ACID/HNN: A Framework for Hierarchical Connectionist Acoustic Modeling

Jürgen Fritsch
Interactive Systems Labs

University of Karlsruhe
76128 Karlsruhe, Germany

Carnegie Mellon University
Pittsburgh, PA 15213, USA

Abstract - We propose the ACID/HNN framework for context-dependent large vocabulary conversational speech recognition (LVCSR) using connectionist acoustic models. Our approach advocates the principles of modularity and hierarchy for the estimation of thousands of context-dependent posterior HMM state probabilities. We argue that a hierarchical organization of the acoustic model is crucial in obtaining competitive performance with connectionist estimators. We introduce ACID, an Agglomerative Clustering scheme based on Information Divergence and use it to induce soft decision trees for hierarchical classification. A Hierarchy of Neural Networks (HNN) is then applied to the estimation of conditional posterior probabilities. We discuss the benefits of hierarchically structured acoustic models for speaker adaptation and scoring speed-up. Finally, we present experiments on the Switchboard conversational telephone speech corpus, currently a major focus of research in the LVCSR community.

1 Introduction

Statistical speech recognition based on hidden Markov models (HMM) currently is the dominating paradigm in the research community, even though lots of limitations of this technique are repeatedly being discussed. Connectionist acoustic models [1] have proven to be able to overcome some of the drawbacks of HMMs. In particular, connectionist acoustic models were shown to outperform traditional mixtures of Gaussians based acoustic models on small, controlled tasks using context-independent HMMs.
However, wide-spread use of connectionist acoustic models is hindered by at least two issues: (1) Training of connectionist acoustic models is much slower, leading to training times of several days, if not weeks, and (2) poor scalability of connectionist acoustic models to larger systems. Refinement of traditional mixtures of Gaussians based acoustic modeling using phonetic decision trees for polyphonic context modeling recently led to systems consisting of thousands of HMM states. Significant gains in recognition accuracy have been observed in such systems. Nevertheless, research in context-dependent

connectionist acoustic models has long concentrated on comparably small systems since it was not clear how to reliably estimate posterior probabilities for thousands of states. Application of a single artificial neural network as in context-independent modeling leads to an unfeasibly large number of output nodes. Factoring posteriors based on context, monophone or HMM state identity was shown to be capable of breaking down the global estimation problem into subproblems of small enough size to allow the application of multiple artificial neural networks [4, 5, 6]. Comparable gains in performance were achieved with context-dependent connectionist acoustic models based on this technique. However, factoring posteriors in terms of monophone and context identity seems to be limited to medium size systems. In large systems, non uniform distribution of the number of context classes again leads to unfeasibly large numbers of output nodes for some of the context networks.
This paper presents a principled hierarchical approach to factoring posteriors for connectionist acoustic modeling. Our approach exhibits full scalability, avoids stability problems due to non-uniform prior distributions and is easily integrated into existing LVCSR systems. Starting from an inital set of decision tree clustered context-dependent subphonetic units, it uses an agglomerative clustering algorithm across monophones to automatically design a tree structured decomposition of posterior probabilities which is instantiated with thousands of small neural network estimators.

2 Hierarchical Connectionist Modeling

Connectionist acoustic modeling in the context of HMM based speech recognition is characterized by discriminative training of observation probability estimates [1]. Instead of using an independent set of parametric distributions to model HMM emission probabilities for HMM states, connectionist acoustic models make use of artificial neural networks to jointly estimate posterior state probabilities. In this paper, we focus on locally discriminant connectionist acoustic models, mainly because of ease of integration into an existing LVCSR system, in our case Janus-3 [3]. However, our approach is rather general and could in principle be applied to estimate global posteriors.

2.1 Hierarchical Decomposition of Posteriors

Using Bayes rule, HMM emission probabilities can be expressed in terms of posterior state probabilities [1]. This is attractive, because it leads to maximum a-posteriori (MAP) instead of standard maximum likelihood (ML) training. According to this setting, scaled likelihoods can be computed from posterior state probabilities by dividing by priors, which are estimated by relative frequencies. The potentially large number of states in context-dependent HMM modeling requires to factor the posterior probability in order to be able to apply estimators such as artificial neural networks.

Let S be a set of HMM states [1] s_k. For the moment, consider we have a method at our disposition which gives us a reasonable partition of such a set 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).

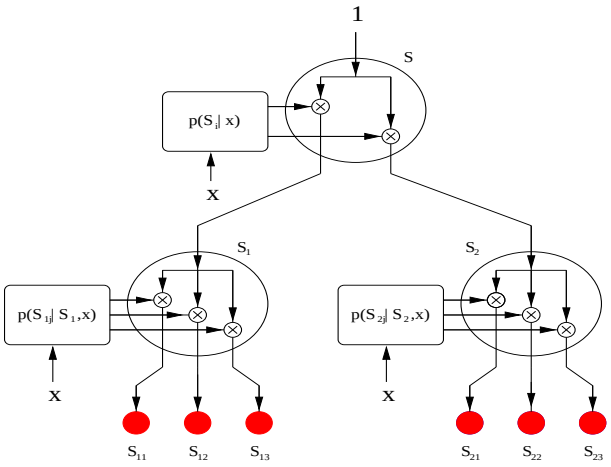

Figure 1: Hierarchical Decomposition of Posteriors

It can be interpreted as a probability mass distribution device [8]. At the root node, an initial probability mass of 1 is fed into the architecture. At each node, the incoming probability mass is multiplied by the conditional posterior probabilities and fed into the children nodes. Eventually, the probability mass is distributed among all the leaves (states) rendering their posterior probabilities. In contrast, typical hierarchical classifiers such as classification trees [2] operate as hard switching devices, allowing only a single path from root node to one of the leaves, depending on the outcome of categorical questions in internal nodes.

Since perfect estimation of (conditional) posterior probabilities can not be achieved in practice, the proposed hierarchical decomposition critically depends on the method used to design the tree structure. One could argue, that

[1] Throughout the paper, the term 'HMM states' refers to a set of tied HMM states, typically clustered by means of phonetic decision trees

such a method is superfluous since we already have available a tree structure in form of the phonetic decision trees used to cluster context-dependent HMM states. However, we prefer not to adopt phonetic decision trees for several reasons: (1) In most cases, separate decision trees are used to independently cluster context classes for each monophone, and (2) phonetic decision trees often are highly unbalanced. Therefore, we propose to apply an unconstrained clustering algorithm that allows to form tree structured hierarchies across phone identities. Furthermore, our algorithm implicitly pursues uniform prior distributions in each node and therefore avoids unbalanced splits which could lead to poorly approximated conditional 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 propose to apply an agglomerative (bottom-up) clustering scheme using the symetric information divergence as a measure of acoustic dissimilarity of subphonetic units. Based on this rather inexpensive distance measure, subphonetic units can be clustered efficiently yielding a suitable hierarchical decomposition of posteriors.

3.1 Information Divergence

Consider the case of two acoustic classes, s_i and s_j which are to be discriminated. Let $p(\mathbf{x}|s_i)$ and $p(\mathbf{x}|s_j)$ be the class conditional likelihoods for s_i and s_j, respectively. The average symetric discriminating information [9], or symetric information divergence between s_i and s_j can then be defined as

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

Now, suppose we model the class-conditional likelihoods using single full covariance multivariate Gaussians with mean vectors μ_i and covariance matrices $\mathbf{\Sigma}_i$. The symetric information divergence between two normally distributed classes s_i and s_j is

$$\begin{aligned} d(s_i, s_j) &= \frac{1}{2} tr\{(\mathbf{\Sigma}_i - \mathbf{\Sigma}_j)(\mathbf{\Sigma}_j^{-1} - \mathbf{\Sigma}_i^{-1})\} \\ &+ \frac{1}{2} tr\{(\mathbf{\Sigma}_i^{-1} + \mathbf{\Sigma}_j^{-1})(\mu_i - \mu_j)(\mu_i - \mu_j)^t\} \end{aligned}$$

To reduce the computational load of a clustering algorithm that utilizes this distance measure, one can restrict the Gaussian covariances to diagonal ma-

trices, resulting in the following distance measure

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

where σ_{ik}^2 and μ_{ik} denote the k-th coefficient of the variance and mean vectors of class s_i, respectively.

3.2 ACID Clustering

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

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

This distance measure is used in the **ACID** clustering algorithm:

1. **Initialize algorithm with n clusters S_i, each containing (1) a parametric model of the class-conditional likelihood and (2) a count C_i, indicating the frequency of class s_i in the training set.**
2. **Compute within cluster priors $p(s_i|S_k)$ for each cluster S_k, using the counts C_i**
3. **Compute the symetric divergence measure $D(S_k, S_l)$ between all pairs of clusters S_k and S_l.**
4. **Find the pair of clusters with minimum divergence, S_k^* and S_l^***
5. **Create a new cluster $S = S_k^* \bigcup S_l^*$ containing all Gaussians of S_k^* and S_l^* plus their respective class counts. The resulting parametric model is a mixture of Gaussians where the mixture coefficients are the class priors**
6. **Delete clusters S_k^* and S_l^***
7. **While there are at least 2 clusters remaining, continue with 2.**

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).

3.3 Hierarchies of Neural Networks (HNN)

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.

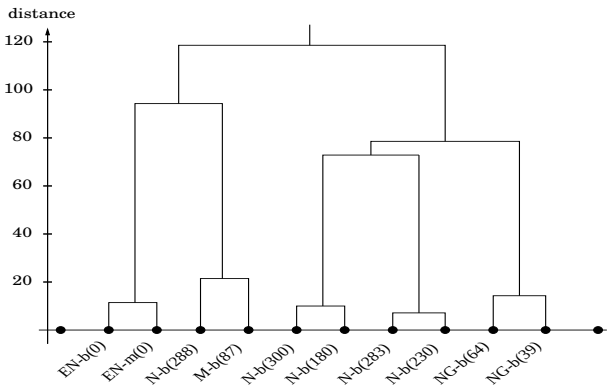

Figure 2: Partial Dendrogram of ACID Clustering

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). 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 Exploiting HNN Structure

The hierarchical structure of ACID/HNN based acoustic models contains information about similarity of acoustic units on a coarse to fine grain scale that is missing in conventional flat organizations of acoustic models. This information can for example be exploited in speed-up and adaptation algorithms where it leads to elegant solutions.

4.1 Speed vs. Accuracy

In contrast to conventional mixtures of Gaussians based acoustic models, the ACID/HNN framework does not require additional structures to reduce the complexity of model evaluation. The tree structure itself can be exploited to control the speed-accuracy trade-off. The evaluation of posterior state probabilities follows a path from root node to a specific leaf in the HNN, multiplying all estimates of conditional posteriors along the way. Subtrees can be pruned by closing paths whenever the partial probability falls below a suitable threshold. This way the evaluation of a significant amount of

networks at the bottom of the HNN can be avoided, possibly at the cost of increased error rate.

4.2 Speaker Adaptation

In order to achieve robust adaptation to specific speakers on limited data, conventional acoustic models usually require additional structure in form of regression trees to assign a small set of adaptation transformations to parameters of HMMs as in the MLLR [7] framework. Such information is readily available in the HNN structure and robust speaker adaptation can be accomplished by simply adapting those networks in the HNN tree that receive enough adaptation data. Individual networks can be adapted by updating weights of either all or some of the layers using error backpropagation on Viterbi state alignments. This scheme automatically adjusts to the amount of available adaptation data. In case of very little data, only a few networks in the vincinity of the root node will get updated. The more data becomes available, the more networks receive enough samples, until eventually all of the networks in the HNN become subject to an update.

5 Experiments

In our initial experiments with the ACID/HNN framework, we were constructing and training hierarchies for 6000 and 24000 HMM states on the Switchboard LVCSR corpus. Approximately 160 hours or 57.6M speech frames were available for training the architecture. Training targets (state alignment labels) were generated using the Janus-3 1997 Switchboard recognizer [3]. Cross validation using 400 utterances was used to monitor performance and to decide when to stop training. Using individually adapted learning rates during training, 1-4 passes through the training data usually suffice to reach a maximum in log-likelihood on the Switchboard corpus.
We integrated the proposed hierarchical connectionist acoustic models into the Janus-3 recognizer such that we could benefit from a dictionary, phonetic decision trees and language models optimized for Switchboard. Competitive performance was achieved with acoustic models based on the ACID/HNN framework, outperforming our earlier approaches to context-dependent connectionist acoustic modeling. The following table gives results for different connectionist acoustic models on the 1996 Switchboard evaluation set:

acoustic model	# HMM states	# networks	# params	word error
CI HME/HMM	166	59	220k	58.6 %
CD HME/HMM	10000	224	1.2M	37.3 %
CD ACID/HNN	**6000**	**962**	**1.6M**	**35.7 %**
CD ACID/HNN	**24000**	**4046**	**2.8M**	**33.3 %**

The first two rows give results obtained with our earlier approach to connectionist acoustic modeling [5]. CI/CD denote context-independent/-dependent systems, respectively. Significant improvements were achieved with the ACID/

HNN framework. It was for the first time possible to succesfully train and test a connectionist acoustic model for as much as 24k HMM states, indicating better scalability of the ACID/HNN framework.

6 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 discriminative training while circumventing the limitations of standard connectionist acoustic models. Furthermore, ACID/HNN acoustic models already incorporate the structure for speaker adaptation and scoring speed-up algorithms that usually require additional effort in traditional mixture densities acoustic models.

Acknowledgments

The author wishes to thank all colleagues in the ISL Switchboard group, especially Michael Finke for fruitful discussions and active support.

References

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

[2] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification and Regression Trees*, Wadsworth International Group 1984.

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

[6] 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.

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

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

[9] J. T. Tou and R. C. Ganzales, *Pattern Recognition Principles*, Addison Wesley, 1974.