# CONTEXT-DEPENDENT HYBRID HME/HMM SPEECH RECOGNITION USING POLYPHONE CLUSTERING DECISION TREES

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

**Interactive Systems Laboratories**
University of Karlsruhe — Germany
Carnegie Mellon University — USA

## ABSTRACT

This paper presents a context-dependent hybrid connectionist speech recognition system that uses a set of generalized hierarchical mixtures of experts (HME) to estimate context-dependent posterior acoustic class probabilities. The connectionist part of the system is organized in a modular fashion, allowing the distributed training of such a system on regular workstations. Context classes are based on polyphonic contexts, clustered using decision trees which we adopt from our continuous density HMM recognizer JANUS [8]. The system is evaluated on ESST, an english speaker-independent spontaneous speech database. Context dependent modeling is shown to yield significant improvements over simple context-independent modeling, requiring only small additional overhead in terms of training and decoding time.

## 1. INTRODUCTION

It was recently shown by a variety of researchers (eg. [1, 2, 4]) that hybrid HMM systems which rely on connectionist discriminative acoustic modeling can be competitive with traditional mixtures of Gaussians based HMM systems, yet requiring orders of magnitude less parameters. Such systems are attractive, because they are compact and offer faster decoding speeds than standard systems. Also, they facilitate the incorporation of additional knowledge sources into the process of computing acoustical scores (e.g. using a window of input frames). However, training the network(s) of hybrid systems generally requires parallel implementations and is often reported to take several days, which is more than one order of magnitude higher than the training time of traditional systems.

We present a system based on modular neural networks, specifically generalized hierarchical mixtures of experts (HME) [5, 6], where gates and experts in the HME tree nodes can contain arbitrary classifiers, as long as they follow a multinomial probability model. The modular aspect of HME's bears similarities to the Meta-Pi paradigm [3] with the difference, that the training data is not partioned a-priori among experts in an HME - Instead, the network learns smooth feature space partitionings without supervision by maximizing the likelihood of a generative statistical model. The HME architecture and its underlying statistical framework offer faster training times than those observed in MLP and recurrent neural network based hybrid systems. In fact, it can be trained in a reasonable amount of time (approx. 2-3 times real-time for one of 2-5 training iterations) on a set of regular workstations.

Modeling of subword units in context is a standard technique which boosts performance of current state-of-the-art HMM recognizers significantly. Relatively simple context-independent hybrid systems were reported to be competitive with more sophisticated context-dependent mixture-of-Gaussian systems [4], but it was shown that hybrid systems also benefit from context modeling [2, 7, 9]. In this paper, we report first results of our ongoing work on connectionist context-modeling for our hybrid HME/HMM system.

## 2. GENERALIZED HIERARCHICAL MIXTURES OF EXPERTS

Jordan and Jacobs [5, 6] introduced the hierarchical mixture of experts as a modular neural network for supervised learning using the divide-and-conquer strategy. The learning task is divided in sets of overlapping regions by a tree-organized hierarchy of gating networks. Expert networks at the leaves of the tree perform the learning task in their specific region of the input space. Expert outputs are blended by the gating networks and proceed up the tree to yield the final output. Expert and gating networks parameters are jointly estimated in order to maximize the likelihood of a generative model, that is, the construction of overlapping regions in which experts act requires no supervision and is part of the learning algorithm. It was shown, that an HME can model discontinuities in the input-output mapping much better than traditional monolithic neural networks.

Fig. 1 shows the structure of a binary branching HME of depth 2. The output vector of such an HME is computed according to

$$\mu = \sum_i g_i(\mathbf{x}) \sum_j g_{j|i}(\mathbf{x}) \mu_{ij}(\mathbf{x})$$

where $g_i(\mathbf{x})$ and $g_{j|i}(\mathbf{x})$ are the outputs of gating networks and $\mu_{ij}(\mathbf{x})$ are the outputs of the expert networks. In our case, HME's are being used in a hybrid NN/HMM speech recognition framework as classifiers, estimating posterior class probabilities. For classification, expert and gating netorks in an HME compute multinomial probability models and are therefore parameterized using the softmax non-linearity ('canonical link' in GLIM theory):

$$z_i(\mathbf{x}) = \frac{\exp y_i(\mathbf{x})}{\sum_j \exp y_j(\mathbf{x})}$$

In [5, 6] the $y_i(\mathbf{x})$ are parameterized as linear models, leading to an efficient EM training algorithm (iteratively re-weighted weighted least squares) for the hierarchy. However, we discovered that it is sometimes advantageous to use more complex parameterizations for gates and experts, eg. multi-layer feed-forward architectures. Such architectures can still be trained efficiently using generalized EM algorithms with on-line updates.

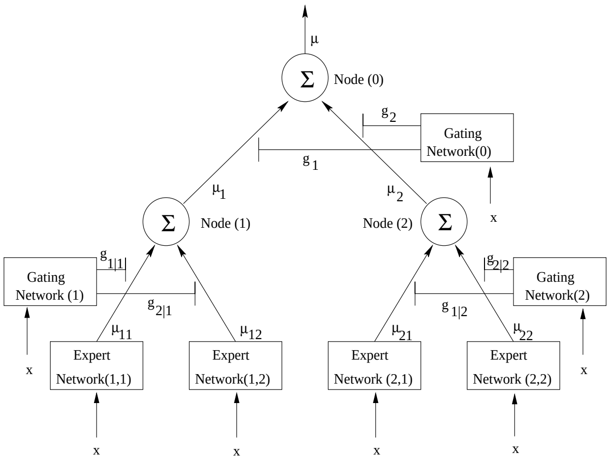


Fig. 1: Hierarchical Mixtures of Experts

## 3. CONNECTIONIST CONTEXT MODELING

Consider an HMM based speech recognition system that models sub-word units (eg. phones, clustered triphones or polyphones) using $m$-state left-right topologies. Such a multi-state HMM model requires the computation of state, monophone and context dependent likelihoods $p(\mathbf{x}|c_j, \omega_i, s_k)$ for each frame $\mathbf{x}$, where $s_k (1 \le k \le m)$ denotes the HMM state, $c_j$ the context class and $\omega_i$ the monophone of a context-dependent acoustic model.

In traditional HMM system, the above likelihoods are modeled independently, estimating separate parametric distributions, usually mixtures of Gaussians, for each model. Applying Bayes' rule and factoring the conditional probabilities, we can reformulate the problem in a way that allows the discriminative estimation of scaled likelihoods in terms of a-posteriori probabilities

$$\begin{aligned} p(\mathbf{x}|c_j, \omega_i, s_k) &= \frac{p(c_j, \omega_i, s_k|\mathbf{x})p(\mathbf{x})}{P(c_j, \omega_i, s_k)} \\ &= \frac{p(c_j, \omega_i|s_k, \mathbf{x})}{P(c_j, \omega_i|s_k)} \; \frac{p(s_k|\mathbf{x})}{P(s_k)} \; p(\mathbf{x}) \\ &= \frac{p(c_j|\omega_i, s_k, \mathbf{x})}{P(c_j|\omega_i, s_k)} \; \frac{p(\omega_i|s_k, \mathbf{x})}{P(\omega_i|s_k)} \; \frac{p(s_k|\mathbf{x})}{P(s_k)} \; p(\mathbf{x}) \end{aligned}$$

All the terms in the denominators are prior probabilities, which can be estimated by relative frequencies. The frame probability $p(\mathbf{x})$ can be dropped, when seeking the model with maximum likelihood. It remains to estimate the posteriors in the numerators.

Starting from the right side, the posteriors $p(s_k|\mathbf{x})$ can be computed by a single neural network, discriminating between the states in an $m$-state HMM topology. Therefore, we call such a network a *state discriminating network (SDN)*.

The posteriors $p(\omega_i|s_k, \mathbf{x})$ are conditioned on the HMM state and the input frame and can be computed by a set of $m$ networks, one for each HMM state. Given a particular HMM state $s_k$, the corresponding network must be trained to discriminate between the monophones $\omega_i$, thus it'll estimate $p_k(\omega_i|\mathbf{x})$.

The posteriors $p(c_j|\omega_i, s_k, \mathbf{x})$ are conditioned on the input frame $\mathbf{x}$, the HMM state $s_k$ and the monophone $\omega_i$. They can be computed by a matrix of networks consisting of $m$ times $n$ networks (where $n$ is the number of monophones). Each of these networks discriminates between all the context classes of a specific monophone in a specific state. The network for state $s_k$ and monophone $\omega_i$ therefore computes $p_{ki}(c_j|\mathbf{x})$.

The following figure gives an overview of the structure of a set of posterior probability estimators (in our case HME's) for a 3-state HMM topology. Each box represents a single HME:

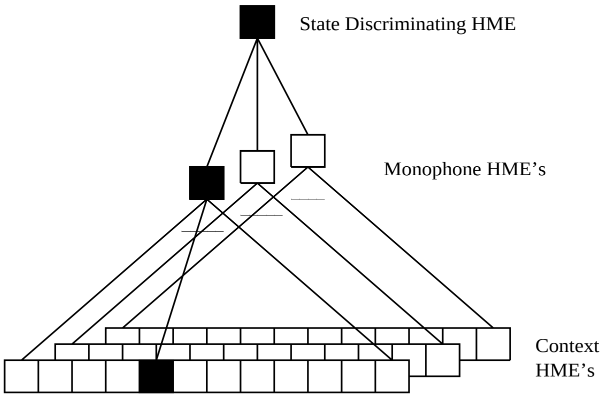


Fig. 2: Overview: Structure of HME set

In order to compute a specific context-dependent posterior class probability for an $m$-state HMM topology, a sequence of three HME evaluations is necessary (depicted as black boxes in the above figure). The resulting network outputs are divided by the respective class priors before being multiplied together to form an estimate of the scaled observation likelihood. Smoothing factors might be introduced for the context-dependent HME's, in order to accomodate different dynamic ranges of context-dependent and context-independent network outputs.

## 4. POLYPHONE CONTEXT CLASSES

For each monophone in each state, we need to define a set of context classes which are to be modeled by the method described above. As in [7], we use phonetic decision trees to cluster phonetic contexts. However, our work differs in two aspects: (1) our system clusters polyphone instead of just triphone contexts, (2) the decision trees are adopted from a continuous density HMM system. The splitting criterion for growing the decision trees is based on weighted gain in entropy between the discrete probability distributions (the mixture coefficients in the Gaussian mixtures) before and after a potential split.

$$\begin{aligned} D(\mathbf{p}, \mathbf{p}_l, \mathbf{p}_r) &= n_l H_l(\mathbf{p}_l) + n_r H_r(\mathbf{p}_r) - nH(\mathbf{p}) \\ \text{with} \quad H_l(\mathbf{p}_l) &= -\sum_i p_{li} \log p_{li} \\ H_r(\mathbf{p}_r) &= -\sum_i p_{ri} \log p_{ri} \\ H(\mathbf{p}) &= -\sum_i p_i \log p_i \end{aligned}$$

where $\mathbf{p}$ is the vector of mixture coefficients before and $\mathbf{p}_l, \mathbf{p}_r$ are the vectors of mixture coefficients resulting from the separate modeling in the two children nodes after a split. Potential splits are generated by asking phonetic questions in polyphonic contexts, with the restriction of only one phone across word boundaries. The following figure shows an example of such a cluster tree. Internal nodes contain phonetic questions (numbers in questions are positions relative to the current monophone), leaves contain model names.

After the polyphone clustering decision tree has been grown within the standard HMM system, a set of corresponding context expert HME's for the hybrid system can be build and trained. In the case of the tree in Fig. 3, we would create and train an HME with 9 output nodes (one for each context class).

## 5. SMOOTHING CONTEXT POSTERIORS

In order to compensate different dynamic ranges of monophone and context posteriors, we are using a smoothing method for context-dependent posteriors based on a binomial model. The likelihood estimation is modified to include a monophone and state dependent scaling factor $\gamma_{ik}$ with $0.0 \leq \gamma_{ik} \leq 1.0$:

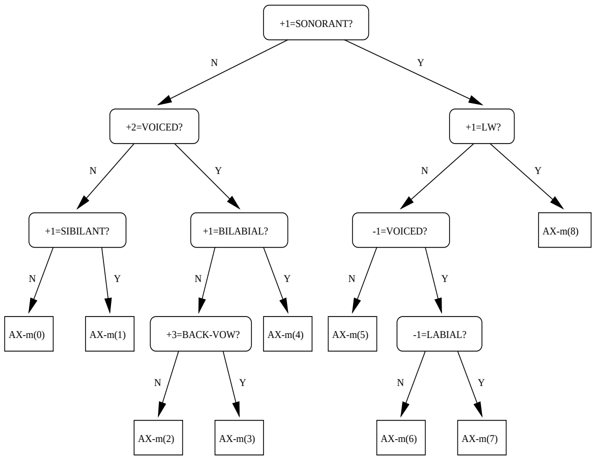


Fig. 3: Example: Polyphone Cluster Tree for middle state of monophone AX

$$p(c_j, \omega_i, s_k | \mathbf{x}) = \left[p_{ijk}^{CD}\right]^{\gamma_{ik}} \left[p_{ik}^{CI}\right]^{1-\gamma_{ik}}$$

$$\text{with} \quad p_{ijk}^{CD} = p(c_j | \omega_i, s_k, \mathbf{x})$$

$$\text{and} \quad p_{ik}^{CI} = p(\omega_i | s_k, \mathbf{x}) \quad p(s_k | \mathbf{x})$$

In log-space, this method of smoothing simplifies to a linear interpolation between the two logarithmized posterior estimates. A smoothing factor $\gamma = 0.5$ corresponds to the original likelihood estimation, where context-dependent and context-independent posteriors are weighted equally. As $\gamma$ goes towards zero, the contribution of the context-dependent HME's is reduced. For $\gamma = 0.0$ the system degenerates to a context-independent system, context-dependent posterior estimates are fully suppressed.

Weighting factors $\gamma_{ik}$ can be estimated iteratively using stochastic gradient descent to minimize a frame classification error function. Using MSE $E^{(t)} = 0.5[p_{ijk}^{(t)} - q_{ijk}^{(t)}]^2$, one can derive the following update rule:

$$\gamma_{ik}^{(t+1)} = \gamma_{ik}^{(t)} - \eta \Delta_{ik}^{(t)}$$

$$\text{with} \quad \Delta_{ik}^{(t)} = (p_{ijk}^{(t)} - q_{ijk}^{(t)}) p_{ijk}^{(t)} (\log p_{ijk}^{CD} - \log p_{ik}^{CI})$$

where $\eta$ is a small learning rate, $q_{ijk}^{(t)}$ is the desired output and $p_{ijk}^{(t)}$ is the smoothed posterior estimate.

## 6. EVALUATION

The hybrid HME/HMM system was implemented as part of the JANUS [8] HMM recognizer and evaluated on the English Spontaneous Scheduling Task (ESST), a 2500 word spontaneous speech database containing over 25 hours of speech. The system uses 3-state left-right

HMM's and 51 monophones. The connectionist part consists of one state discriminating HME, 3 monophone HME's and 3x51 context expert HME's. To reduce training and testing complexity, our context HME's consist of only one multinomial GLIM node. This allows us to train the context HME's in about 4-6 hours on a standard workstation. For each CI system, we evaluated two context systems with 500 and 1000 context classes, respectively. For the CI systems, we experimented with four different architectures: Two GLIM-based HME systems, one with HME's of depth 1, branching factor 16, the other with HME's of depth 2, branching factor 4, and two MLP-based HME systems with HME's of depth 1, branching factor 4. Training of these HME's took between 24 and 40 hours, also on standard workstations. The HME's were trained along labels which were generated by our continuous-density HMM recognizer. The following table shows results for the different systems, numbers are word accuracies (WA). The system name is encoded as [node-parameterization]-[depth]-[branching factor] (GD denotes a gender-dependent system).

| | CI | CD-500 | CD-1000 |
|---|---|---|---|
| GLIM-1-16 | 57.5% | 60.6% | 63.0% |
| #param | 370k | 420k | 510k |
| GLIM-2-4 | 57.7% | 60.8% | 63.8% |
| #param | 420k | 500k | 580k |
| MLP-1-4 | 60.8% | 61.7% | 64.1% |
| #param | 962k | 1.06M | 1.14M |
| MLP-1-4-GD | 63.2% | 66.5% | 68.3% |
| #param | 2.0M | 2.16M | 2.32M |

We achieved our best results with the GD-MLP-based HME's. Note, that the additional context modeling improves performance by as much as 10.3 %, relative to the CI system. A continuous density HMM JANUS system which models 5 times more context classes (5000) achieves 73.1% WA on this task (containing 4.26M acoustic parameters) at the expense of higher decoding time requirements. Decoding speed is about 2-5 times faster for the hybrid system.

We started to investigate the effect of smoothing of context-dependent posterior estimates as proposed earlier. Here, we report first results, where we used a single smoothing factor $\gamma = \gamma_{ik}$ for all context HME's.

The effect of this kind of smoothing can be seen in Fig. 4, which shows the word accuracy for different global smoothing factors applied to the MLP-1-4 CD-1000 system. A smoothing factor of $\gamma = 0.8$ yielded an absolute increase in WA of 1.1%.

## 7. CONCLUSIONS

We presented a highly modular context-dependent hybrid HMM system, which outperforms its context-independent version significantly. This encourages us to further investigate and improve the hybrid system. The ultimate goal is, to improve overall performance by combining HME- and Gaussians-based scoring the same way, expert networks are combined in an HME.

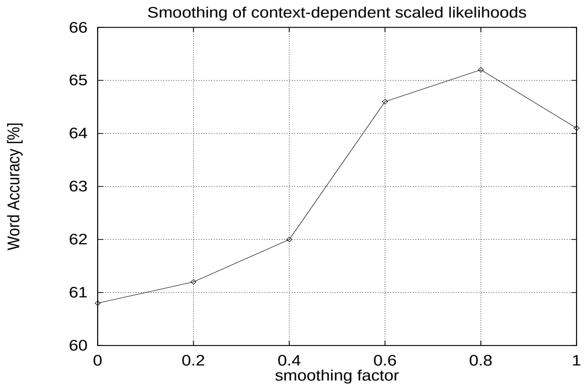


Fig. 4: Smoothing context-dependent posteriors

## REFERENCES


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

[2] Franco, H., Cohen, M., Morgan, N., Rumelhart, D. & Abrash V. (1994) *Context-dependent connectionist probability estimation in a hybrid Hidden Markov Model – Neural Net speech recognition system.* Computer Speech and Language, Vol. 8, No 3, 211-222, July 1994.

[3] Hampshire II, J. B., Waibel A. H. (1989) *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.

[4] Hochberg, M. M., Cook, G. D., Renals, S. J., Robinson, A. J. & Schechtman, R. S. (1995) *The 1994 ABBOT Hybrid Connectionist-HMM Large-Vocabulary Recognition System.* In Spoken Language Systems Technology Workshop, 170-176, ARPA, Jan. 1995.

[5] Jordan, M. I., Jacobs, R. A. (1992) *Hierarchies of adaptive experts.* In Advances in Neural Information Processing Systems 4, J. Moody, S. Hanson & R. Lippmann, eds., pp. 985-993. Morgan Kaufmann, San Mateo, CA.

[6] Jordan, M. I., Jacobs, R. A. (1994) *Hierarchical Mixtures of Experts and the EM algorithm.*, Neural Computation 6, 181-214, MIT Press.

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

[8] Waibel et. al. (1996) *JANUS-II - Advances in Spontaneous Speech Translation.* Internat. Conf. Acoustics, Speech and Signal Proc., May 1996, Atlanta, Georgia.

[9] Zhao, Y., Schwartz, R., Sroka, J. & Makhoul, J. (1995) *Hierarchical Mixtures of Experts Methodology Applied to Continuous Speech Recognition.* Internat. Conf. Acoustics Speech and Signal Proc., Vol 5, 3443-3446, May 1995.