Dimensionality Reduction of Speech Features using Nonlinear Principal Components Analysis

Stephen A. Zahorian[1], Tara Singh[2], Hongbing Hu[1]

[1] Department of Electrical and Computer Engineering, Binghamton University, Binghamton, NY, USA
[2] Department of Electrical and Computer Engineering, Old Dominion University, Norfolk, VA, USA
zahorian@binghamton.edu, tarasingh04@gmail.com, hongbing.hu@binghamton.edu

Abstract

One of the main practical difficulties for automatic speech recognition is the large dimensionality of acoustic feature spaces and the subsequent training problems collectively referred to as the "curse of dimensionality." Many linear techniques, most notably principal components analysis (PCA) and linear discriminant analysis (LDA) and several variants have been used to reduce dimensionality while attempting to preserve variability and discriminability of classes in the feature space. However, these orthogonal rotations of the feature space are suboptimal if data are distributed primarily on curved subspaces embedded in the higher dimensional feature spaces. In this paper, two neural network based nonlinear transformations are used to represent speech data in reduced dimensionality subspaces. It is shown that a subspace computed with the explicit intent of maximizing classification accuracy is far superior to a subspace derived as to minimize mean square representation error.

Index Terms: dimensionality reduction, nonlinear principal components analysis

1. Introduction

Methods in multivariate statistical analysis are essential for working with large amounts of geophysical data, data from observational arrays, from satellites, or from numerical model output. In classical multivariate statistical analysis, there is a hierarchy of methods, starting with linear regression at the base, followed by principal component analysis (PCA) and finally canonical correlation analysis (CCA). A multivariate time series method, the singular spectrum analysis (SSA), has been a fruitful extension of the PCA technique [4].

The common drawback of these classical methods is that only linear structures can easily be extracted from the data. For example, linear PCA yields a k-dimensional linear subspace of features that best represents the full data according to minimum square error criterion. If the data represents the complicated interaction of features, then the linear subspace may be a poor representation and a nonlinear subspace may be needed. Figure 1 illustrates one potential limitation of linear PCA [1]. The straight line fit to the data obtained by linear PCA does not provide much information about the original curve of the data. However, a nonlinear method might "discover" the curve that the data lies on.

For data that lies on curved subspaces, the basic concept is not to apply PCA directly to the given data but rather to a transformed version of the data [6]. More precisely, a nonlinear transformation can be described as

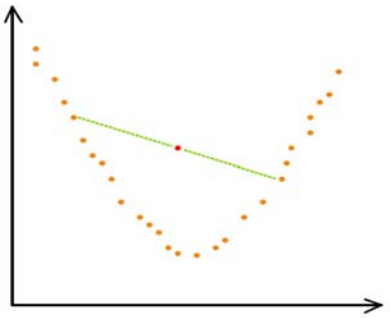

Figure 1: Straight line obtained using linear PCA, a poor representation of the nonlinear data.

$$\phi(.): R^D \to R^M \qquad x \to \phi(x), \tag{1}$$

such that the structure of the resulting data $\phi(x)$ becomes significantly more linear. In machine learning, $\phi(x)$ is called the feature of the data point *x*, and R^M is called the feature space. This feature space can be organized as a matrix:

$$\Phi = [\phi(x_1), ..., \phi(x_N)] \in R^{M*N}. \tag{2}$$

The principal components of the feature space are given by the eigenvectors of the sample feature covariance matrix:

$$\sum_{\phi(x)} = (\frac{1}{N}) \sum_{i=1}^{N} \phi(x_i)\phi(x_i)^T = (\frac{1}{N})\Phi\Phi^T \in R^{M*M}. \tag{3}$$

Let $v_i \in R^M$ and λ_i be the eigenvectors and eigenvalues of $\sum_{\phi(x)}$:

$$\sum_{\phi(x)} v_i = \lambda_i v_i, \quad i = 1, ..., M, \tag{4}$$

then the nonlinear principal components y_i of every data point *x* are given by

$$y_i = v_i^T \phi(x) \in R, \quad i = 1, ..., d. \tag{5}$$

In many cases, searching for the proper map $\phi(.)$ is a difficult task limiting the use of nonlinear PCA. However, in some practical applications, good candidates for the map can be found from the nature of the problem. For an arbitrary nonlinear relationship expressed by $\phi(.)$, a neural network is an excellent approach to use because of its universal approximation property [3].

It is widely known that mapping performed by a neural network can approximate any continuous function with arbitrarily desired accuracy [2]. The concept of extracting features from highly nonlinear data has been discussed by a number of researchers with most techniques reported in the

literature based upon artificial neural networks [5]. This is possible due to the capability of neural networks to provide a nonlinear transformation of a feature space.

2. NLPCA approaches

The two methods for nonlinear PCA (NLPCA) investigated in this paper are based on neural networks. In particular, a bottleneck neural network as shown in Figure 2, is used to perform the dimensionality reduction.

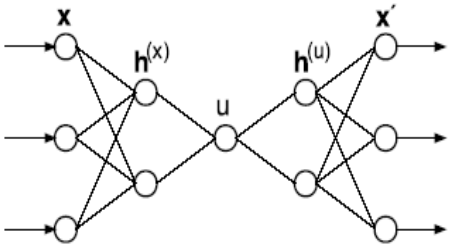

Figure 2: Bottleneck neural network.

Two approaches were used for training the bottleneck neural network. In the first approach, which we refer to as NLPCA1, the neural network is trained as an identity map. That is, the network is trained to minimize mean square error using targets that are the same as the inputs. As was mentioned in [4], training with regularization is often needed to "guide" the network to a better minimum in error. In the second approach, referred to as NLPCA2, the network structure shown in Figure 2 is trained as classifier. This second method could also be viewed as nonlinear discriminant analysis, since the network is trained to maximize discrimination. In both cases, the data at the output of the middle hidden layer, the bottleneck layer, is the reduced dimensionality data.

Figures 3 and 4 illustrate the potential of NLPCA1 to determine a curved subspace embedded in a higher dimensional space. In particular for the case of Figure 3, two-dimensional pseudo-random data was generated that is clustered about a parabolic curve in the 2-D space. A neural network with 1 hidden node was able to determine this underlying data structure.

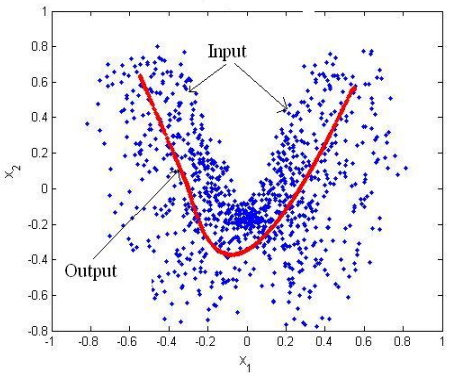

Figure 3: Plot of input and output data for semi-random 2-D data. The output data is a plot of reconstructed data obtained after passing the input data through the trained neural network.

Figure 4 illustrates a case where pseudo-random 3-D data is constrained to lie on a Gaussian surface, and a neural network with 2 hidden nodes is shown to be able to determine this underlying data structure.

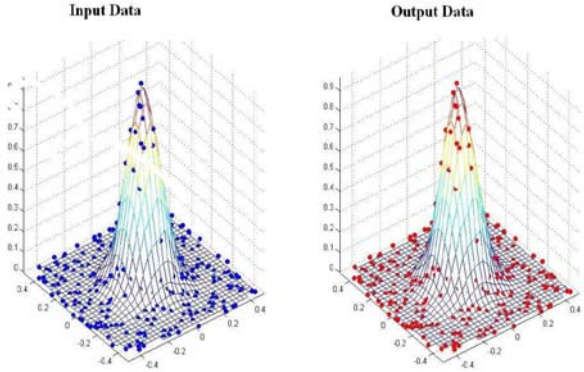

Figure 4: An example with 3D data. Input and output plot of 3-D Gaussian data before and after using neural network for NLPCA.

3. Experimental evaluation

The two versions of NLPCA as well as linear PCA and LDA were compared with vowel classification experiments for various numbers of features. The 10 steady-state vowels /ah/, /ee/, /ue/, /ae/, /ur/, /ih/, /eh/, /aw/, /uh/, and /oo/ were extracted from the NTIMIT database and used [8]. All the training sentences (4620 sentences) were used to extract a total of 31,300 vowel tokens for training. All the test sentences (1680 sentences) were used to extract a total of 11,625 vowel tokens for testing. For each vowel token, 39 DCTC-DCS features were computed, as described in [7].

For all cases, including original features, LDA, PCA, and the two versions of NLPCA, a neural network classifier with 100 hidden nodes and 10 output nodes, trained with backpropagation, was used as the classifier. In addition, a maximum likelihood Mahalanobis distance based Gaussian assumption classifier (MXL) was used for evaluation.

For the NLPCA cases, the first and third hidden layers had 100 nodes (empirically determined). The number of hidden nodes in the second hidden layer was varied from 1 to 39, according to the dimensionality being evaluated. For the case of NLPCA2, the network used for dimensionality reduction was also a classifier. For the sake of consistency, the outputs of the hidden nodes from the bottleneck neural network were used as features for a classifier, using either another neural network or the MXL classifier. [1]

3.1. Experiment 1

In the first version of this experiment, all training data were used to train the transformations including LDA, PCA and two NPLCAs as well as the classifiers.

Figure 5 shows the results based on the neural network and MXL classifiers for each transformation method in terms of classification accuracy, as the number of features varies

[1] It was, however, experimentally verified that classification results obtained directly from the bottleneck neural network were nearly identical to those obtained with this other network.

from 1 to 39. For both the neural network and MXL classifiers, highest accuracy was obtained with NLPCA2, especially with a small numbers of features. For the MXL classifier, NPLCA2 features result in approximately 10% higher classification accuracies as compared to all other features. For both the neural network and MXL classifiers, accuracy with NLPCA1 features was very similar to that obtained with linear PCA.

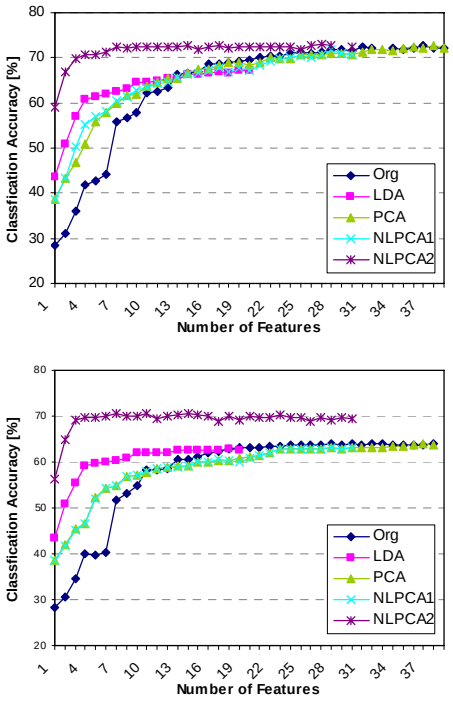

Figure 5: Classification accuracies of neural network (top panel) and MXL (bottom panel) classifiers with various types of features.

Another evaluation was conducted with a small amount of training data, since dimensionality reduction is often not needed with enough training data and a powerful classifier. Here, experiments were conducted using 1%, 2%, 5%, 10%, 25%, 50% and 100% of the available training data.

Figure 6 shows the results for 2% and 50% training data. The classification accuracies of the original and NLPCA2 reduced features were compared using both the neural network and MXL classifiers. Using 2% of the available training data (approximately 626 vowel tokens total), overall best accuracy is obtained with a parametrically based classifier (MXL) and a large number of features (18 or more) For the experiment with 50% of the available training data, the accuracy with NLPCA2 features is substantially higher than those obtained with original features, at least for 12 or fewer features. With a large number of features, accuracy is approximately the same with either original or NLPCA2 features for both classifiers. Overall lowest accuracy is obtained with original features and the MXL classifier. The superiority of NLPCA2 with few features was also found using 10% and 25% of the training data.

Thus, the overall conclusion of experiment 1 is that the NLPCA2 method is quite effective for improving classification accuracy with a small number of features, but does not result in an improvement with a large number of features.

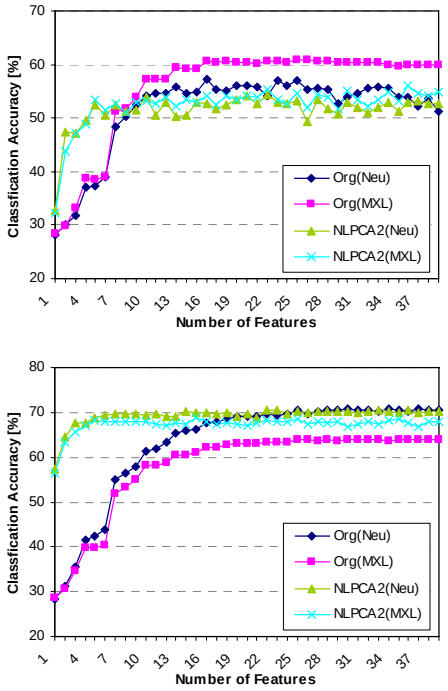

Figure 6: Classification accuracies of neural network and MXL classifiers based on original features and NLPCA2 reduced features with 2% (top panel) and 50% (bottom panel) of the training data.

3.2. Experiment 2

As mentioned above, dimensionality reduction is not usually advantageous (in terms of accuracy) for classifiers trained with enough data. However, for the case of complex automatic speech recognition systems, there is generally not enough training data. To simulate lack of training data, from a different perspective, another experiment was conducted.

In the experiment, the training data was separated into two groups, with about 50% in each group. One group of data (group 1) was used for training transformations while the other data (group 2) was used for training classifiers. In contrast to experiment 1 for which the same percentage of data was used for both the training of transformations and classifiers, for experiment 2, a fixed 50% of the training data was used for training transformations and a variable percentage, ranging from 1% to 100% of the other half of the training data, was used for training classifiers.

The results obtained with the neural network and MXL classifiers using 10% of the group 2 training data (that is, 5% of the overall training data) are shown in Figure 7. The numbers of features evaluated are 1, 2 4, 8, 16 and 32. For both the neural network and MXL classifiers, NLPCA2 clearly performs much better than the other transformations or the original features.

Figure 8 shows the classification accuracies of transformations with various percentages of training data. The neural network and MXL classifiers both used 4 features obtained from dimensionality reduction. NLPCA2 yields the best performance, with about 68% accuracy for both cases. Similar trends were also observed for 1, 2, 8, 16, and 32

features. However, the advantage of NLPCA2 decreases with an increasing number of features.

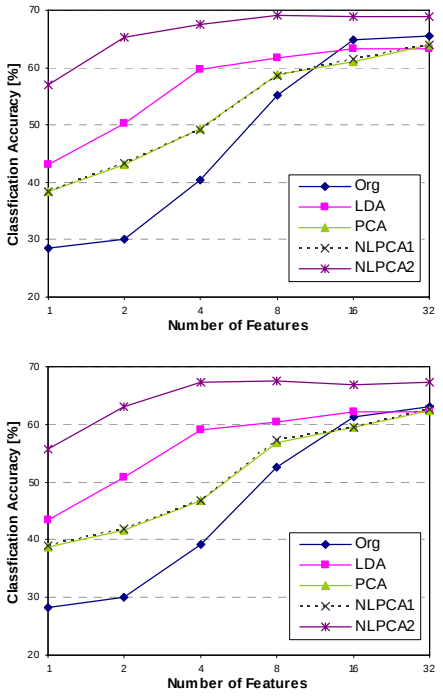

Figure 7: Classification accuracies of neural network (top panel) and MXL (bottom panel) classifiers using 10% of group 2 training data for training classifier.

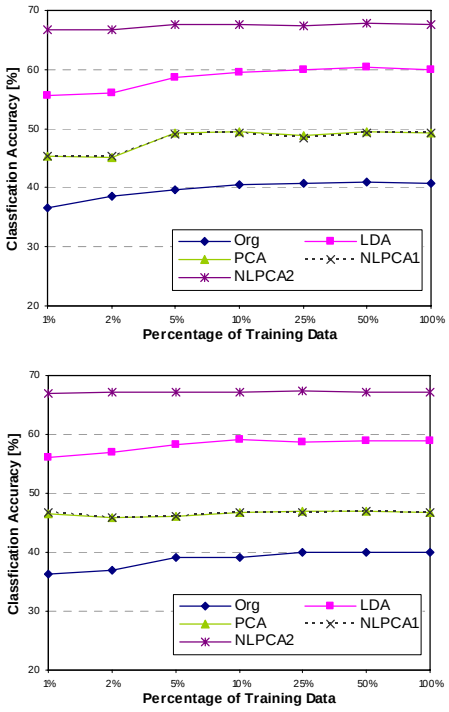

Figure 8: Classification accuracies of neural network (top panel) and MXL (bottom panel) classifiers with various percentages of classifier training data using 4 features.

The results of experiment 2 illustrate that the NLPCA2 method is much more effective at capturing the structure of data than PCA, LDA, or NLPCA1, in terms of classification accuracy, using a small amount of additional data to train a classifier. However, since group 1 training data was used to form the transformations, overall, over 50% of the training data was used, directly or indirectly.

4. Conclusions

Two nonlinear methods based on neural networks were presented as feature dimensionality reduction techniques and experimentally compared with linear methods for dimensionality reduction. A nonlinear technique which minimizes mean square reconstruction error from a reduced dimensionality space can be very effective for representing data which lies in curved subspaces, but does not appear to offer any advantages over linear dimensionality reduction methods for a speech classification task. A nonlinear technique for dimensionality reduction based on minimizing classification error is quite effective for accurate classification in low dimensionality spaces. For the case of vowel classification, from telephone speech, approximately 8 dimensions are nearly as effective as using all 39 original dimensions. Additionally, the reduced features appear to be well modeled as Gaussian features with a common covariance matrix. Additional testing is needed to determine if there are advantages to using this dimensionality reduction method with HMMs in a more complex ASR task.

5. Acknowledgements

This work was partially supported by JWFC 900.

6. References

[1] Chapel, L., "Dimension Reduction by Non Linear Methods," http://www.cs.unc.edu/Courses/comp290-90-f03/DimReduction2.pdf, Fall 2003.

[2] Dong, D. and McAvoy, T.J., "Nonlinear Principal Component Analysis-Based on Principal Curves and Neural Networks," Proceedings of the American Control Conference, pp.1284-1288, June 1994.

[3] Hornik, K., Stinchcombe, M., and White, H., "Multilayer Feedforward Neural Networks are Universal Approximators," Neural Networks, Vol.2, pp.359-366, 1989.

[4] Hsieh, W. W., "Nonlinear Multivariate and Time Series Analysis by Neural Network Methods," Review of Geophysics, Vol.42, March, 2004.

[5] Jang C.S. and Un, C.K., "A new parameter smoothing method in the hybrid TDNN/HMM architecture for speech recognition," Speech Communication. vol. 19 (4) pp.317-324, 1996.

[6] Matthias, S., Fatma, K., Charles, L. G., Joachim, K., and Joachim, S., "Non-linear PCA: a Missing Data Approach," Bioinformatics, Vol.21, pp.3887–3895, August 2005.

[7] Zahorian, A. S. and Jagharghi, A., "Spectral-shape Features versus Formants as Acoustic Correlates for Vowels," J. Acoust. Soc. Amer., vol. 94, pp 1966-1982, 1992.

[8] Zahorian A. S., Zimmer M., and Meng F., "Vowel Classification for computer-based visual feedback for speech training for the hearing impaired," ICSLP, 2002.