HIERARCHICAL TANDEM FEATURE EXTRACTION

Sunil Sivadas[1] *and Hynek Hermansky*[1,2]

[1]Oregon Graduate Institute of Science and Technology, Portland, Oregon, USA.
[2]International Computer Science Institute, Berkeley, California, USA.
email: {sunil,hynek}@ece.ogi.edu

ABSTRACT

We present a hierarchical architecture for tandem acoustic modeling. In the tandem acoustic modeling paradigm a Multi Layer Perceptron (MLP) is discriminatively trained to estimate phoneme posterior probabilities on a labeled database. The outputs of the MLP after nonlinear transformation and whitening are used as features in a Gaussian Mixture Model (GMM) based recognizer. In this paper we replace the large monolithic MLP with hierarchies of MLP experts .We apply this approach on Speech in Noisy Environments (SPINE1) evaluation conducted by the Naval Research Laboratory (NRL). We observe a reduction in word error rate of 30% with context-independent models and 5% WER with context-dependent models relative to PLP features.

1. INTRODUCTION

In the tandem approach [1, 2, 3] a MLP classifier is first trained to estimate the context-independent phoneme posterior probabilities. The probability vectors are gaussianised and decorrelated and used as features for GMM system. The MLP and GMM are trained independently. This approach performed best in the ETSI Aurora evaluation [4], a continuous digit recognition task in noisy environments and achieved significant reduction in word error rate with context-independent models in large vocabulary SPINE1 evaluation [3].

Modular and hierarchical neural networks have been studied extensively in pattern recognition literature [5, 6]. These networks divide the overall classification task among several networks. The decisions from networks are combined in a hierarchical manner to arrive at the overall network output. Thus the task of classifying a global set of classes, context-independent phonemes in the case of tandem approach, is divided into subsets. The partition is based on prior knowledge about the task or by data-driven clustering algorithms. For example, natural choice of first partitioning in the case of phonemes will be speech and silence.

The hierarchical systems have shorter training times and can have fewer parameters than the monolithic neural networks. This technique has been applied to build connectionist acoustic models [7]. In this paper we investigate the effectiveness of hierarchical approach in feature extraction under tandem framework. This is implemented as hierarchies of MLPs. We make soft splits of data using soft classification trees. This is based on the statistical method of factoring posteriors [8] which is explained in the next section. Section 3 describes the design of hierarchical tree. Experimental results are presented in section 4. In section 5 we discuss the results, followed by our conclusions.

2. HIERARCHICAL CLASSIFICATION

2.1. Factoring Posterior Probabilities

Let L denote the set of classes λ_k to be discriminated. Consider the partition of L into M disjoint and non-empty subsets L_i such that members of L_i are least confused with members of L_j $(\forall j \neq i)$. A particular class λ_k will now be a member of L and only one of the subsets L_i. Therefore, we can rewrite the posterior probability of class λ_k as a joint probability of the class and the corresponding subset L_i and factor it according to

$$\begin{aligned} p(\lambda_k|\mathbf{x}) &= p(\lambda_k, L_i|\mathbf{x}), \lambda_k \in L_i \\ &= p(L_i|\mathbf{x})p(\lambda_k|L_i, \mathbf{x}). \end{aligned}$$

Thus, the global task of discriminating between all the classes in L has been converted into discriminating subsets L_i and independently discriminating the classes λ_k remaining within each of the subsets L_i. Recursively repeating this process yields a hierarchical tree-organized structure. The posterior probability for a specific class can be computed by multiplying all the conditional posteriors from root node to the leaf corresponding to the specific class.

Conditional node posteriors can be estimated by restricting the training set of the corresponding MLP to the subset L_i on which the probability is conditioned. Thus the training data for each node is shared among all its child nodes according to the partitioning of classes and the amount of training data decreases with increase in specialization. Due to the diminishing training data as we traverse down the tree

0-7803-7402-9/02/$17.00 ©2002 IEEE

and the errors in posterior estimation, the the design of hierarchical structure become crucial.

3. HIERARCHICAL TANDEM SYSTEM

3.1. Hierarchical Tree Structure

If all the nodes in the tree would compute true conditional posteriors, the tree structure would have no influence on the classifier performance because any kind of factoring yields an exact decomposition of the class posteriors.Since this not true in practice, the choice of tree structure is important. Due to the large number of choices at each node it is impossible to find an optimal structure through an exhaustive search. Hence we apply evidence from data and heuristics to design the tree structure.

In speech recognition the obvious first partitioning is speech and silence. At the root of the tree we discriminate speech and background noise. This is motivated by the observation that these classes are easy to distinguish acoustically. The speech subset is further split into voiced and unvoiced classes. The leaf nodes of the tree compute monophone posteriors conditioned on voiced and unvoiced classes. Figure 1 shows the topology of the hierarchy. In this paper we design a hierarchical tree with three levels. Table 1 shows the hierarchical splitting of classes. "Tandem 0" system is the basic tandem system with single MLP. "Tandem 1" has two levels of hierarchy and "Tandem 2" has three levels of hierarchy.

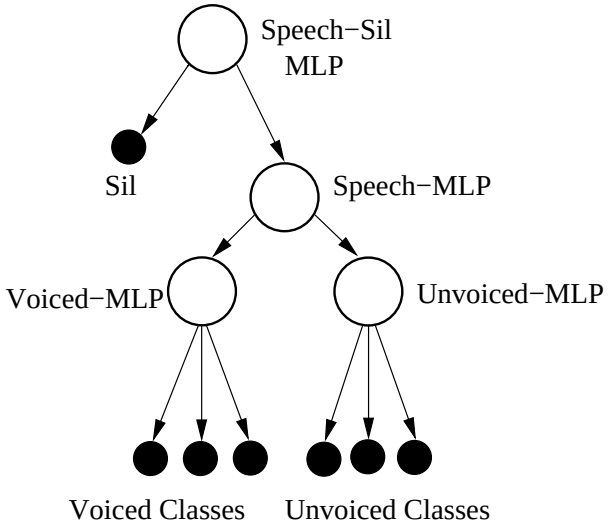

Fig. 1. Hierarchies of MLPs

Classifier	Hierarchy	Classes
Tandem 0	0	monophones + sil
Tandem 1	0	speech - sil
	1	monophones
Tandem 2	0	speech - sil
	1	voiced - unvoiced
	2	voiced classes
	2	unvoiced classes

Table 1: Hierarchical splitting of classes.

3.2. Postprocessing of Posteriors

The posterior probabilities have a skewed distribution, making them harder to be modeled by mixture of Gaussian components. Different postprocessing methods to warp the posteriors into a different domain has been tried [1]. Replacing the softmax nonlinearity at the output layer with a linear function is shown to make the distribution more Gaussian. This retains the rank ordering of posteriors. In the hierarchical architecture the class posterior probabilities are computed by a cluster of MLPs. We cannot remove the softmax nonlinearity of each of them as this does not retain the ranking of posteriors. Removing the softmax is equivalent to the logarithm of the posteriors with a normalization constant.

$$l_i = log(p(C_i|\mathbf{x})) - K, 0 \leq i \leq N-1$$

where l_i is the linear output corresponding to class C_i and $K = log(\sum_{i=0}^{N-1} exp(l_i))$. Since we have no means of obtaining K from $p(C_i|\mathbf{x})$ we approximate it by the average of the log posteriors.

$$K = \frac{1}{N} \sum_{i=0}^{N-1} log(p(C_i|\mathbf{x})), 0 \leq i \leq N-1$$

The distribution of resulting features is found to be similar to the one obtained by removing the softmax. Diagonalization of the global covariance matrix of the features by Karhunen-Loeve (KL) transformation improves the performance because the GMM assumes that features are uncorrelated. We retain all the feature components after KL transformation.

4. EXPERIMENTAL EVALUATION

4.1. System Description

We tested the hierarchical system on SPINE1 task [3]. The task focuses on transcribing speech produced in noisy environments with emphasis on noisy military environments. It involves a medium-sized vocabulary of about 5000 words. The data consists of conversations between two communicators working on a collaborative, Battleship-like task in which they seek and shoot at targets.Each person is seated

in a sound chamber in which a previously recorded military background noise environment is accurately reproduced. The speech is sampled at 16KHz.

Perceptual Linear Prediction (PLP) cepstral features are extracted from a frame of 25 ms of speech, every 10ms. The feature vector consists of 13 PLP coefficients augmented by deltas and double-deltas. They are then normalized over the utterance to zero mean and unit variance. The input to each MLP is a window of 9 successive feature vectors.

The labels for training MLP are generated by the process of forced alignment as explained in [3]. From ICSI56 context-independent phoneme set a subset of 50 phonemes occurring in SPINE1 data was derived. Each MLP in the hierarchy is trained by backpropagation with a minimum-cross-entropy criterion to 'one-up' targets obtained from the labels. The outputs from the MLPs are fed to the GMM system after the postprocessing. The GMM system is trained according to the standard EM algorithm. We used CMU SPHINX-III recognizer with 3 states per context-dependent triphone with 2600 tied states, each modeled by a mixture of 8 Gaussians.The context-independent phonemes are also modeled using 3 state HMMs with 8 Gaussians per state. The tandem MLP and GMM are trained independently and use different number of context-independent phonemes. Table 2 shows the architecture of each MLP in the hierarchy.

Classifier	Hierarchy	No. of Classes	IU	HU	OU
Tandem 0	0	50	351	1000	50
Tandem 1	0	2	351	500	2
	1	49	351	750	49
Tandem 2	0	2	351	500	2
	1	2	351	500	2
	2	37	351	500	37
	2	12	351	500	12

Table 2: Architecture of MLPs in the hierarchy. IU stands for number of input units, HU for hidden units and OU for output units.

4.2. Results

The SPHINX system was trained on 8 hours of data. Models were trained for three tandem features and the PLP features. Recognition was performed on 9 hours of evaluation data. The word error rates for all the systems are shown in table 3. The recognizer was not tuned to improve the performance of individual systems.

It can be seen that all the tandem systems outperform the PLP system when CI models are used for decoding. The word error rates are 30%, 27% and 26% lower than PLP system for Tandem 0, Tandem 1 and Tandem 2 respectively. The increase in error rate with hierarchical structure is discussed in the next section.

The performance of the systems tend to converge when context-dependent models are used for recognition. The tandem systems are marginally better than the PLP system. However, Tandem 2 system performs 3% better than the Tandem 0 system and 5% better than the PLP system. This reversal of trend in performance of the tandem system compared to context-independent models is discussed in the following section.

Type of feature	Dimensions	CI	CD
PLP with Δ and Δ^2	39	71.6	39.1
Tandem 0	50	50.5	38.3
Tandem 1	50	52.0	38.2
Tandem 2	50	53.5	37.1

Table 3: Word error rates (%) with SPHINX-III system for various feature sets. CI stands for context-independent and CD for context-dependent.

5. DISCUSSION

We find that the tandem systems perform significantly better than the PLP system with context-independent models whereas only marginal improvement is obtained with context-dependent models. It is observed that the hierarchical tandem system perform marginally better than the monolithic classifier based system with context-dependent models and worse when context-independent models are used.

In [3] we interpreted the MLP in tandem modeling as a transformation of the feature space that magnifies regions around phonetic boundaries and suppressing the non-phonetic variability due to speaker and noise within the region corresponding to class. In the tandem approach we train the MLP to maximize the separability of context-independent phonemes with a block of 9 successive frames of feature vectors as input. The target phoneme corresponds to the frame at the center of the window. This introduces shift-invariance and suppresses the context, speaker and environmental variability. Thus there is little information left to be modeled by context-dependent GMMs. This explains why the advantages obtained by context-independent models did not carry over to context-dependent models.

Dividing the data reduces the bias of an estimator, but it generally increases the variance [5]. We make "soft" splits of data i.e., allowing the data to lie simultaneously in multiple regions. This has a variance decreasing effect since many classifiers contribute to the final output. The increase in word error rate of the hierarchical tandem system compared to the single MLP with context-independent models may be due the effective increase in variance. The low complexity context-independent modeling is unable to model this additional variability caused by data splitting. But this may be helping the context-dependent modeling, as can be seen from table 3.

It can be observed from table 2 that the number of parameters in Tandem 3 system is half of that in Tandem 0.

This has reduced the training time and system complexity without affecting the performance.

To investigate further the reasons for the disparity in improvements from context-independent and context-dependent models we tested the features with GMMs of varying complexity. Figure 2 shows the performance curve of PLP system and Tandem 3 system for different number of Gaussians/state keeping the number of states per model unit same (=3). Increasing the number of Gaussians gives GMM additional parameters to model the variability in feature space within each phoneme. It can be seen from the figure that the performance of the tandem system and PLP system tend to converge with increasing number of Gaussians/state. The word error rate of PLP system reduced by 25% from 1 Gauss/state to 8 Gauss/state whereas the tandem system improved by only 10%. This is consistent with the earlier observation that the tandem MLP suppresses the context variability within the phoneme.

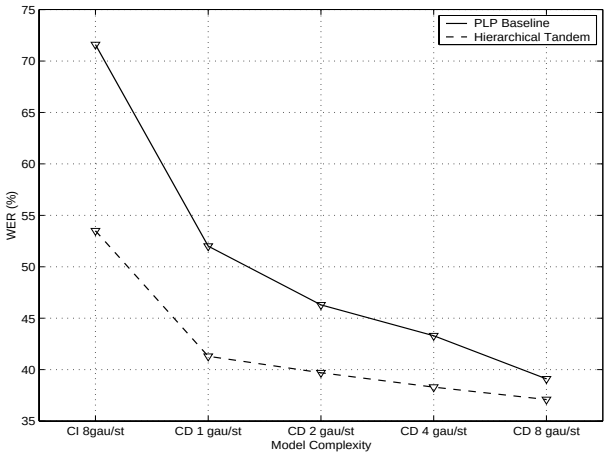

Fig. 2. Word error rates (%) of feature sets vs. Complexity of models

6. CONCLUSIONS

We presented a preliminary study on hierarchical feature extraction under tandem framework using a MLP tree. Hierarchical modeling offers a power method of combining multiple classifiers into a tree structure. We have shown that it achieves comparable word error rates to a monolithic MLP, with fewer parameters. The design of tree was based on the prior knowledge of classes. A much more structured approach to the design of classifier tree could improve the performance. Although we achieved significant reduction in word error rate with context-independent models further work needs to be done to extend this performance to context-dependent models. We conclude that tandem modeling approach offers considerable advantages for low complexity systems with few subword classes especially when signal to noise ratio is low.

7. ACKNOWLEDGMENT

The research was supported by DARPA under R16007-01 and by an industrial grant from Qualcomm Inc. We thank Rita Singh of CMU for making the SPHINX-III system available and for the insights to improve its performance. We thank Stephane Dupont of ICSI for helpful discussions.

8. REFERENCES

[1] H. Hermansky, D. Ellis and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems", *in Proc. ICASSP'00,* Istanbul, Turkey, June 2000.

[2] S. Sivadas, P. Jain and H. Hermansky, "Discriminative MLPs in HMM-based recognition of speech in cellular telephony", *in Proc. ICSLP'00,* Beijing, China, October 2000.

[3] D.W.P. Ellis, R. Singh and S. Sivadas, "Tandem Acoustic Modeling in Large-Vocabulary Recognition", *in Proc. ICASSP'01,* Salt Lake, City, Utah, USA, May 2001.

[4] S. Sharma, D. Ellis, S. Kajarekar, P. Jain and H. Hermansky, "Feature Extraction using non-linear transformation for robust speech recognition on the Aurora database", *in Proc. ICASSP'00,* Istanbul, June 2000.

[5] M.I. Jordan and R.A. Jacobs, "Hierarchical Mixtures of Experts and the EM algorithm", *in Neural Computation,* vol.6, pp.181-214, 1994.

[6] M.I. Jordan and R.A. Jacobs, "Modular and Hierarchical Learning Systems", *in M.A. Arbib (Ed) The Handbook of Brain Theory and Neural Networks,* pp 579-581, 1995.

[7] J. Fritsch, *Modular Neural Networks for Speech Recognition,* Diploma Thesis, July 1996, Carnegie Mellon University, Pittsburgh, USA.

[8] J. Schuermann and W. Doster, "A Decision Theoretic Approach to Hierarchical Classifier Design", *in Pattern Recognition,* 17, 3, pp.359-369, 1984.