Hierarchies of Neural Networks for Connectionist Speech Recognition

Jürgen Fritsch, Alex Waibel

Interactive Systems Laboratories

University of Karlsruhe
76128 Karlsruhe, Germany

Carnegie Mellon University
Pittsburgh, PA 15213, USA

Abstract. We present a principled framework for context-dependent hierarchical connectionist HMM speech recognition. Based on a divide-and-conquer strategy, our approach uses an *A*gglomerative *C*lustering algorithm based on *I*nformation *D*ivergence (ACID) to automatically design a soft classifier tree for an arbitrary large number of HMM states. Nodes in the classifier tree are instantiated with small estimators of local conditional posterior probabilities, in our case feed-forward neural networks. Our framework represents an effective decomposition of state posteriors with advantages over traditional acoustic models. We evaluate the effectiveness of our *H*ierarchies of *N*eural *N*etworks (HNN) on the Switchboard large vocabulary conversational speech recogntion (LVCSR) corpus.

1. Introduction

In hybrid NN/HMM connectionist speech recognition, parametric mixture densities that are typically applied to model observation probabilities in hidden Markov models (HMM) are replaced by connectionist estimators of posterior probabilities. Experiments with such systems (e. g. [1]) indicated an advantage of hybrid models in terms of discriminative power, required number of parameters and decoding speed. However, despite the success of such models in a wide range of speech recognition tasks, current state-of-the-art systems for large vocabulary conversational speech recognition (LVCSR) almost entirely rely on the conventional paradigm for acoustic modeling. What are the reasons for this preference towards traditional acoustic models?

First, training of connectionist acoustic models usually is computationally more expensive. Second, context modeling in continuous density HMMs has evolved significantly since the advent of hybrid NN/HMM models. The application of decision trees to the clustering of polyphones recently led to systems consisting of thousands of HMM states. Since modeling of observation probabilities using mixture densities is independent for each state, an increase in the number of states imposes no conceptual problem. In contrast, connectionist acoustic models jointly estimate posterior state probabilities and are much harder to scale to larger systems. Often, context-modeling is avoided at all. Nevertheless, significant improvements in recognition accuracy can be gained through context modeling in both traditional and connectionist acoustic modeling [3, 5, 7]. However, the number of HMM states and therefore the level of context-dependence has been limited to medium size systems.

This paper presents the ACID/HNN [4] framework, a highly modular and scalable approach to connectionist acoustic modeling. Viewing the estimation of posterior state probabilities as a hierarchical process, an automatically clustered tree structured ensemble of neural networks is applied to estimate state posteriors. Although similar in spirit, earlier approaches [5, 6, 9] lack a principled treatment of decomposition. We present experiments on the Switchboard LVCSR corpus, demonstrating that state-of-the-art performance can be achieved with our framework.

2. Hierarchical Acoustic Modeling

Connectionist acoustic modeling for hybrid NN/HMM systems is characterized by the estimation of posterior state probabilities using one or several neural networks. Integration of this model into the HMM framework is justified by the application of Bayes rule

$$p(\mathbf{x}|s_i) = \frac{p(s_i|\mathbf{x})}{P(s_i)}\, p(\mathbf{x})$$

to get estimates of the state observation likelihood $p(\mathbf{x}|s_i)$ given an acoustic feature vector $\mathbf{x}$. Usually, the term $p(\mathbf{x})$ is neglected because it is constant for all states and does not influence the outcome of a Viterbi decoder. Therefore, scaled observation likelihoods can be computed from state posteriors by dividing by state priors $P(s_i)$. For context-independent systems, the number of HMM states is small enough to apply a single neural network to jointly estimate the posterior state probabilities. However, introducing context-dependence increases the number of states significantly and training a single neural network becomes prohibitive. A decomposition can be gained by factoring the posterior state probabilities [3, 5, 7]. Typically, posterior state probabilities are factored according to monophone identity. Here, we present a more principled approach where factoring is guided by an agglomerative clustering process.

Let S denote the set of all (decision tree clustered) HMM states s_k. Consider a partition of S into M disjoint and non-empty subsets S_i. A particular state s_k will now be a member of S and exactly one of the subsets S_i. Therefore, we can rewrite the posterior probability of state s_k as a joint probability of state and appropriate subset S_i and factor it according to

$$\begin{aligned} p(s_k|\mathbf{x}) &= p(s_k, S_i|\mathbf{x}) \qquad \text{with} \qquad s_k \in S_i \\ &= p(S_i|\mathbf{x})\, p(s_k|S_i, \mathbf{x}) \end{aligned}$$

Thus, the global task of discriminating between all the states in S has been converted into (1) discriminating between subsets S_i and (2) independently discriminating between the states s_k contained within each of the subsets S_i. Recursively repeating this process yields a hierarchical tree-organized structure (see Fig. 1). The effectiveness of any such hierarchical decomposition of posteriors crucially depends on the tree design method [8] since local estimators of conditional posterior probabilities can only be trained to approximate the true distributions.

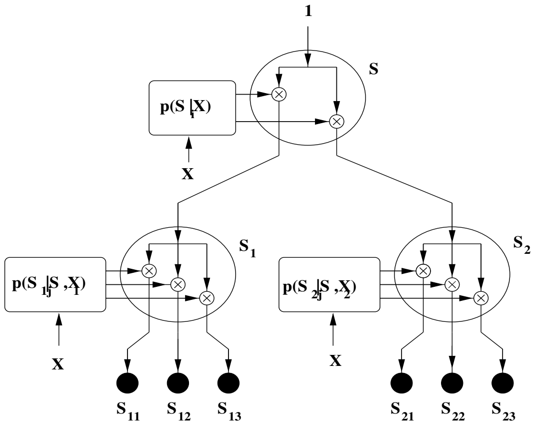

Figure 1: Conditional Factoring of Posteriors

3. The ACID/HNN Framework

When dealing with a rather large number of classes, several thousands in our case, evaluation of all possible configurations for a hierarchical decomposition of the posterior class probabilities becomes intractable. Also, common heuristic top-down approaches based on examination of the class confusion matrix of pre-trained monolithic classifiers are problematic. We therefore apply an agglomerative (bottom-up) clustering scheme using the symmetric information divergence

$$d(s_i, s_j) \quad = \quad \int_{\mathbf{x}} (p(\mathbf{x}|s_i) - p(\mathbf{x}|s_j)) \log \frac{p(\mathbf{x}|s_i)}{p(\mathbf{x}|s_j)} \, d\mathbf{x}$$

as a measure of acoustic dissimilarity of subphonetic units. Based on this rather inexpensive distance measure, even large amounts of subphonetic units can be clustered efficiently. We typically model the class-conditional likelihoods using single diagonal covariance multivariate Gaussians with mean vectors μ_i and variance vectors σ_i^2. In this case, the symmetric information divergence between two states s_i and s_j amounts to

$$d(s_i, s_j) = \frac{1}{2} \sum_{k=1}^{n} \frac{(\sigma_{jk}^2 - \sigma_{ik}^2) + (\sigma_{ik}^2 + \sigma_{jk}^2)(\mu_{ik} - \mu_{jk})^2}{\sigma_{ik}^2 \sigma_{jk}^2}$$

Making the simplifying assumption of linearity of information divergence, we can define the following distance measure between clusters of states S_k and S_l

$$D(S_k, S_l) = \sum_{s_i \in S_k} p(s_i|S_k) \sum_{s_j \in S_l} p(s_j|S_l) d(s_i, s_j)$$

The ACID algorithm uses the above distance measure in a standard bottom-up agglomerative clustering method. Note that this algorithm clusters HMM states without knowledge of their phonetic identity solemnly based on acoustic dissimilarity. Fig. 2 illustrates ACID clustering on a very small subset of initial clusters. The ordinate of the dendrogram plot shows the information divergence at which the merger occured. Names encode monophone, state (begin,middle,end) and context id (numeric).

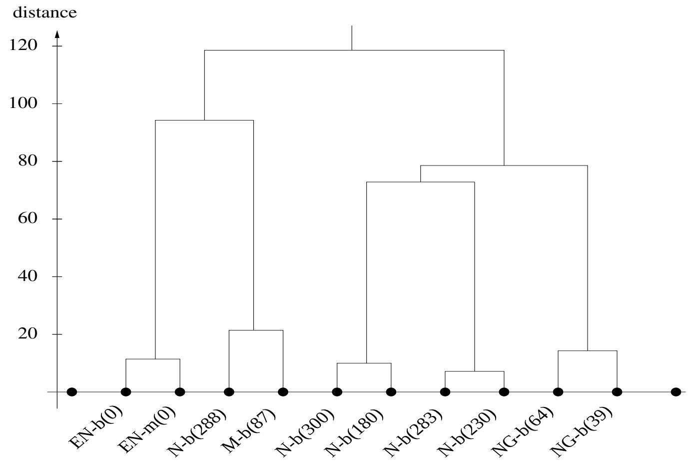

Figure 2: Partial Dendrogram of ACID Clustering

Each node in an ACID-clustered tree structure represents conditional posteriors when interpreted as a hierarchical decomposition. Estimators such as polynomial regressors, radial basis functions or feed-forward networks can potentially be trained to estimate such posteriors. We are currently experimenting with 2-layer MLPs, trained in the framework of a generalized EM algorithm using error backpropagation. Therefore, we term the complete connectionist acoustic model a Hierarchy of Neural Networks (HNN), see Fig. 3.

Challenging aspects of such an architecture are model complexity and adaptation of learning rates during training. While the network in the root node is trained on all of the training data, networks deeper down the tree receive less training data than their predecessors. We found that it is advantageous to reduce the number of networks in an HNN by applying a greedy bottom-up node merging algorithm as a second step of ACID clustering. Using this strategy, we typically increase the average arity of the HNN tree from 2 to about 8.

4. Experiments

Experiments with the ACID/HNN approach were carried out on the Switchboard LVCSR corpus. We chose Switchboard, because it consists of very noisy spontaneous speech in telephone quality requiring excessive modeling of coarticulation to achieve state-of-the-art performance. Switchboard also is a comparably hard speech recognition task. Current best systems based on traditional HMM approaches achieve word error rates in the vicinity of 30-40% while typically running 150-300 times slower than real time.

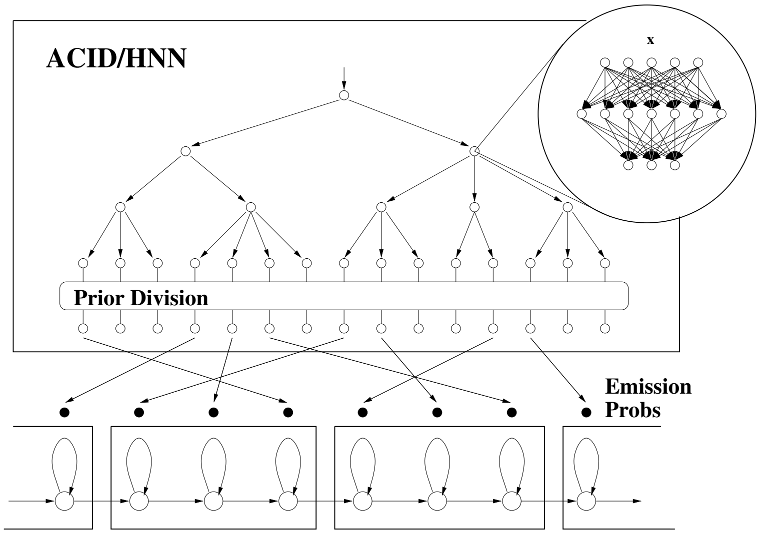

Figure 3: ACID clustered Hierarchy of Neural Networks

The following table summarizes results for various hybrid NN/HMM models focusing on the ACID/HNN framework. The models were trained on 170 hours of Switchboard training data corresponding to roughly 60 million patterns. Recognition experiments were performed with the Janus-RTk [2] Switchboard recognizer on the 1997 development test set, consisting of 40 unseen speakers. The first two rows give earlier results that we obtained with hybrid HME/HMM models [5]. CI denotes context-independent, CD context-dependent modeling. Apart from word error rates, the table gives number of HMM states, effective number of evaluated networks per frame, number of parameters and real time factors for each system.

acoustic model	# states	# NNs	# params	xRT	word error
CI HME/HMM	166	59	220k	80	58.6 %
CD HME/HMM	10000	224	1.2M	130	37.3 %
CD ACID/HNN	**6000**	**962**	**1.6M**	**120**	**35.7 %**
CD ACID/HNN	**24000**	**4046**	**2.8M**	**145**	**33.3 %**
adapted ACID/HNN	**24000**	**4046**	**2.8M**	**130**	**31.8 %**
pruned ACID/HNN	**24000**	**$\approx$500**	**2.8M**	**26**	**33.6 %**

Obviously, context-dependent modeling improves performance vastly. We trained two ACID/HNN acoustic models with 6k and 24k tied states, respectively, to demonstrate the scalability of the proposed approach. Furthermore, our results indicate that going from 6k to 24k HMM states improves performance significantly. Unsupervised speaker adaptation can be applied very easily to our model by simply retraining those networks in the HNN that receive more than a certain amount of adaptation data (the ones at the top of the tree). An additional gain of 1.5% in accuracy was achieved using this simple algorithm. Finally, ACID/HNN models allow to trade off accuracy against decoding speed by simply pruning the evaluation of the HNN tree in each frame based on partial posteriors. This way, a significant amount of network evaluations can be omitted with almost no loss in accuracy. In contrast, traditional acoustic models usually require much more effort to achieve the same goal.

5. Conclusions

We present a novel framework for connectionist acoustic modeling and demonstrate its viability on the Switchboard LVCSR task. Based on the principle of divide and conquer, it allows to build and robustly estimate connectionist acoustic models for arbitrary large sets of context-dependent HMMs. Our approach maintains the advantages of discriminatively trained acoustic models while circumventing the limitations of standard hybrid NN/HMM architectures. On the 1997 Switchboard development test set, we achieve a competitive word error rate of 31.8% with an ACID/HNN based acoustic model. Furthermore, our approach simplifies important algorithms such as speaker adaptation and scoring speed-up.

References

[1] H. Bourlard and N. Morgan, *Connectionist Speech Recognition – A Hybrid Approach*, Kluwer Academic Press, 1994.

[2] M. Finke, J. Fritsch, P. Geutner, K. Ries and T. Zeppenfeld, "The JanusRTk Switchboard/Callhome 1997 Evaluation System", *Proceedings of LVCSR Hub5-e Workshop*, Baltimore 1997.

[3] H. Franco, M. Cohen, N. Morgan, D. Rumelhart and V. Abrash, "Context-dependent connectionist probability estimation in a hybrid Hidden Markov Model – Neural Net speech recognition system", *Computer Speech and Language*, Vol. 8, No 3, 1994.

[4] J. Fritsch, "ACID/HNN: A Framework for Hierarchical Connectionist Acoustic Modeling", In *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, 1997.

[5] J. Fritsch, M. Finke and A. Waibel, "Context-Dependent Hybrid HME/HMM Speech Recognition using Polyphone Clustering Decision Trees", *Proc. of ICASSP'97*, Munich 1997.

[6] J. Hampshire II, A. Waibel, "The Meta-Pi Network: Building Distributed Knowledge Representations for Robust Pattern Recognition", *Tech. Rep. CMU-CS-89-166*, Carnegie Mellon University, Pittsburgh PA, August 1989.

[7] D. J. Kershaw, M. M. Hochberg and A. J. Robinson, "Context-Dependent Classes in a Hybrid Recurrent Network HMM Speech Recognition System", *Tech. Rep. CUED/F-INFENG/TR217*, CUED, Cambridge, England 1995.

[8] J. Schürmann and W. Doster, "A Decision Theoretic Approach to Hierarchical Classifier Design", *Pattern Recognition 17 (3)*, 1984.

[9] A. Waibel, H. Sawai and K. Shikano, "Consonant Recognition by Modular Construction of Large Phonemic Time-Delay Neural Networks", *Proc. of ICASSP'89*, Glasgow 1989.