DIMENSIONALITY REDUCTION METHODS FOR HMM PHONETIC 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

This paper presents two nonlinear feature dimensionality reduction methods based on neural networks for a HMM-based phone recognition system. The neural networks are trained as feature classifiers to reduce feature dimensionality as well as maximize discrimination among speech features. The outputs of different network layers are used for obtaining transformed features. Moreover, the training of the neural networks uses the category information that corresponds to a state in HMMs so that the trained networks can better accommodate the temporal variability of features and obtain more highly discriminative features in a low dimensional space. Experimental evaluation using the TIMIT database shows that recognition accuracies with the transformed features are slightly higher than those obtained with original features and considerably higher than obtained with linear dimensionality reduction methods. The highest phone accuracy obtained with 39 phone classes and TIMIT was 74.9% using a large number of training iterations based on the state-specific targets.

Index Terms— nonlinear discriminant analysis, neural networks, dimensionality reduction, HMMs

1. INTRODUCTION

Over the past two decades, there has been a lot of research effort devoted to combining HMMs and Neural Networks (NN) with a single, hybrid architecture, called hybrid NN/HMM speech recognition [1]. These hybrid systems attempt to take advantage of both HMMs and neural networks to improve flexibility and recognition performance. For instance, the hybrid system proposed by Bourlard and Morgan [2] applied a neural network to estimate the posterior probabilities of HMM states. Recently, the so-called TANDEM recognition approach introduced by Hermansky et al. [3] has shown a large improvement in recognition performance. This approach with neural networks and HMMs connected in tandem uses neural networks to obtain discriminative features as the input features for Gaussian Mixture Models (GMMs) of HMMs.

In this paper, we focus on the dimensionality reduction ability of neural networks and propose two neural network based NonLinear Discriminative Analysis (NLDA) transformations for a HMM-based phone recognition system. In contrast to many linear dimensionality reduction techniques including Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA), neural network based nonlinear transformation methods are able to form a dimensionally-reduced representation for complex data, while preserving variability and discriminability of the original data [4]. The outputs from different layers of a neural network are used as dimensionality reduced features for HMMs. In the first approach, which we refer to as NLDA1, the transformed features are produced from the final output layer of the network. In the second approach named NLDA2, the outputs of the middle hidden layer are used as transformed features. In addition, with the training independent of the HMM training, the neural network based feature transformations described in this paper could easily be combined with other processing methods.

The remainder of this paper is organized as follows: In Section 2, we describe the NLDA dimensionality reduction methods. Section 3 summarizes the evaluation of the proposed approaches using the TIMIT database with various neural network and HMM configurations. The conclusion is given in Section 4.

2. NLDA DIMENSIONALITY REDUCTION

2.1. NLDA methods

As illustrated in Fig. 1, the NLDA dimensionality reduction approach is based on a multilayer neural network and performs a nonlinear transformation for a lower dimensional representation of input features. The outputs of the network are further processed by PCA to create transformed features to be the inputs of an HMM recognizer.

The neural network used in NLDA includes an input layer, hidden layers and an output layer. The numbers of nodes contained in the input and output layers respectively correspond to the dimensions of the input features and the number of categories in the training target data.

NLDA1 obtains dimensionality reduced features at the output layer of a neural network. A linear output layer is used for the feature transformation in order to obtain output features which can be better modeled by HMMs, although all sigmoid nonlinear layers are used for the NLDA1 training.

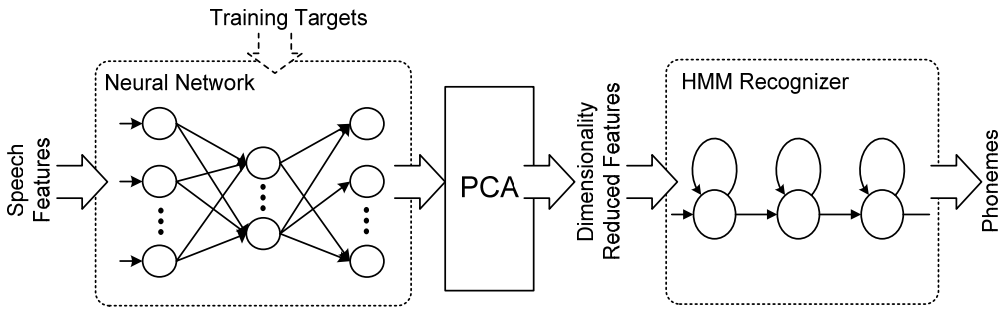

Fig. 1. Overview of the NLDA transformation for speech recognition

In contrast, NLDA2 uses the outputs of the middle hidden layer since the activations of the middle layer represent the internal structure of the input features. The dimensionality of the reduced feature space is determined only by the number of nodes in the middle layer. Therefore, an arbitrary number of reduced dimensions can be obtained, independent of the input feature dimensions and the nature of the training targets. Unlike NLDA1, all layers including the middle hidden layer are nonlinear in both feature transformation and training.

A PCA processing is applied to the output of the neural network as was used in [3]. PCA performs a Karhunen-Loeve (KL) transform in order to reduce the correlation of the network outputs and improve their match to a GMM. Since PCA itself is a good dimensionality reduction method, the dimension of the network outputs in NLDA1 is further reduced using PCA, while the PCA processing in NLDA2 is only used for feature decorrelation.

The transformed features are used as the inputs to a HMM with each state modeled as a GMM. Phoneme HMMs are used in this paper for phonetic recognition experiments.

2.3. Network training with state specific targets

The original features are scaled using the mean and standard deviation vectors of the training data so that all components have the same mean and variance as described in [5].

The training of the neural network requires the category information for creating training targets. In this paper, as a baseline, we use a number of output nodes equal to the number of phone categories, with a value of 1 for the target category and 0 for the non-target categories. The phone labeling information is the same as that used for the HMM training, thus 48 phone categories are used for the network training.

Due to the nonstationarity of speech signals, a speech signal varies even in a very short time interval, e.g. a phoneme. In order to accommodate this variability, multiple neural networks are employed in [6] and each corresponds to a state in a HMM. In contrast, in this paper a single neural network is trained using some "don't care" states for each phoneme model, so that one neural network trained with the targets can generate state dependent outputs.

As illustrated in Fig. 2, the phone-specific training target for a phone "Δ" in a simple two-phone example is expanded to include specific targets for each of the 3 states. In the training process, for each point in time, one state target will be considered as a "1," and the other two state targets for that point in time will be considered as a "don't care," and the state targets for all other categories will be consider as "0" value targets. As time progresses during a phone, the "1" moves from state 1 to state 2, to state 3. For the actual work reported in this paper, 48 phonemes were modeled with 3 state models, thus resulting in a neural network trained with 144 outputs rather than 48.

Training target for Δ { 1 0 }

State level targets for Δ

State 1: { 1 0 # 0 # 0 }

State 2: { # 0 1 0 # 0 }

State 3: { # 0 # 0 1 0 }

#: *Don't care*

Fig. 2. State level training targets

This strategy was based on the thinking that the features transformed by the neural network should be distinctive for each phoneme state, but the boundaries between states are likely to be indistinct. Thus "don't cares" are used in training so that there are no errors computed for the "don't" cares output nodes.

Two approaches are used to expand a phoneme level label to a state level label. The first approach uses a fixed state length ratio for all phonemes —assuming the first part of each phone is state 1, central section state 2, and last part state 3. The second one determines state boundaries by using the HMM-based Viterbi alignment, using an already trained HMM. The latter approach also provides global training between HMMs and neural networks by iteratively training the two components. For instance, a neural network is first trained with a fixed state length targets, then HMMs are trained, and then alignment based targets are used to train the neural networks. The HMM training and neural network training steps are iterated until some point of convergence is reached.

3. EXPERIMENTAL EVALUATION

3.1. Experimental setup

Several experiments based on the TIMIT database were conducted to investigate the two NLDA methods. The SA sentences were removed from the database, resulting in 3696 sentences for training and 1344 sentences for test. The original TIMIT 62 phone set was mapped to the reduced 48 phone set as described in [7]. There are seven groups where within-group confusions are not counted: {sil, cl, vcl, epi}, {el, l}, {en, n}, {sh, zh}, {ao, aa}, {ih, ix}, {ah, ax}. Thus, a total of 39 phone classes were used for the evaluation.

For all data, the modified Discrete Cosine Transformation Coefficients (DCTCs) and Discrete Cosine Series Coefficients (DCSCs) [8] were extracted as original features. DCTCs are used for representing speech spectra, and DCSCs are used to represent spectral trajectories. For all experiments, 13 DCTCs and 6 DCSCs were computed using 10 ms frames with a 2 ms frame spacing and a 1s block length, for a total of 78 DCTC-DCSC features.

Left-to-right Markov models with no skip were used and a total of 48 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.

A neural network with 3 hidden-layers (500-36-500 nodes) was used for determining feature transformations. The numbers of nodes in the input layer was 78 corresponding to the dimensionality of the original features. The output layer used 48 nodes for the phoneme level targets or 144 nodes for the state level targets.

3.2. Control experiment

The intent of this experiment was to establish a baseline for the evaluation of the proposed methods using the original DCTC-DSCS features as well as the PCA and LDA dimensionality reduced features. The original features were reduced to 20 and 36 dimensions both by PCA and LDA.

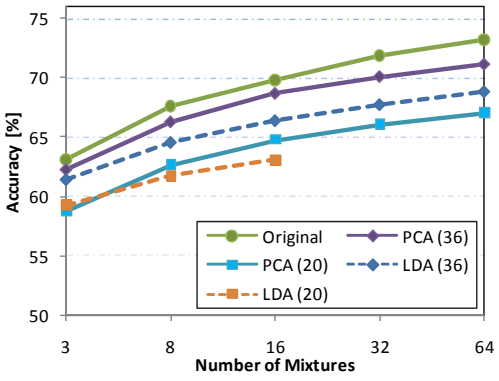

Fig. 3. Recognition accuracies using the original, PCA, and LDA features.

Fig. 3 shows recognition accuracies of the original and the reduced features using various numbers of mixtures in 3-state HMMs. The original features show the highest accuracy of 73.2% using 64-mixture HMMs. In all cases, the dimensionality reduced features lead to a degradation in accuracy. Compared to the original features, the PCA 36-dimensional features result in approximately 2% lower accuracy with 64 mixtures and 1% lower with 3 mixtures.

3.3. NLDA1 and NLDA2 experiments

Experiments were conducted to evaluate NLDA1 and NLDA2 using 36-dimensional reduced features. The number 36 was chosen based on pilot experiments, showing that typically highest accuracy was obtained with this number of features. The 48 phoneme level targets were used in the training of the network. The features which are direct outputs of the network without PCA processing were also evaluated.

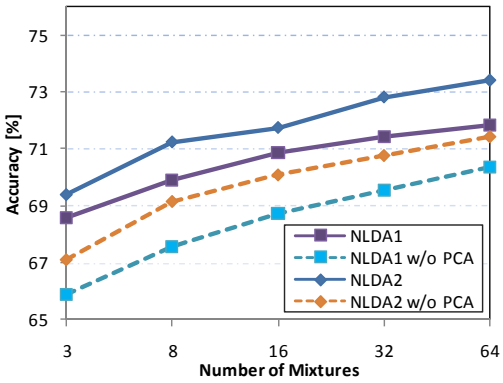

Fig. 4. Recognition accuracies using the NLDA dimensionality reduced features ("NLDA1" and "NLDA2") and the features without the PCA processing ("NLDA1 w/o PCA" and "NLDA2 w/o PCA").

Fig. 4 shows accuracies using 3-state HMMs with varying number of mixtures per state. NLDA2 performed better than NLDA1 for all conditions with about 2% higher accuracy. The NLDA2 transformed features resulted in the highest accuracy of 73.4% with 64 mixtures, which is slightly higher than the original features and considerably higher than the PCA and LDA features reported in Fig 3.

These results imply that the middle layer outputs of a neural network are able to better represent original features in a dimensionality-reduced space than are the outputs of the final output layer. The accuracies were also improved about 2% with PCA incorporated.

3.4. Experiments with state specific training targets

The NLDA methods trained with the state level targets were evaluated in this experiment. The network had 144 output nodes instead of 48 nodes used in 3.3 and was trained with a large number of iterations ($4x10^7$ weight updates).

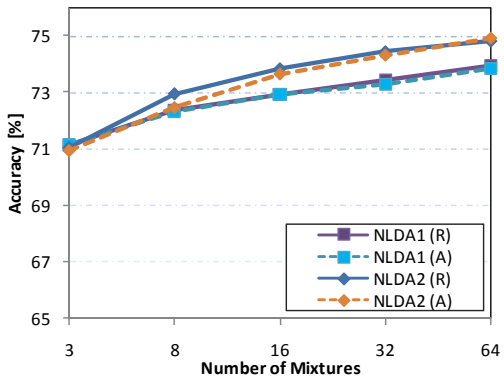

Fig. 5. Recognition accuracies of the NLDA dimensionality reduced features using the state level targets. "(R)" and "(A)" indicate the training targets obtained with the constant length ratio and forced alignment respectively.

Recognition accuracies of reduced 36-dimensional features using a constant state length ratio (ratio for 3 states: 1:4:1) and the Viterbi forced alignment for expanding the targets are shown in Fig. 5.

Both NLDA1 and NLDA2 using the expanded targets lead to approximate 2% higher in accuracy than using the phoneme level targets reported in Fig. 4. The use of forced alignment for state boundaries resulted in the highest accuracy of 74.9% with 64 mixtures, but slightly lower accuracies than using a fixed ratio of state lengths at the other conditions.

Comparing these results with those from Fig. 3, the NLDA2 features in a reduced 36-dimensional space achieved a substantial improvement versus the original features, especially when a small number of mixtures used. These results show the NLDA methods based on the state level training targets are able to obtain highly discriminative features in a dimensionality reduced space.

4. LITERATURE COMPARISON

Table 1 lists some recognition accuracies based on the TIMIT database as reported in the literature. The best result obtained in this paper is higher than all others, except for that of the Tandem NN [12] in which multiple neural networks and higher dimensional features were used.

Table 1. TIMIT results reported in literature

Study	Feature	Recognizer	Acc. (%)
Somervuo [9]	MFCC	HMM	68.5
Ketabdar et al. [10]	PLP	MLP-HMM	71.5
Pinto et al. [11]	LPC	HMM-MLP	74.6
Schwarz et al. [12]	MFCC	Tandem NN	78.5
Zahorian et al. [8]	DCTC/DCSC	HMM	73.9
This study	DCTC/DCSC	NN-HMM	74.9

5. CONCLUSIONS

Two feature dimensionality reduction methods based on neural networks were presented in this paper. In order to train neural networks with state dependent targets, "don't cares" are used where boundaries are not likely to be distinct, e.g., between states within a phone.

Experimental evaluation using TIMIT showed that very high recognition accuracies with the NLDA dimensionality reduced features were obtained, especially when using the outputs of network middle layer as in NLDA2. Recognition accuracies were improved using the state specific targets and a large number of iterations in the network training. The highest accuracy of 74.9% was obtained with the NLDA2 features using 3-state HMMs with 64 mixtures per state. These results showed that the presented methods are able to produce a low-dimensional effective representation of speech features, thus improving the performance of continuous speech recognition using HMMs.

6. REFERENCES

[1] Trentin, E., and Gori, M., "A Survey of Hybrid ANN/HMM Models for Automatic Speech Recognition," Neurocomputing, 37(1), 2001, pp.91-126.

[2] Bourlard, N., and Morgan, N., "Continuous Speech Recognition by Connectionist Statistical Methods," IEEE Trans. Neural Networks, 4(6), 1993, pp.893-909.

[3] Hermansky, H., Ellis, P.W. D., and Sharma S., "Tandem Connectionist Feature Extraction for Conventional HMM Systems," Proc. ICASSP, 2000, pp.1635-1638.

[4] Zahorian, A. S., Singh, T., and Hu, H., "Dimensionality Reduction of Speech Features using Nonlinear Principal Components Analysis," Proc. Interspeech, 2007, pp.1134-1137.

[5] Hu, H., and Zahorian, A. S., "A Neural Network Based Nonlinear Feature Transformation for Speech Recognition," Proc. Interspeech, 2008, pp.1533-1536.

[6] Chung, J. Y., and Un, K. C., "An MLP/HMM hybrid model using nonlinear predictors," Speech Communication, 19, 1996, pp.307-316.

[7] Lee, K. F., and Hon, H. W., "Speaker-Independent Phone Recognition Using Hidden Markov Models," IEEE Trans. Acoust., Speech, Signal Proc., 37(11), 1989, pp.1641-1648.

[8] Zahorian, A. S., Hu, H., Chen, Z., and Wu, J., "Spectral and Temporal Modulation Features for Phonetic Recognition," Proc. Interspeech, 2009.

[9] Somervuo, P., "Experiments with Linear and Nonlinear Feature Transformations in HMM based Phone Recognition," Proc. ICASSP, 2003, Vol. I, pp.52-55.

[10] Ketabdar, H., and Bourlard, H., "Hierarchical Integration of Phonetic and Lexical Knowledge in Phone Posterior Estimation," Proc. ICASSP, 2008, pp.4065-4068.

[11] Pinto, J., and Hermansky, H., "Combining Evidence from a Generative and a Discriminative Model in Phoneme Recognition," Proc. Interspeech, 2008, pp.2414-2417.

[12] Schwarz, P., Matejka, P., and Cernocky J., "Hierarchical Structures of Neural Networks for Phoneme Recognition," Proc. ICASSP, 2006, Vol. 1, pp. 325-328.