ACID/HNN: CLUSTERING HIERARCHIES OF NEURAL NETWORKS FOR CONTEXT-DEPENDENT CONNECTIONIST ACOUSTIC MODELING

Jürgen Fritsch, Michael Finke

Interactive Systems Labs
University of Karlsruhe, Germany and Carnegie Mellon University,USA
fritsch@ira.uka.de, finkem@cs.cmu.edu

ABSTRACT

We present the ACID/HNN framework, a principled approach to hierarchical connectionist acoustic modeling in large vocabulary conversational speech recognition (LVCSR). Our approach consists of an Agglomerative Clustering algorithm based on Information Divergence (ACID) to automatically design and robustly estimate Hierarchies of Neural Networks (HNN) for arbitrarily large sets of context-dependent decision tree clustered HMM states. We argue that a hierarchical approach is crucial in applying locally discriminative connectionist models to the typically very large state spaces observed in LVCSR systems. We evaluate the ACID/HNN framework on the Switchboard conversational telephone speech corpus. Furthermore, we focus on the benefits of the proposed connectionist acoustic model, namely exploiting the hierarchical structure for speaker adaptation and decoding speed-up algorithms.

1. INTRODUCTION

A few years back, several researchers (e.g. [1]) succesfully experimented with neural networks as probabilistic estimators for hidden Markov models. This approach is often termed hybrid NN/HMM since the usual parametric mixture densities to model HMM observation probabilities are replaced by connectionist estimators of posterior probabilities. The experiments indicated an advantage of hybrid models in terms of discriminative power, required number of parameters and decoding speed. These results led to extensive research in connectionist acoustic modeling for HMM systems with promising and sometimes superior results compared to conventional acoustic modeling. However, despite the success in applying connectionist acoustic modeling to a wide range of speech recognition tasks, current state-of-the-art systems for large vocabulary conversational speech recognition (LVCSR) on corpora such as Switchboard and Broadcast News almost entirely rely on the conventional paradigm for acoustic modeling (with the exception of [2, 3]). What are the reasons for this preference towards traditional acoustic models?

First, training of connectionist acoustic models usually is computationally very expensive. Compared to mixture density models, training times often are orders of magnitude higher for neural networks. This fact is most obvious in LVCSR tasks, where the largest amounts of training data are available. Furthermore, in contrast to mixtures of Gaussians, connectionist acoustic models are mostly trained with on-line stochastic gradient optimization techniques that require hand-optimization of learning parameters such as gain and momentum factor.

Second and much more critical, context modeling with traditional continuous density HMMs has evolved significantly since the advent of hybrid NN/HMM models. The application of decision trees to the clustering of tri-, quint- and even septphones 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. Factoring of state posteriors according to monophone and context identity can be applied to modularize the connectionist estimation process into several networks [4, 6, 8]. However, the number of HMM states and therefore the level of context-dependence that can be modeled with such an approach is limited.

What is missing for connectionist LVCSR is a principled approach that scales well to the large number of HMM states that are typically required to achieve competitive performance. This paper presents the ACID/HNN [5] framework which aims at providing such an approach. 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 . We give results on the Switchboard LVCSR corpus and experimentally demonstrate how the hierarchical structure of such an acoustic model can directly and efficiently be exploited for purposes such as speaker adaptation and decoding speed-up.

2. HIERARCHICAL CONNECTIONIST MODELING

Connectionist acoustic modeling for hybrid NN/HMM system 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 train-

ing a single neural network becomes prohibitive due to the large amount of output nodes that would be necessary. A distributed representation can be realized by factoring the posterior state probabilities [4, 6, 8]. Typically, posterior state probabilities are factored according to monophone and context-identity. In contrast, we present a more principled approach where factoring is guided by an agglomerative clustering process.

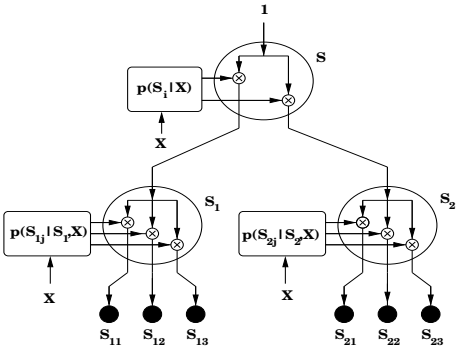

Figure 1: Hierarchical Decomposition of Posteriors

Let S denote the set of all (decision tree clustered) HMM states s_k. Consider we partition 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 (Fig. 1). One of the critical aspect of such a hierarchical decomposition of posteriors is the strategy for partitioning state sets [10].

3. THE ACID ALGORITHM

ACID is an agglomerative clustering algorithm that is based on a measure of the symmetric information divergence between sets of HMM states. The first step in the ACID algorithm is to estimate the parameters of a simple parametric model of the likelihood for each state, in our case diagonal covariance Gaussians. For this kind of model, the symmetric information divergence between two states s_i and s_j amounts to

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

where σ_{ik}^2 and μ_{ik} denote the k-th coefficient of the variance and mean vectors of class s_i, respectively. Making the simplifying assumption of linearity of information divergence, we define the following distance measure between sets of states S_k and S_l

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

where $p(s_i|S_k)$ is the prior probability of state s_i within the set S_k. Initially, each state represents an individual set or cluster. Agglomerative clustering then iteratively merges the pair of sets with minimum distance according to the defined divergence measure. Eventually, the algorithm terminates with a single cluster containing all states. The hierarchical structure that evolves during clustering constitutes a suitable decomposition of the posterior state probabilities.

4. HIERARCHIES OF NEURAL NETWORKS (HNN)

Fig. 2 gives a schematic overview of an HNN based acoustic model. For the estimation of conditional posteriors in each node we are using small 2-layer perceptrons.

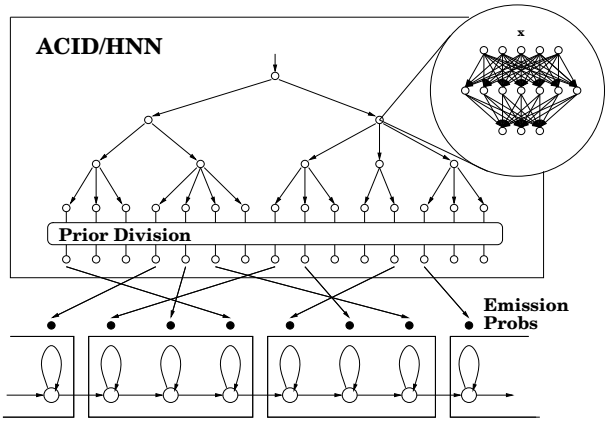

Figure 2: Hierarchy of Neural Networks

The outcome of the ACID clustering algorithm is a binary tree structure. For a given number of HMM states n, such a hierarchy requires to train $n-1$ neural networks. In order to get a more compact HNN with less internal nodes (networks), we apply a greedy bottom-up node merging algorithm with a constrained maximum branching factor b ($b \leq 10$) to obtain the final HNN structure. Such a strategy also has implications on the training of a hierarchy of networks. At the root node, we have to train a network to approximate unconditional posteriors which can be accomplished by training on the full training set. However, going further down the tree, the available training data has to be distributed among children nodes so that the corresponding networks learn to approximate the required *conditional* posteriors. Therefore, networks at the bottom of an HNN potentially receive very little training data. By merging nodes in an HNN, we increase the amount of available training data for the respective network.
After training has converged, the posterior probability of a specific leaf/state in the HNN can be evaluated by traversing the tree from root node to the leaf, evaluating the networks and multiplying the conditional posteriors along the way.

5. EXPERIMENTS ON SWITCHBOARD

Switchboard is a large corpus of conversational American English dialogs, recorded in telephone quality all over the US. It consists of about 170 hours of speech. Switchboard is a comparably hard task, current best systems achieve word error rates of 30-40% in NIST's annual evaluations on this corpus. For our experiments on Switchboard, we integrated the ACID/HNN framework into the Janus-

RTk Switchboard 1997 recognizer [3] and replaced the mixtures of Gaussians with the proposed hierarchical connectionist acoustic modeling part.

5.1. System Setup

Preprocessing consists of extracting an MFCC based feature vector every 10 ms. The final feature vector is computed by a truncated LDA transform on the concatenation of MFCCs and their respective deltas and delta-deltas. Vocal tract length normalization and cepstral mean subtraction are used to extenuate speaker and channel differences. For acoustic modeling, we use allophonic decision trees to cluster 3-state HMMs. The overall model consists of 24000 distinct states. Using this model and mixtures of Gaussians for observation probability estimation the Janus-RTk recognizer scored tied first in NIST's 1997 Switchboard evaluation.

5.2. Clustering and Training of HNNs

Applying the ACID/HNN framework to our system requires to design an HNN with 24000 leaves. Using the ACID algorithm, we constructed an initial binary tree structure with depth 18. Node merging with $b = 10$ was then applied, resulting in the final HNN with depth 5 and a total of 4046 internal nodes. The following table summarizes the structure of this HNN:

tree level	# NN	min/max children	# params
1	1	10/10	4514
2	10	7/10	42560
3	77	4/10	167529
4	524	3/10	671748
5	3434	3/10	1957432
total	4046	–	2.8 M

For training the HNN, we used 87000 utterances or 60 million training patterns. Accurate training labels were available from the standard mixture of Gaussians Janus-RTk system. The parameters of all 4046 networks in the HNN could be jointly estimated with only 3 passes through the training data using stochastic gradient ascent in log-likelihood. Fig. 3 shows the evolution of the log-likelihood of cross-validation data during training.

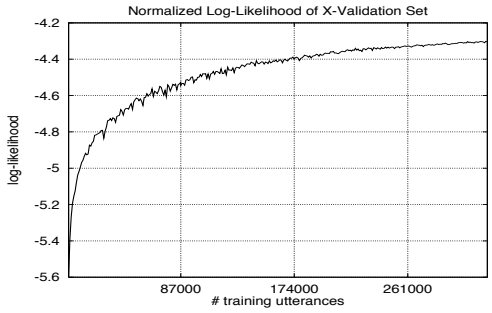

Figure 3: Typical HNN Training Curve

For testing our hierarchical acoustic model, we were using the 1997 Switchboard development test set, consisting of 30 seconds speech from 40 speakers from the Switchboard (SWB) and 40 speaker from the CallHome (CH) corpus. Using an interpolated trigram language model, our gender-independent 24k ACID/HNN system achieves a word error rate of 33.3% on SWB and 43.9% on CH.

6. EXPLOITING HIERARCHICAL STRUCTURE

The key advantage of the ACID/HNN framework is its hierarchical structure. In the remainder, we will experimentally demonstrate how algorithms such as speaker adaptation and decoding speed-up can benefit from this structure.

6.1. Speaker Adaptation

Nowadays, techniques for adapting an acoustic model to specific speakers and/or recording conditions such as MLLR [9] are widely used and consistently improve performance. However, due to limited amounts of adaptation data these techniques usually require to build additional structure, for instance in form of regression trees, to cluster acoustically similar models so that unseen models can be adapted too. In an ACID clustered HNN, there is no need for additional structure since parameter sharing is inherently realized. For instance, adaptation can consist of simply adapting the parameters of the root node, since this node is shared by all models and will receive the highest amount of adaptation data.
For unsupervised speaker adaptation experiments on Switchboard, we were decoding approximately 3 minutes of speech from each of the 40 SWB test speakers. The resulting hypothesis were aligned with the input data to get a labeled training set and filtered by a confidence measure to discard regions that we suspect contain errors. For each speaker, we adapt not only the network in the root node but all those networks in the HNN that receive at least 1000 patterns in order to guarantee sufficient generalization. Adaptation consists in backpropagating errors to the hidden layer and training only the weights from input to hidden layer. Fig 4 gives results of unsupervised adaptation for the individual speakers.

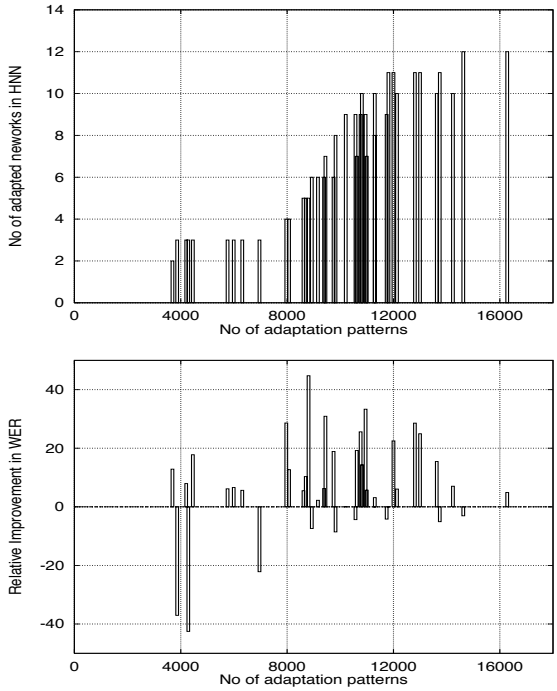

Figure 4: Unsupervised Speaker Adaptation – 40 SWB speakers

As we expected, the number of adapted networks in the HNN increases in proportion to the available adaptation data. However, while the word error rate improves for most of the speakers, adaptation is observed to be counterproductive for some of the speakers

with a small amount of adaptation data. Overall, adaptation decreases word error rate by 5% relative to the unadapted baseline HNN. When restricting adaptation to those speakers with more than 8000 patterns of adaptation data, the word error rate drops by 8% relative.

6.2. Speed vs. Accuracy

Many techniques have recently been invented to speed-up the evaluation of acoustic models during recognition. Techniques such as the BBI algorithm [7] for mixtures of Gaussians (among others) are required to speed-up the recognition process of most speech recognition systems to make them useful in practical applications. Interestingly, many speed-up techniques impose additional hierarchical structures on the set of acoustic models to be able to quickly determine a reduced set of models with high probability. Why not organizing the acoustic model in a hierarchical fashion in the first place and benefit from this structure later on?

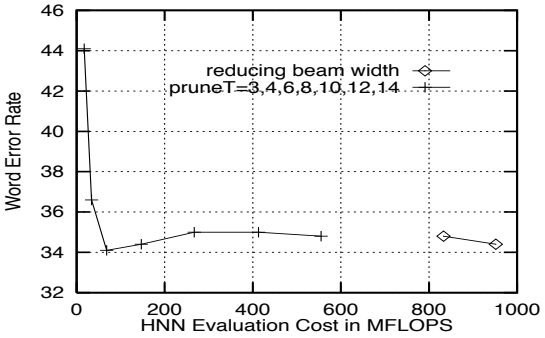

Figure 5: Word Error vs. Acoustic Model Evaluation Cost

With an ACID clustered HNN, speeding up the evaluation of acoustic models can simply be done by pruning subtrees of the hierarchy using a lower bound on the partial posterior probability. In the following experiments we were testing on 12 representative speakers from SWB. Fig. 5 shows the cost of evaluating the HNN when varying a negative logarithmic pruning threshold `pruneT`.

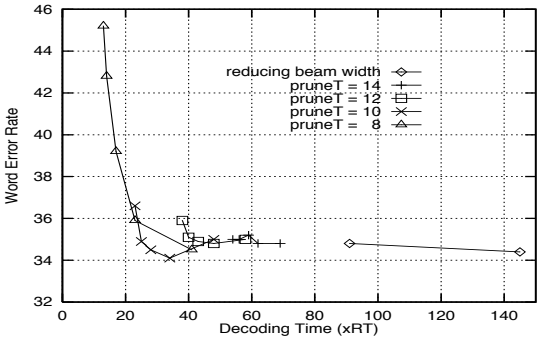

Figure 6: Word Error vs. Decoding Time

Obviously, this cost can be reduced by more than a factor of 10 without affecting the word error rate. However, pruning the acoustic model by imposing a threshold in our case overestimates the posterior probabilities. This in turn slows down a Viterbi beam search decoder such that we do not get the full speed-up of acoustic model evaluation in decoding. To avoid this, we artificially lower the posterior probabilities of pruned models by multiplying them with some factor $c < 1$. Fig. 6 demonstrates the implications of this technique on both speed and accuracy for some values of `pruneT` with varying parameter c ($10^{-7} \leq c \leq 1$).
Using such a simple technique (essentially one line of code), the decoding time of the presented ACID/HNN based acoustic model can be reduced from over 140xRT to about 24xRT without affecting the word error rate at all. Allowing a word error rate of 45%, the speed of the recognizer can be increased further to 14xRT.

7. SUMMARY OF RESULTS

The following table summarizes results for the ACID/HNN based connectionist acoustic models on the SWB and CH test sets.

Decoding Conditions	Word Error		
	SWB	CH	SWB+CH
base LM, no adapt	34.4%	47.8%	40.4%
interp LM, no adapt	33.3%	43.9%	38.6%
base LM, adapt	32.7%	44.6%	38.2%
interp LM, adapt	31.8%	43.3%	37.2%

8. CONCLUSIONS

We experimentally demonstrated the viability of the ACID/HNN framework for connectionist acoustic modeling in an LVCSR system, emphasizing the benefits resulting from the hierarchical structure. Due to discriminative training and an effective pruning technique the presented NN/HNN hybrid recognizer for Switchboard achieves a competitive word error rate of 34.5% with a factor 6 speed-up in decoding time over the baseline system. Applying the proposed method for unsupervised speaker adaptation to the baseline system, we achieve a word error rate of 31.8%.

9. REFERENCES

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

[2] G. D. Cook, D. J. Kershaw, J. D. M. Christie, C. W. Seymour, S. R. Waterhouse, "Transcription of Broadcast Television and Radio News: The 1996 Abbot System", *Proc. of ICASSP'97*, Munich 1997.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] J. Fritsch, I. Rogina, "The Bucket Box Intersection (BBI) Algorithm for Fast Approximative Evaluation of Diagonal Mixture Gaussians", *Proc. of ICASSP'96*, Atlanta 1996.

[8] 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.

[9] C. J. Leggetter and P. C. Woodland, "Speaker Adaptation of HMMs using Linear Regression", *Tech. Rep. CUED/F-INFENG/TR181*, CUED, Cambridge, England 1994.

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