A Neural Network Based Nonlinear Feature Transformation for Speech Recognition

Hongbing Hu, Stephen A. Zahorian

Department of Electrical and Computer Engineering, Binghamton University,
Binghamton, NY 13902 USA

Hongbing.hu@binghamton.edu, Zahorian@binghamton.edu

Abstract

A neural network based feature dimensionality reduction for speech recognition is described for accurate phonetic speech recognition. In our previous work, a neural network based nonlinear principal component analysis (NLPCA) was proposed as a dimensionality reduction approach for speech features. It was shown that the reduced dimensionality features are very effective for representing data for vowel classification. In this paper, we extend this neural network based NLPCA approach for phonetic recognition using continuous speech. The reduced dimensionality features obtained with NLPCA are used as the features for HMM phone models. Experimental evaluation using the TIMIT database shows that recognition accuracies with NLPCA reduced dimensionality features are higher than recognition rates obtained with original features, especially when a small number of states and mixtures are used for HMM phonetic models.

Index Terms: feature dimensionality reduction, neural networks, HMMs, speech recognition

1. Introduction

Accurate automatic speech recognition (ASR) requires both speech features that are highly discriminative and a recognition model which can form arbitrary boundaries in the feature space. Highly discriminative features are required to incorporate nonlinear frequency scales and time dependency to efficiently present the spectral/temporal characteristics of speech. In [9], nonlinear warping functions are used to obtain smoothed speech features for highly discriminative features. The use of factors that represent dynamic information also increases recognition accuracy. Another important consideration for features is the large dimensionality of feature spaces, leading to problems referred to as the "curse of dimensionality" [2, 3]. That is, for a fixed set of training data, as additional features or dimensions are added, recognition performance improves on the training data but sometimes degrades on test data. Therefore, a compact set of highly discriminative features is very important for accurate ASR.

The hidden Markov model (HMM) has been one of the most successful recognition models since it has good time alignment capability, a well-developed mathematical framework, and convenient mechanisms for incorporating language models. On the other hand, recognition systems using neural networks have some advantages over HMMs. Neural networks are not based on any statistical assumptions, and have good discriminative power. However, neural networks do have limitations such as the lack of ability to model temporal variations.

In our previous work [10], a neural network based nonlinear principal component analysis (NLPCA) was presented as a dimensionality reduction approach for speech features. It was shown that the reduced dimensionality features using this approach can be very effective for representing data for vowel classification. In this paper, we extend this neural network based NLPCA approach to a more complex speech recognition system in which a HMM recognition model is used for phonetic recognition in continuous speech.

This combination of neural networks and HMMs described in this paper can be considered as a hybrid NN/HMM recognition system [8]. In contrast to other hybrid methods [1, 2, 4, and 5] where neural networks are closely integrated with HMMs in the training process, the recognition system described in this paper uses neural networks as a form of preprocessing. This approach results in a simple and fast process, and gives the flexibility and potential to combine the neural network with other processing methods.

The remainder of this paper is organized as follows: In Section 2, we give an overview of the system architecture, which consists of feature transformation (dimensionality reduction) with NLPCA and feature recognition with a HMM. The NLPCA training is described in Section 3. Section 4 presents the recognition performance using the TIMIT database for various conditions, followed by the conclusion in Section 5.

2. NLPCA for HMM Recognition

The architecture of the HMM recognition using NLPCA is illustrated in Figure 1.

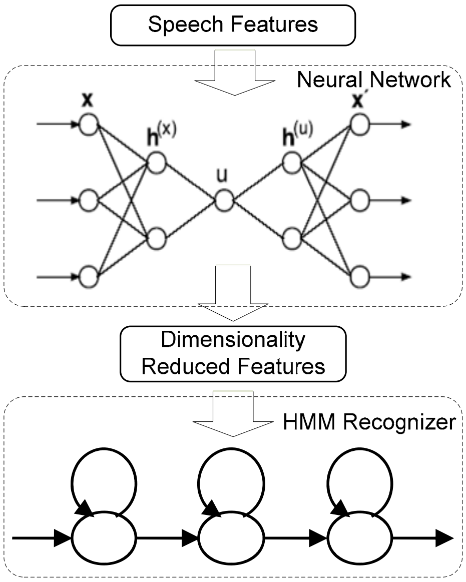

Figure 1: Architecture of HMM recognition using NLPCA

As reported in [10], NLPCA is based on a bottleneck neural network and uses the activations from the middle hidden layer of the network as the reduced dimensionality data. The reduced features are then recognized as phonemes using HMMs as well as language model information.

Since the activations of the middle layer represent the internal structure of the input features, the reduced features created by the NLPCA network may be more suitable for recognition or classification than the input features. The dimensionality of the reduced feature space is determined only by the number of nodes in the middle layer. Thus, an arbitrary number of reduced dimensions can be obtained, independent of the input feature dimensions and the nature of the training targets. This flexibility of dimension determination allows dimensionality to be adjusted so to optimize overall ASR accuracy.

Phoneme HMMs are used in this paper for phonetic recognition experiments, although other recognition units could be used. In the calculation of the emission probability for each state in a HMM, the reduced dimensionality features are used instead of the original features, based on the following Gaussian Mixture Model (GMM). Given a feature vector **o** at time i, the emission probability $b_j(\mathbf{o}_i)$ of the j'th model is

$$b_j(\mathbf{o}_i) = \sum_{m=1}^{M} c_{jm} N(\mathbf{o}_i; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}) \ ,$$

where M is the number of mixture components, c_{jm} is the weight of the m'th component, and $N(\mathbf{o}; \boldsymbol{\mu}, \Sigma)$ is a multivariate Gaussian with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, that is

$$N(\mathbf{o}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{o}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{o}-\boldsymbol{\mu})} \ ,$$

where n is the dimensionality of feature vector **o**.

After emission and transition probabilities are estimated, the Viterbi algorithm is used to calculate the overall probabilities of a feature vector over HMMs and find the HMM with highest probability as the recognition result.

The parameters of the HMMs are trained by the Baum-Welch estimation algorithm. Some global optimization training methods have been used in hybrid NN/HMMs recognition models [1, 7]. However, in this work, the weight training of neural network in NLPCA, as described later, is conducted separately from the HMM training.

3. NLPCA Training

In our previous work [10], two approaches were used for training the bottleneck neural network in NLPCA. One approach (NLPCA1) is to train the neural network as an identity map. However, this approach was not as effective as a second approach (NLPCA2), for which the neural network is trained as classifier. NLPCA2 was also found to be superior to linear PCA. Therefore, only the NLPCA2 method was used for all work reported in this paper.

A difficulty in neural network training is that the input data has a wide range of means and variances for each feature component. These wide ranges can lead to difficulties in training. In order to avoid this difficulty, the input data is scaled so that all feature components have the same mean and variance. In particular, a input feature vector **x** at time i is scaled using

$$\mathbf{o}_i = \frac{\mathbf{x}_i - \boldsymbol{\mu}}{r\boldsymbol{\sigma}} ,$$

where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\sigma}$ is the standard deviation vector of input features. The scale factor r was set to 5; with this scaling the mean of each scaled vector component of **o** is 0.0, and the standard deviation of each component of **o** is 0.2, thus resulting in a range of approximately [-1,1] for all the scaled feature vectors. The mean vector $\boldsymbol{\mu}$ and standard deviation vector $\boldsymbol{\sigma}$ are computed from the training data. However, both training and test data were scaled the same way.

The training targets for NLPCA2 correspond to category indices, that is, the phoneme label for the work reported in this paper. An appropriate representation of the target is thus required for the output layer of the network. We use a number of output nodes equal to the number of phone categories (39 for this paper), with a value of 1 for the target category and 0 for the non-target categories. For instance, in a four-category problem, if the pattern is in category 3, the target vector is set as **t** = (0, 0, 1, 0). In every target vector, one component is 1 and all the rest are 0.

The weights of the neural network are estimated using a standard Back-propagation algorithm to minimize the distance between the scaled input features and target data described above. Since the target data represents the category information of the input data, the neural network is trained to have good classification ability as well as reduce feature dimensionality.

4. Experimental Evaluation

Several experiments were conducted to evaluate the proposed method based on the widely used TIMIT database, consisting of 630 speakers and 10 sentences from each speaker. A total of 4620 sentences were used for training, and the remaining 1680 sentences were used for testing. A reduced 39 phone set as used in [6] was mapped down from the original TIMIT 62 phone set and used in the experiments.

For both training and testing data, the modified Discrete Cosine Transformation Coefficients (DCTC) and Discrete Cosine Series Coefficients (DCSC) were extracted as original features. As in our work with vowel classification [9], the modified DCTC is used for representing speech spectra, and the modified DCSC is used to represent spectral trajectories. For the all experiments, 13 DCTCs and 7 DCSCs were computed using 20 ms frames with 10 ms frame spacing, for a total of 91 features. In Experiment 1, the DCSC term was computed using 10 frames per block (block length of 100 ms). Then, this block length was varied for the evaluations in Experiments 2 and 3.

Left-to-right Markov models with no skip were used and a total of 39 monophone HMMs were created from the training data using the HTK toolbox (Ver3.4). The bigram phone information extracted from the training data was used as the language model. Various numbers of states and mixtures were evaluated as described in the following experiments.

The neural network used for NLPCA had 3 hidden-layers with 500 nodes in the first and third hidden layers. The number of hidden nodes in the second hidden layer (the bottleneck) was varied from 4 to 91, according to the reduced dimensionality being evaluated. The numbers of nodes in the input and output layers were 91 and 39 respectively, with 91 corresponding to the dimensionality of the original features and 39 determined by the number of phone categories.

4.1. Experiment 1

In the first experiment, NLPCA was evaluated with various dimensions in the reduced feature space. The HMMs were trained with 1, 3 and 5 states, and with 1, 2, 5 and 10 mixtures for each state.

Figure 2 shows the results based on 1-state and 3-state HMMs, with various numbers of mixtures, in terms of recognition accuracy, as the reduced dimensionality of features varies from 4 to 91. The dotted lines are the results with the original 91 features. For the case of 1-state HMMs, the highest recognition accuracy of 65.46% was obtained with a 15-dimensional feature space and 10 mixtures. For the simplest HMM case, with 1 state and 1 mixture, the 15-dimensional NLPCA features result in approximately 20% higher accuracy than obtained with the original 91 features. For the case of 1 state and 10 mixtures, the 15-dimensional NLPCA space results are about 5% higher than those obtained with the original 91 features. For the 3-state HMMs, the highest accuracy was obtained with 10 mixtures using 20-dimensional NLPCA features. Furthermore, accuracy obtained with the 20-dimensional NLPCA features and 5 mixtures is similar to that obtained with original features and 10 mixtures.

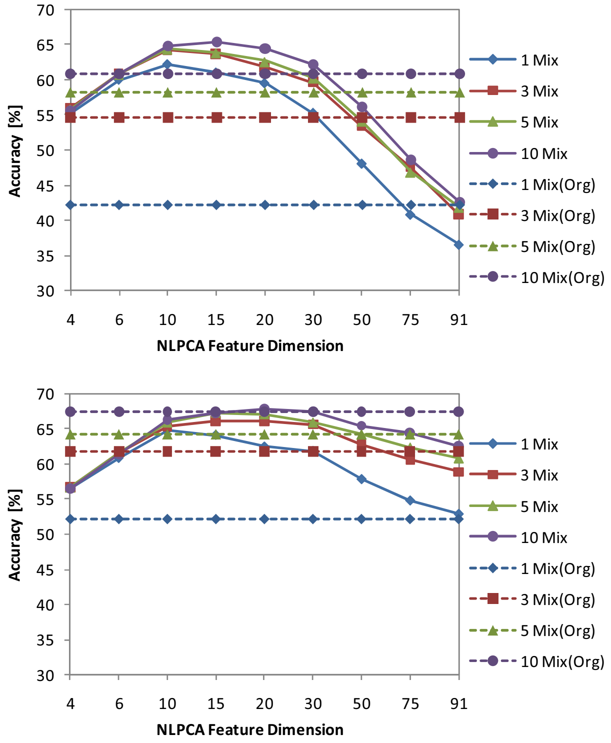

Figure 2: Recognition accuracies of 1-state HMMs (top panel) and 3-state HMMs (bottom panel) with various reduced feature dimensions.

For both 1-state and 3-state HMMs, best accuracy was obtained with feature dimensionality between 10 and 30, especially when a small number of mixtures are used. In a similar experiment, the 5-state HMMs also resulted in higher accuracy with NLPCA reduced dimensionality spaces versus the original 91 dimensionality space.

These results imply that the NLPCA is able to represent the complexity of original feature in a reduced dimensionality space. Furthermore, with an HMM speech recognizer, the reduced features result in high accuracy using a small number of mixtures and states in the HMM phone models.

4.2. Experiment 2

In the previous experiment, a fixed block length of 100 ms was used for DCTC-DCSC feature calculations. In this experiment, various block lengths were evaluated to optimize the block length that would maximize the accuracy of the entire system. The 20-dimensional NLPCA features that gave the best performance in Experiment 1 were used to compare with the original features.

The recognition accuracies obtained with 1-state and 3-state HMMs with various numbers of mixtures are shown in Figure 3. The block lengths evaluated were 70, 100, 150, 200, 250 and 300 ms. For both 1-state and 3-state HMMs, highest accuracy was obtained with reduced dimensionality features using block lengths between 100 and 200 ms. In contrast to the original features which largely degrade performance with an increasing block length, the 20-dimensional NLPCA features lead to a small degradation.

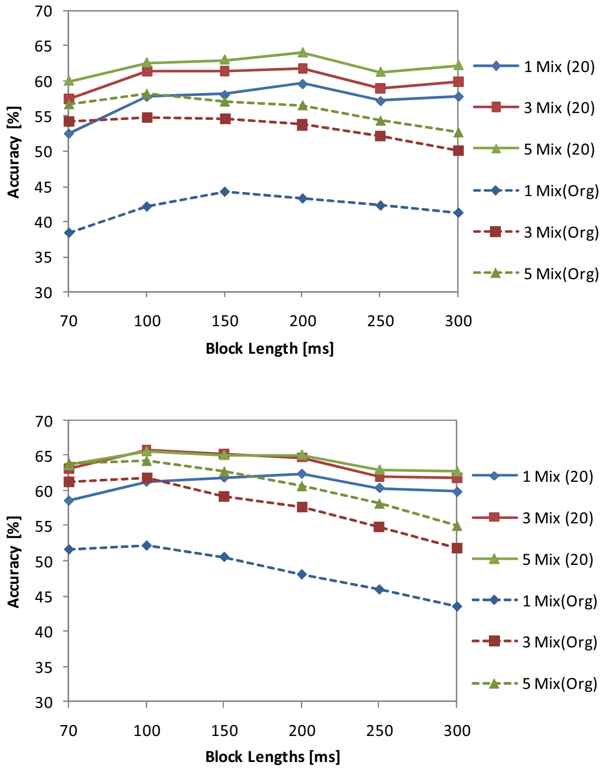

Figure 3: Recognition accuracies of 1-state (top panel) and 3-state HMMs (bottom panel) using the original and reduced features with various block lengths.

These results show that the NLPCA features are better able to represent speech information over longer segment lengths than are the original features. For example, best results for 1-state HMMs with NLPCA features are obtained with a block length on the order of 200 ms, versus best results with original features for the shortest block length tested (70 ms). Additional, a very simple 1-state 5-mixture monophone HMM model based on 20 NLPCA features is able to achieve phone accuracy only slightly worse (63.99%) than the best 3-state model tested (65.68%). These results thus imply that the NLPCA can account for some of the temporal information accounted with HMMs, thus potentially simplifying the HMM configuration.

4.3. Experiment 3

A case could be made that the advantages of the NLPCA features over the original features are only due to the large number of neural network parameters obtained from training data. This also leads to the question of whether NLPCA neural network weights trained from one set of data would generalize well to an HMM trained with another set of data. Therefore experiments were conducted with 50% of the training data used for NLPCA and the other 50% of the training data for HMMs.

For this experiment, the 4620 training sentences were equally separated into two groups so that each group had 2310 sentences with 5 sentences for each speaker. One group of data was used for training NLPCA while the other group was transformed by the trained NLPCA and then used for training HMMs. As for experiment 2, the 20-dimensional NLPCA features were used to compare with the original 91 features with varying block length.

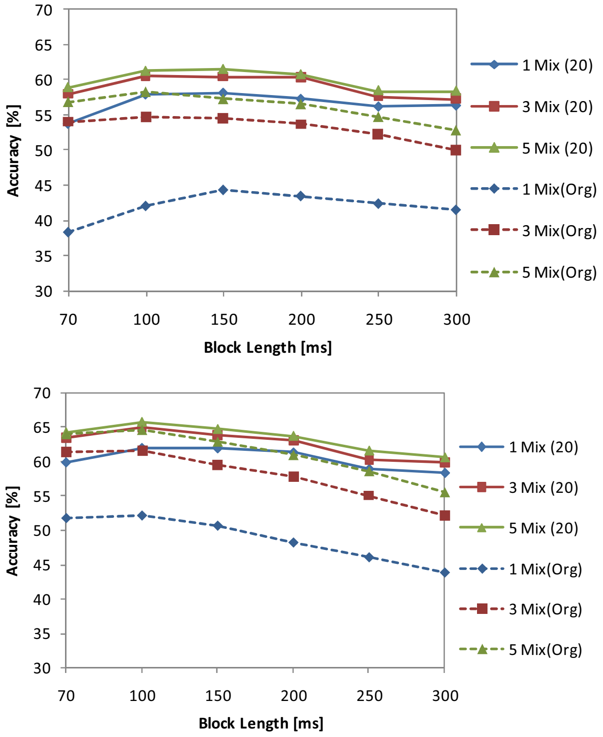

Figure 4: Recognition accuracies of 1-state (top panel) and 3-state HMMs (bottom panel) using the original and reduced features for various block lengths.

Figure 4 depicts the accuracies of 1 and 3 states HMMs with 1, 3 and 5 mixtures respectively. For the case of 1-state HMMs, the NLPCA features performed considerably better than the original features although the advantage decreases with an increasing number of mixtures. The highest accuracy of about 62% was obtained for the NLPCA features using a block length of 100 ms. The advantage of the NLPCA features is also shown for the 3-state HMM case. However, in contrast to the 1-state HMM case for which similar accuracies were obtained with the block lengths between 100 and 200 ms, the NLPCA features using a block length of 100 ms resulted in highest accuracy, presumably because of the state transitions in the HMM models and the shorter duration of each state for a 3-state model.

Comparing results from Figure 3, for which all the training data was used for both NLPCA training and HMM training, with the results in Figure 4 using the partitioned and effectively reduced set of training data, the best 1 state HMMs results are about 2% lower (61.48% vs. 63.99%) and the best 3 state results are slightly lower (65.58% vs. 65.73%). Thus there is only a small degradation in test accuracy due to the reduced size of the training set.

5. Conclusions

In this paper, a nonlinear feature transformation based on neural networks is presented. This feature transformation method is incorporated with an HMM recognition model for continuous speech phonetic recognition. In the method presented, the activations from the middle layer of a bottleneck neural network are used as reduced dimensionality features. The neural network is trained to recognize phonetic categories.

Experimental evaluation using the TIMIT database showed that recognition accuracies with NLPCA reduced dimensionality features are higher than recognition rates obtained with original features, especially when a small number of states and mixtures are used for HMM phonetic models. For phone recognition using a 3-state 5-mixture HMM, the accuracy obtained with the reduced 20-dimensional features was about 4% higher than that obtained with the original 91-dimensional features. Additionally, the NLPCA features are able to well represent spectral-temporal information in segments as long as 200 ms, thus potentially reducing HMM model complexity. Although the NLPCA training is relatively time consuming, the entire recognition system could benefit from low-dimensionality features both in terms of processing time and recognition accuracy.

6. References

[1] Bengio, Y., De Mori R., Flammia, G., and Kompe, R., “Global Optimization of a Neural Network – Hidden Markov Model Hybrid,” IEEE Trans. Neural Networks, 3(2), 1992.

[2] Chung, Y. J., and Un, C. K., “An MLP/HMM Hybrid Model using Nonlinear Predictors,” Speech Communication, 19(4), 1996.

[3] Duda, O. R., Hart, E. P., and Stork, G. D., *Pattern Classification*, Wiley-Interscience, New York, 2000.

[4] Hsieh, W. W., “Nonlinear Principal Component Analysis of Noisy Data,” Proc. IJCNN, 2006.

[5] Jang, C. S., and Un, C. K., “A New Parameter Smoothing Method in the Hybrid TDNN/HMM Architecture for Speech Recognition,” Speech Communication, 19(4), 1996.

[6] Lee, K.F., and Hon, H.W., “Speaker-Independent Phone Recognition Using Hidden Markov Models,” IEEE Trans. Acoust., Speech, Signal Processing, 37(11), 1989.

[7] Rigoll, G., Neukirchen, C., and Rottland, J., “A New Hybrid System based on MMI-neural Networks for the RM Speech Recognition Task,” Proc. ICASSP, 1996.

[8] Trentin, E., and Gori, M., “A Survey of Hybrid ANN/HMM Models for Automatic Speech Recognition,” Neurocomputing, 37(1), 2001.

[9] Zahorian, A. S., and Nossair, B. Z., “A Partitioned Neural Network Approach for Vowel Classification Using Smoothed Time/Frequency Features,” IEEE Trans. Speech and Audio Processing, 7(4), 1999.

[10] Zahorian, A. S., Singh, T., and Hu, H., “Dimensionality Reduction of Speech Features using Nonlinear Principal Components Analysis,” Proc. Eurospeech (Interspeech 2007), 2007.