It is due to this bottleneck that we have been unable to run concrete experiments and have kept our discussion at a purely theoretical level. We hope to have efficient code for factorized latent spaces in the near future.

REFERENCES

[1] S. E. Levinson, *Mathematical Models for Speech Technology.* West Sussex, England: John Wiley and Sons, 2005.

[2] A. Adjoudani and C. Benoit, “On the integration of auditory and visual parameters in an HMM-based ASR,” in *Speechreading by Humans and Machines: Models, Systems, and Applications*, D. G. Stork and M. E. Hennecke, Eds. New York, NY: Springer, 1996, pp. 461–471.

[3] A. A. Wrench and K. Richmond, “Continuous speech recognition using articulatory data,” in *ICSLP*, 2000, pp. 145–148.

[4] A. Navot, “On the role of feature selection in machine learning,” Ph.D. dissertation, Hebrew University, Jerusalem, Israel, 2006.

[5] R. Haeb-Umbach and H. Ney, “Linear discriminant analysis for improved large vocabulary continuous speech recognition,” in *ICASSP*, 1992, pp. 13–16.

[6] K. Livescu and M. Stoehr, “Multi-view learning of acoustic features for speaker recognition,” in *ASRU*, 2009, pp. 82–86.

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

[8] 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, June 1994.

[9] J. Frankel and S. King, “ASR - articulatory speech recognition,” in *Eurospeech*, 2001, pp. 599–602.

[10] K. Markov, J. Dang, and S. Nakamura, “Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework,” *Speech Communication*, vol. 48, pp. 161–175, 2006.

[11] L. Deng, G. Ramsay, and D. Sun, “Production models as a structural basis for automatic speech recognition,” *Speech Communication*, vol. 33, pp. 93–111, 1997.

[12] 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, pp. 621–624.

[13] S. M. Kakade and D. P. Foster, "Multi-view regression via canonical correlation analysis," in *COLT*, 2007, pp. 82–96.

[14] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *ICASSP*, 2000, pp. 1635–1638.

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

[16] 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.

[17] T. D. Bie and B. D. Moor, "On the regularization of canonical correlation analysis," in *ICA*, 2003, pp. 785–790.

[18] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *ICML*, 2009, pp. 129–136.

[19] 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.

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

[21] M. Liu, Y. Fu, and T. S. Huang, "Audio-visual fusion framework with joint dimensionality reduction," in *ICASSP*, 2008, pp. 4437–4440.

[22] C. H. Ek, P. H. Torr, and N. D. Lawrence, "Ambiguity modelling in latent spaces," in *MLMI*, 2008, pp. 62–73.

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

[24] S. Borys and M. Hasegawa-Johnson, "Distinctive feature based svm discriminant features for improvements to phone recognition on telephone band speech," in *Interspeech*, 2005, pp. 697–700.

[25] K. Kirchhoff, "Robust speech recognition using articulatory information," Ph.D. dissertation, University of Bielefeld, Bielefeld, Germany, 1999.

[26] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, and L. Goldstein, "Retrieving tract variables from acoustics: A comparison of different machine learning strategies," *IEEE Journal of Selected Topics on Signal Processing*, vol. 4, pp. 1027–1045, 2010.

[27] V. Mitra, "Articulatory information for robust speech recognition," Ph.D. dissertation, University of Maryland, College Park, Maryland, 2010.

[28] Y. Jia, M. Salzmann, and T. Darrell, "Factorized latent spaces with structured sparsity," in *NIPS*, 2010, pp. 982–990.