EFFECTIVE STRUCTURAL ADAPTATION OF LVCSR SYSTEMS TO UNSEEN DOMAINS USING HIERARCHICAL CONNECTIONIST ACOUSTIC MODELS

Jürgen Fritsch, Michael Finke, Alex Waibel
fritsch@ira.uka.de

Interactive Systems Labs
University of Karlsruhe, Germany

Interactive Systems Inc.
Pittsburgh, USA

ABSTRACT

We present an approach to efficiently and effectively downsize and adapt the structure of large vocabulary conversational speech recognition (LVCSR) systems to unseen domains, requiring only small amounts of transcribed adaptation data. Our approach aims at bringing todays mostly task dependent systems closer to the aspired goal of domain independence. To achieve this, we rely on the ACID/HNN framework [2, 3], a hierarchical connectionist modeling paradigm which allows to dynamically adapt a tree structured modeling hierarchy to differing specifity of phonetic context in new domains. Experimental validation of the proposed approach has been carried out by adapting size and structure of ACID/HNN based acoustic models trained on Switchboard to two quite different, unseen domains, Wall Street Journal and an English Spontaneous Scheduling Task. In both cases, our approach yields considerably downsized acoustic models with performance improvements of up to 18% over the unadapted baseline models.

1. INTRODUCTION

Despite the success of current HMM based technology, speech recognition systems still suffer from domain dependence. Over the years, the community has validated and emerged the technology based on standardized training and test sets in restricted domains, such as Wall Street Journal (WSJ) (business newspaper texts), Switchboard (SWB) (spontaneous telephone conversations) and Broadcast News (BN) (radio/tv news shows). Performance of systems trained on such domains typically drops significantly when applied to different domains [5], especially with changing speaking style, e.g. when moving from read speech to spontaneous speech. For instance, performance of a recognizer trained on WSJ typically decreases severely when decoding SWB data. Several factors can be held responsible for the strong domain dependence of current statistical speech recognition systems:

- Constrained quality, type or recording conditions of domain specific speech data (read, conversational, spontaneous speech / noisy, clean recordings / presence of acoustic background sources, etc.)
- Vocabulary and language model dependence of phonetic context modeling based on phonetic decision trees. This implies a strong dependence of allophonic models on the specific domain.
- Domain dependent optimization of size of acoustic model based on amount of available training data and/or size of vocabulary.

While the first of the above mentioned factors is typically addressed by some sort of speaker and/or environment adaptation technique, the latter two factors usually miss an adequate treatment in cross-domain applications.

In this paper, we present an approach that allows to effectively adapt the structure and size of trained acoustic models to unseen domains with only small requirements regarding the amount of adaptation data. It is based on a previously proposed architecture for connectionist acoustic modeling - the ACID/HNN framework [2, 3] - benefiting from a multi-level, hierarchical representation of context-dependent acoustic models. In contrast to approaches based on acoustic adaptation only, our approach uses an estimate of the a-priori distribution of modeled HMM states on the new domain to dynamically downsize/prune the tree structured acoustic model. This way, we can account for differences in vocabulary size and adjust to the specifity of phonetic contexts observed in the new domain.

Consider the scenario of porting a trained recognizer to a different domain within the same language. Usually, a phonetic dictionary for the new domain based on the set of phones modeled by the recognizer can be constructed relatively easily using a large background dictionary and, if necessary, applying a set of phone mapping rules. Also, we consider it justifiable to assume that enough text data is available, such that we can train a statistical language model for the new domain. What typically makes porting efforts expensive and time consuming is the adaptation of the acoustic model. The most common approach of applying supervised acoustic adaptation techniques requires large amounts of transcribed speech data from the new domain in order to capture the differing statistics reasonably well. We will show how additionally adapting the specifity of phonetic context modeling leads to improved performance with very little requirements on the amount of adaptation data. Furthermore, our approach compensates overfitting effects particularly when targeting a domain with much smaller vocabulary.

Although we focus on adapting the structure of our modeling tree to compensate allophonic mismatches between training and new domain, our approach may also be applied to downsize an ACID/HNN based acoustic model to any desired size in order to accomodate computing and/or memory resource limitations comparable to [4].

2. HIERARCHICAL CONNECTIONIST ACOUSTIC MODELING

In connectionist acoustic modeling, one tries to benefit from discriminatively trained neural networks estimating posterior state probabilities. To accomodate the HMM framework, these posteriors have to be scaled by dividing by prior probabilities. The resulting scaled likelihoods can then be used to replace or enrich standard Gaussian mixture modeling.

It is well known that detailed phonetic context modeling is one of the key techniques for achieving state-of-the-art performance in large vocabulary conversational speech recognition (LVCSR) systems, particularly with ever increasing training data sets. Connectionist acoustic models proved to be much harder to scale to the typically very large state/model spaces of LVCSR systems.

Recently, we have developed a framework based on data-driven hierarchical factoring of posterior probabilities [2, 3] which shows good scalability and additionally offers some attractive properties absent in traditional acoustic models.

2.1. The ACID/HNN Framework

In context-dependent connectionist acoustic modeling, the global task of discriminating between all (tied) HMM states modeled by the recognizer can be decomposed into a tree structured configuration of conditional posterior probability estimators. Fig. 1 depicts this divide-and-conquer technique which is based on factoring the states' posterior probability distribution.

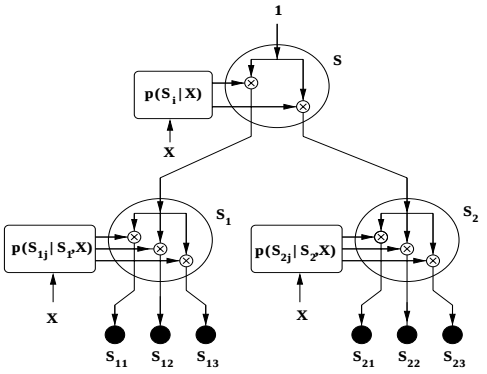

Fig. 1: Hierarchical Decomposition of Posterior Probabilities

The posterior probability of a particular HMM state can now be computed by multiplying all conditional posteriors estimated by the nodes along the path from root node to the leaf representing the state. Applying small neural networks to estimate the local conditional posterior probabilities in each node yields what we call a Hierarchy of Neural Networks (HNN). We use Agglomerative Clustering based on Information Divergence (ACID) [2] to construct HNNs automatically using Gaussian statistics gathered for all distinct allophonic HMM states modeled by the system (which will become leaves in the HNN tree).

In [2, 3], we have demonstrated that the ACID/HNN framework allows to construct competitive connectionist acoustic models for as much as 24000 allophonic HMM states. Furthermore, the hierarchical structure allows to dynamically prune model evaluation and supports acoustic adaptation naturally. Note that for a given acoustic feature vector $\mathbf{x}_t$, posterior $p(s_i|\mathbf{x}_t)$, prior $P(s_i)$ and scaled likelihood $\hat{p}(\mathbf{x}_t|s_i)$ of an HNN leaf modeling state s_i can be computed incrementally in log space:

$$\begin{aligned}
\log p(s_i|\mathbf{x}_t) &= \sum_{k=0}^{D(i)-1} \log p(N_i(k+1)|N_i(k), \mathbf{x}_t) \\
\log P(s_i) &= \sum_{k=0}^{D(i)-1} \log P(N_i(k+1)|N_i(k)) \\
\log \hat{p}(\mathbf{x}_t|s_i) &= \sum_{k=0}^{D(i)-1} \Big[\log p(N_i(k+1)|N_i(k), \mathbf{x}_t) \\
&\quad - \log P(N_i(k+1)|N_i(k))\Big]
\end{aligned}$$

where $D(i)$ denotes the depth of leaf s_i in the HNN tree, $N_i(k)$ denotes the tree node at depth k along the path from root node to leaf s_i, and the $p(l|m, \mathbf{x}_t)$ and $P(l|m)$ denote local conditional posteriors and priors of node l given node m, respectively. Since the conditional log posteriors and log priors are all negative, partial posteriors and priors of leaf nodes decrease monotonically when traversing the tree and computing the above sums. This property can for instance be exploited in posterior pruning which typically yields significant savings in computational load.

The following Fig. 2 gives an overview of how the HNN architecture is applied to the estimation of HMM emission probabilities using phonetic decision trees to assign scaled likelihoods at HNN leaves to actual HMM states.

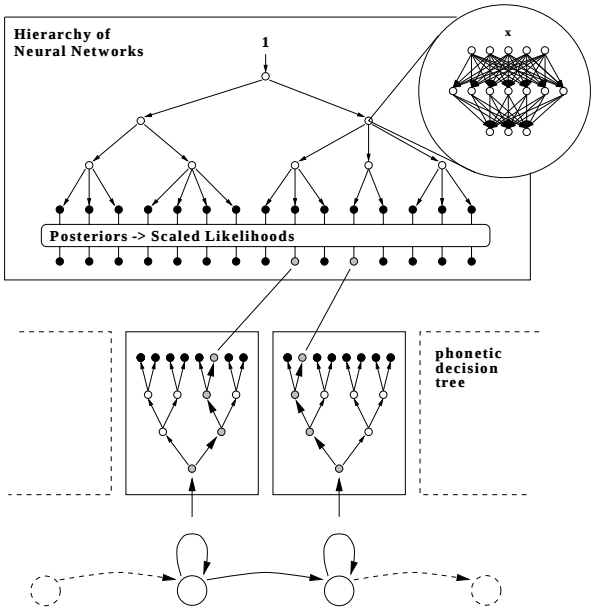

Fig. 2: Integration of ACID/HNN Architecture into LVCSR System

3. STRUCTURAL ADAPTATION

An interesting property of HNNs that we exploit for structural adaptation is the fact that partially computed posterior probabilities at all crossed paths in every horizontal cross section of the tree constitute a legal posterior probability distribution over a reduced (merged) set of leaves. Starting point for structural adaptation is a Hierarchy of Neural Networks constructed and trained on a domain exhibiting sufficiently rich diversity in phonetic context to provide a basis for any new, unseen domain. When adapting this baseline tree structure to a new, smaller domain typically exhibiting a very different specifity of phonetic context, we perform the following steps (see Fig. 3)

1. Take the baseline HNN tree (circles=nodes, squares=leaves)
2. Select nodes that receive more than a predetermined, sufficiently large amount of adaptation data (*mincount*) and adapt their local estimators of conditional posteriors *and* priors using adaptation data from the new domain.
3. Remove all nodes that receive less than a predetermined amount of adaptation data. Create new leaf nodes (squares) in place of the root nodes of pruned subtrees
4. Finally, merge leaf nodes of pruned subtrees. Tie all HMM states corresponding to the leaves of pruned subtrees in the original tree such that they share a single model, represented by the newly created leaves

Although step 2 appears to operate similar to adaptation techniques such as regression tree based MLLR, its effects are actually quite different due to the possibility and necessity of adapting the priors too, a feature that is unique to connectionist architectures. By adapting the local conditional priors, step 2 already modifies the structure of HNNs implicitly by, for instance, cutting off subtrees whose models could not be observed in adaptation data. In addition, step 3 and 4 are used to control the size of the resulting HNN by merging the models with the smallest prior probability in the target domain. Furthermore, computational complexity of model evaluation can be traded off against recognition accuracy. In fact, it turns out that in many cases, one can heavily downsize the HNN tree without loosing recognition accuracy.

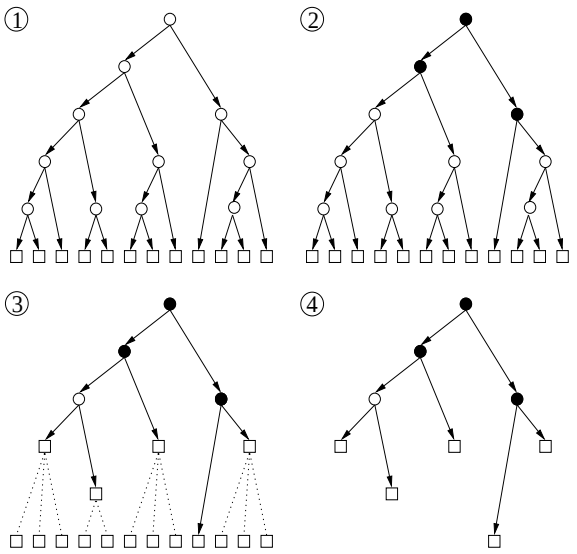

Fig. 3: Structural Adaptation of HNN trees (see text)

4. EXPERIMENTS

As baseline recognizer for our experiments, we use a variant of the JanusRTk based 1997 CMU-ISL Switchboard/Callhome recognizer [1] with ACID/HNN based acoustic modeling. For details concerning the construction and training of the ACID/HNN models for this recognizer we refer the reader to [2, 3]. We chose Switchboard as our training domain since it consists of very sloppy conversational speech showing many spontaneous effects such as false starts, hesitations, interjections, etc. . Furthermore, recording conditions vary extremely since Switchboard was collected over the public telephone network. In summary, we believe that Switchboard offers immense phonetic and acoustic variety and serves well as a data set for training a baseline system for the English language.

4.1. Target Domains

We report results of experiments in applying our structural adaptation approach to adapting the baseline Switchboard recognizer to two quite different target domains.

	SWB	WSJ	ESST
style	conv.	read	conv.
rec. quality microphone	noisy telephone	clean telephone	clean Sennheiser
vocab size w/ variants	14959 29573	4999 10170	2851 4636
# adapt spks adapt data		10 50 min	18 60 min
# test spks test data		10 27 min	14 18 min

Tab. 1: Domain Overview

The first one is taken from the Wall Street Journal (WSJ) task, representing a domain of read speech. We chose the official 1993 WSJ Spoke 6 data set which consists of recordings from a telephone handset [1] and therefore matches the bandwidth of SWB data. As second target domain, we chose a subset of an English Spontaneous Scheduling Task (ESST) collected at CMU. ESST consists of conversational speech recorded in high quality (16kHz/16bit). Tab. 1 gives details about all three domains considered in our experiments. Speakers used for adaptation are different from test speakers in both cases.

[1] in contrast to the majority of WSJ data which is recorded from Sennheiser high quality microphones.

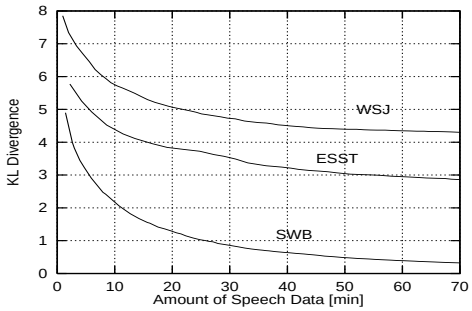

Fig. 4: Mismatch to trained SWB 24k CD models' prior distribution

The plot in Fig. 4 demonstrates that there is great mismatch in specifity of phonetic context between baseline and target domains. It plots the Kullback-Leibler (information) divergence between the a-priori distribution of the baseline recognizers' 24k distinctly modeled allophonic states (estimated from the full SWB corpus) and a-priori distributions of the same set of states estimated from a varying amount of SWB, WSJ and ESST data. As expected, the SWB curve quickly approaches zero. WSJ and ESST, however, exhibit a certain bias in divergence that will never be remedied by more data. Interestingly, WSJ shows the largest mismatch in prior distribution which might be attributed to very different vocabulary and speaking style.

4.2. Structural Adaptation

In evaluating structural adaptation, we were following the outline presented in section 3 using adaptation and test data from the new domains as stated in Tab. 1. As a baseline, we first computed the performance of the unadapted SWB recognizer on the target domains. Next, we aligned the available adaptation data using the SWB models in order to get state alignments for adapting individual node classifiers in the HNN tree. Adaptation of conditional posteriors was done by continuing to train the networks of selected nodes on available adaptation data until convergence of likelihood (max. of 20 iterations to avoid overfitting). Priors were adapted by replacing the SWB based estimates in selected nodes by estimates based on the adaptation data. Tab. 2 gives an overview of the three mincounts investigated for selecting NN nodes for adaptation (see step 2 in section 3) and the number of NN nodes actually selected and adapted in each case.

mincount	WSJ	ESST
500	240	280
1000	101	120
2000	72	71

Tab. 2: Number of Adapted NN Nodes in HNN Tree

After adapting the selected networks, we experimented with various degrees of pruning resulting in ever smaller HNN trees. Fig. 5 shows word error rate results obtained for adapting and pruning the SWB recognizers ACID/HNN based acoustic model to WSJ and ESST, respectively. Note the single marks for 24000 models, indicating the performance of the unadapted baseline recognizer. Note also that already after HNN node adaptation (no explicit pruning yet) the number of remaining models (having non-zero priors) dropped significantly (see rightmost point on the curves), particularly in the case of ESST. Starting at the right end of the curves, results plotted towards the left of the graph were obtained with increasing pruning thresholds. The three curves in

each graph represent the three mincounts for acoustic adaptation that were examined.

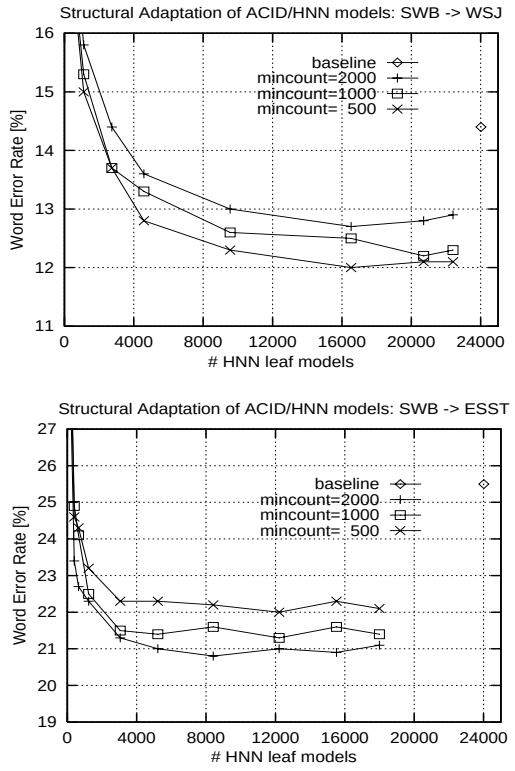

Fig. 5: Structural Adaptation from SWB to WSJ and ESST

While a mincount of 500 frames for adaptation proved optimal in the case of WSJ, a mincount of 2000 gave best results on ESST. However, the effects of varying mincount are reasonably small to consider its choice rather uncritical. Tab. 3 gives details about the configuration of the ACID/HNN models for the settings resulting in optimal performance on the target domains.

AM	SWB	WSJ	ESST
baseline [WER]	34.4%	14.4%	25.5%
baseline [#nodes]	4046	4046	4046
baseline [#models]	24016	24016	24016
adapted [WER]		12.0%	20.8%
adapted [#nodes]	—	2645	1366
adapted [#models]		16532	8411
WER improvement	—	**16.7%**	**18.4%**

Tab. 3: Structural Adaptation – ACID/HNN Models

In both cases, WSJ and ESST, performance could be consistently improved by 16-18% through structural adaptation. On WSJ, the size of the resulting optimal HNN tree is only 65% of the original size. On ESST, the optimal tree is even smaller. Only 33% of the original tree nodes remain after optimal pruning in that case. Interestingly, the WSJ and ESST HNN trees can be pruned further to only about 15% and 10% of the original size with very modest increases in word error rate. Finally, our results compare favourably with those achieved by systems trained specifically on large training corpora from the target domains. For instance, the best reported result in 1994 on the WSJ'93 Spoke 6 test set was 12.5% WER [6]. Using a Gaussian mixtures based recognizer trained on ESST training data we achieved an error rate of 19.5% on the test set used in our experiments.

4.3. Contrast Experiment

Since the presented approach interweaves acoustic and structural adaptation indivisibly, we consider it important to compare it to an approach which performs acoustic adaptation only, in order to assess the impact of adapting the structure. For instance, can we get the same performance improvements by adapting a set of Gaussian mixture distributions using supervised MLLR?

To answer such questions, we ran additional experiments using the same recognizer setup but replacing the ACID/HNN based acoustic model by our standard SWB acoustic model based on mixtures of Gaussians (MOG). The baseline results with unadapted models are slightly better than those obtained with ACID/HNN models due to the better baseline performance of the MOG models on SWB. However, since we are interested in relative performance improvements through adaptation we did not try to equal baseline model performance. Supervised MLLR adaptation of Gaussian means was performed on the target domains' adaptation sets, carefully adjusting adaptation parameters such that the number of MLLR transformations applied (≈ 100) roughly matched the number of nodes adapted in the ACID/HNN model. Tab. 4 reports results of these experiments.

AM	SWB	WSJ	ESST
baseline [WER]	31.5%	13.3%	24.8%
adapted [WER]	—	12,0%	23.0%
WER improvement	—	**9.8**%	**7.3%**

Tab. 4: MLLR – 24k Gaussian Mixture Models

Obviously, adapting solemnly the estimators of acoustic observation probabilities yields only about half the improvements obtained by adapting both acoustics and structure. MLLR based adaptation seem to have much more difficulties coping with severe mismatches in the prior distribution of modeled states (particularly on ESST), as opposed to structural adaptation of ACID/HNN models.

5. CONCLUSIONS

We presented an approach for effectively adapting the structure of a tree-structured hierarchical connectionist acoustic model to unseen new domains. In contrast to existing architectures and adaptation techniques, our approach not only compensates for mismatches in acoustic space but furthermore adapts to differing specifity of phonetic context in unseen domains by adapting node priors and pruning defective parts of the modeling hierarchy.

6. REFERENCES

[1] 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.

[2] J. Fritsch, "ACID/HNN: A Framework for Hierarchical Connectionist Acoustic Modeling", *Proceedings of IEEE ASRU Workshop*, Santa Barbara, 1997.

[3] J. Fritsch, M. Finke, "ACID/HNN: Clustering Hierarchies of Neural Networks for Context-Dependent Connectionist Acoustic Modeling", *Proceedings of ICASSP'98*, Seattle, 1998.

[4] M. Hwang, X. Huang, "Dynamically Configurable Acoustic Models for Speech Recognition", *Proceedings of IEEE ICASSP'98*, Seattle 1998.

[5] D. L. Thomson, "Ten Case Studies of the Effect of Field Conditions on Speech Recognition Errors", *Proceedings of IEEE ASRU Workshop*, Santa Barbara, 1997.

[6] *Proceedings of ARPA Spoken Language Systems Technology Workshop*, Princeton 1994.