IMPROVING PERFORMANCE ON SWITCHBOARD BY COMBINING HYBRID HME/HMM AND MIXTURE OF GAUSSIANS ACOUSTIC MODELS

Jürgen Fritsch, Michael Finke
{fritsch,finkem}@ira.uka.de

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

ABSTRACT

This paper presents results of our efforts on combining standard mixture of Gaussians acoustic modeling [10] with a context-dependent hybrid connectionist HME/HMM architecture [3, 4] for the Switchboard corpus. Using a score normalization scheme which is independent of the stream's modeling paradigm and adaptive methods for combining multiple probability distributions, we achieve a relative decrease in word error rate of 3.5% and 9.3%, compared to each of the single stream systems. As opposed to multiple acoustic streams based on mixture of Gaussians, the integration of hybrid NN/HMM based modeling appears to be advantageous since the differences in modeling techniques and training algorithms allow to capture different aspects of the speech signal. Small dependence among emission probability estimates is considered essential for potential gains in interpolated systems.

1. INTRODUCTION

Recognizing spontaneous conversational telephone speech is one of the most challenging fields being tackled by the speech recognition community. Sites achieved word error rates ranging from 38.8% to 47.1% (Switchboard 1996, [8]) and from 44.9% to 51.6% (Switchboard + CallHome 1997) in NIST's recent Hub-5E LVCSR evaluations based on the Switchboard and CallHome spontaneous telephone speech corpora. These figures compare to only 7-10% error rates achievable on high-quality read speech (e.g. on the Wall Street Journal domain). Reasons for the large gap in performance include differences in channel quality, variances in word pronunciations due to regional dialect and speaking rate and the casual style of conversation. Lots of spontaneous speech phenomena such as false-starts and interjections can be observed in the Switchboard and CallHome corpora.

Our recognizer for the Switchboard domain is based on the Janus Recognition Toolkit (JRTk) [10]. We recently extended the acoustic modeling techniques available in JRTk by integrating a context-dependent hybrid neural network/HMM system [3, 4]. The hybrid is based on generalized Hierarchical Mixtures of Experts (HME) [7, 3] which are used as statistical estimators of HMM observation probabilities. Using polyphonic decision trees and a factorization technique for posterior probabilities, acoustic contexts can be modeled accurately within the NN/HMM framework. Using this technology, the HME/HMM system has recently become competitive with systems based on mixture of Gaussians acoustic models.

Work by other groups (e.g. [11]) has shown that combinations of different architectures for acoustic modeling can improve the overall performance, even in the case when only static linear interpolation with constant weighting factors is applied. When trying to combine heterogenous systems such as a mixture of Gaussians and a hybrid NN/HMM system, one has to take care of the fact that the latter is estimating scaled likelihoods which can not easily be combined with 'true' likelihoods. In our experiments we are therefore applying a score normalization technique, before combining multiple acoustic streams.

Viewing streams based on different modeling paradigms as experts for acoustic observations, we can introduce a gating mechanism similar to the one used in HME's or in the Meta-Pi [5] architecture, to merge the outputs of such experts. A major difficulty of this technique, known in statistics as linear opinion pooling [6], is the need for relatively independent expert opinions. We are trying to address this issue by combining heterogenous systems. However, independence of observation probability estimates can not be expected since the systems are trained on the same data.

In the remainder, we present details of systems and normalization/combination methods, finally reporting first results of our experiments on Switchboard.

2. MIXTURE OF GAUSSIANS MODELING

This section describes the acoustic modeling part of our current Janus Recognition Toolkit (JRTk) based mixture of Gaussians recognizer for the Hub-5E Switchboard + CallHome task.

- **Preprocessing:** Input coefficients consist of 13 MFCC's plus power and the first and second derivatives. We normalize for speaker dependent vocal tract lengths by frequency warping the spectrum based on a maximum likelihood approach (ML-VTLN). Cepstral mean normalization is used to compensate for different channels. After transforming the resulting 42 dimensional feature vector using an LDA matrix, the first 32 coefficients are used as the final feature vector.
- **Acoustic/phonetic modeling:** The system uses 3-state left to right tied-state continuous density HMM's. Context-dependent acoustic modeling has been augmented from standard triphone to polyphone modeling. In the case of the Switchboard recognizer, polyphone modeling is realized by allowing questions in the allophonic decision tree refering to a maximum of 2 phones to each side of a phone. Clustering of the polyphonic decision tree uses an entropy based splitting criterion and is carried out in a two pass procedure, where codebooks are clustered in an initial step, followed by the clustering of a larger number of distributions which are sharing codebooks [1]. The system used for the experiments reported here

consists of 24k distributions and 6k codebooks, each containing 16 diagonal-covariance Gaussians.

- **Adaptation:** The system uses MLLR based unsupervised adaptation on confidence measure weighted first hypotheses generated by the baseline system. Adaptation training data are confined to words which are considered error free according to the confidence measure. The number of MLLR transformations actually used to adapt to a specific speaker is determined automatically using a clustering algorithm on the pooled set of Gaussians. This algorithm adapts the number of MLLR transformations to the amount of adaptation data available [10].

This recognizer was one of the best performing systems participating in the Hub-5E 1996 and 1997 evaluations [8]. Since at the time of the experiments for this paper, adaptation was not available for the hybrid HME/HMM system, we were not using MLLR for the mixture of Gaussians system either.

3. HYBRID HME/HMM MODELING

The hierarchical mixtures of experts (HME) architecture is a modular neural network suitable for supervised learning. Jordan and Jacobs [7] introduced the hierarchical mixtures of experts for solving non-linear regression problems emphasizing the divide-and-conquer strategy. In our system, HME's are used as classifiers which requires different expert parameterizations. In an HME, the learning task is divided into 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 means, 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.

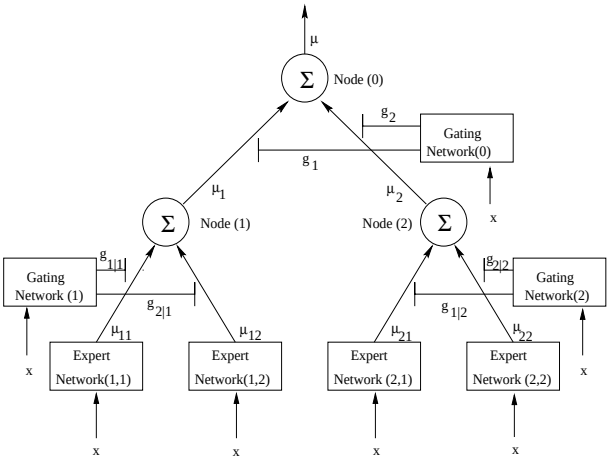

Fig. 1: Hierarchical Mixtures of Experts

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 networks in an HME model multinomial probabilities and are therefore parameterized using the softmax nonlinearity ('canonical link' in GLIM theory):

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

In [7] the $y_i(\mathbf{x})$ are parameterized as linear models, leading to an efficient EM training algorithm (iteratively reweighted weighted least squares) for the hierarchy. However, we discovered that it is sometimes advantageous to use more complex parameterizations for gates and experts. In our system, arbitrary multi-layer feed-forward architectures using projective and/or radial kernels can be applied. Such generalized HME architectures can still be trained efficiently using generalized EM algorithms with on-line updates.

So far, we have described the basic modeling units of the hybrid HME/HMM recognizer. Many research groups have experimented with hybrid systems based on all kinds of neural network models, mostly using a single network to estimate monophone posteriors (sometimes augmented by additional context networks). In our system, we are using a more general setup for the estimation of scaled likelihoods which allows to model any number of acoustic models by adopting a decision tree based context clustering as being used by most standard HMM systems. Given a specific number of acoustic models, a tree of neural networks (in our case generalized HME's) is used to model the required posteriors, which are then converted to scaled likelihoods by dividing by priors. The application of a tree of networks is justified by a factorization of acoustic model posteriors. For example, consider a system which models a set of monophones ω_i in particular contexts c_{ij} using multi-state HMM's with states s_{ijk}. The HMM requires the modeling of the following conditional likelihood, which can be factorized as follows:

$$\begin{aligned} & p(\mathbf{x}|\omega_i, c_{ij}, s_{ijk}) \\ = & \frac{p(\omega_i, c_{ij}, s_{ijk}|\mathbf{x})}{P(\omega_i, c_{ij}, s_{ijk})} \quad p(\mathbf{x}) \\ = & \frac{p(c_{ij}, s_{ijk}|\omega_i, \mathbf{x})}{P(c_{ij}, s_{ijk}|\omega_i)} \quad \frac{p(\omega_i|\mathbf{x})}{P(\omega_i)} \quad p(\mathbf{x}) \\ = & \frac{p(s_{ijk}|\omega_i, c_{ij}, \mathbf{x})}{P(s_{ijk}|\omega_i, c_{ij})} \quad \frac{p(c_{ij}|\omega_i, \mathbf{x})}{P(c_{ij}|\omega_i)} \quad \frac{p(\omega_i|\mathbf{x})}{P(\omega_i)} \quad p(\mathbf{x}) \end{aligned}$$

In our system, this particular factorization (other orders of factorization are possible) would be modeled by a depth-3 tree of neural networks, having a single monophone discriminator at the root node, a set of context discriminators (one network for each monophone) in the first level and a set of state discriminators (e.g. modeling the probability of being in the beginning, middle or end state of an HMM) in the second level. Computation of a specific conditional likelihood then requires the evaluation of three HME's. Class priors are estimated by their relative frequency in the training set.

For our Switchboard HME/HMM system, we are using a more refined factorization corresponding to a deeper tree of networks. The following figure shows parts of this tree structure. Each rectangular node contains a single HME

with different architecture, depending on the amount of training data available (HME architectures are ranging from single node GLIM's used as state discriminators up to a depth-2 two-layer MLP based HME with branching factor 4 as monophone discriminator). Also, as with all hybrid systems based on relatively large neural networks, we have to find a trade-off between model complexity and available computing resources. However, the multi-network tree can easily be trained in parallel on multiple workstations since the NN nodes are independent.

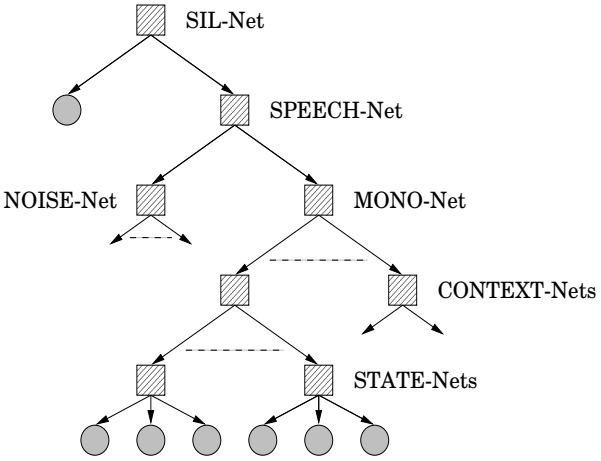

Fig. 1: Tree of networks for the HME/HMM system

The HME/HMM hybrid used for the experiments in this paper did not contain any unsupervised adaptation procedure. We are currently investigating the viability of linear front end adaptation networks.

4. STREAM NORMALIZATION

Since the hybrid HME/HMM system is estimating scaled likelihoods instead of 'true' likelihoods, it is difficult to compare and combine scores of heterogenous streams directly. This fact is illustrated in the following figure, which plots the empirical distribution of stream scores (negative log-domain) for the hybrid HME/HMM and the mixture of Gaussians (MOG) based systems.

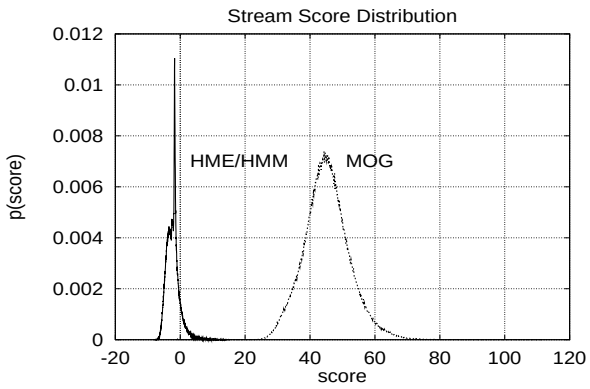

Fig. 1: Stream score distributions

While the distribution of MOG scores seems to smoothly follow a Gaussian like statistic, the distribution of HME/HMM scores contains bumps and a very strong peak near zero. This peak is attributable to the (very frequent) silence model and the MAP training process, which leads to good discrimination of silence and speech. To compensate the differences in distribution, we are using a histogram based normalization technique, which produces estimates of the following probability mass function on the stream scores s_i (assuming negative (scaled) log probabilities):

$$p(s_i > S|\mathbf{x}, \Theta_i)$$

Since lower values of s_i correspond to better scores, this technique effectively normalizes any kind of stream scores to the range [0,1]. The following figure shows the resulting mapping functions for the two systems considered in this paper.

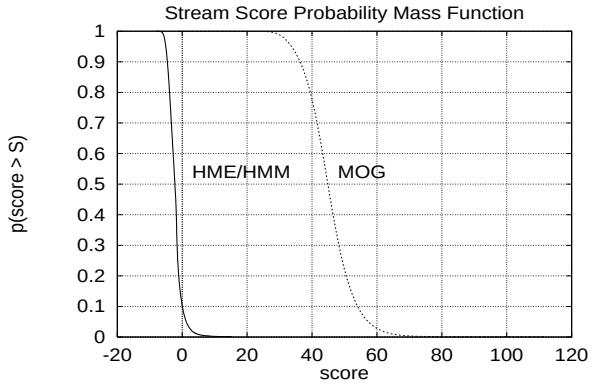

Fig. 1: Empirical stream normalization functions

Finally, we note that this normalization method corresponds to a monotonic mapping which preserves the order of scored models while maximizing the entropy of the resulting normalized stream scores by approximating a uniform distribution. For the following section on combination methods, we are assuming that stream scores have been normalized using the above method.

5. COMBINATION METHODS

A simple, yet effective combination method is the application of static linear interpolation of normalized scores p_i:

$$p(\mathbf{x}|\Theta) = \sum_i \gamma_i p_i(\mathbf{x}|\Theta_i) \quad with \quad \sum_i \gamma_i = 1$$

When using a single set of γ_i's shared among all acoustic models, one can empirically determine the best interpolation weights. The above technique can be enhanced by allowing different acoustic models to have different interpolation weights. This, however, requires a learning algorithm to adjust these parameters in order to maximize an objective function such as the likelihood of the combined system. Work by other people (e.g. [9]) has shown that the estimation of model-dependent interpolation weights requires a discriminative training procedure which can become computationally prohibitive.

The above model of static linear interpolation (also called linear opinion pooling) can be modified to allow for interpolation weights g_i which are functions of some feature vector $\mathbf{z}$:

$$p(\mathbf{x}|\Theta) = \sum_i g_i(\mathbf{z}) p_i(\mathbf{x}|\Theta_i) \quad with \quad \sum_i g_i(\mathbf{z}) = 1$$

This formulation of stream combination is equivalent to a mixtures of experts approach and we can therefore use

the same framework for stream interpolation as for the training of (H)ME's. A reasonably complex gating network can be used to learn gating probabilities by first computing stream posteriors

$$h_i = \frac{g_i(\mathbf{z})p_i(\mathbf{x}|\Theta_i)}{\sum_j g_j(\mathbf{z})p_j(\mathbf{x}|\Theta_j)}$$

which are then acting as target values during the training of the network. Again, this method of adaptive interpolation can be made model-dependent by allowing the application of different gating networks for different acoustic models.

If word confidence measures are available by the recognizer, the combination of streams can be guided by the confidence of each one of the acoustic streams into hypotheses obtained by separate decoding passes using each one of the acoustic streams stand-alone. The individual confidence scores in each frame can either be converted to gating posteriors or be used to decide on switching entirely from one stream to the other.

Finally, all of the presented combination methods require to convert the interpolated stream scores back to the log-domain for decoding.

6. EXPERIMENTS & RESULTS

All experiments were carried out on the official 1997 SWB+CH development test set which consists of 80 conversation sides. In order to be able to compare different approaches in a reasonable amount of time, we restricted this test set further, by using only the first 30 seconds of speech from each conversation side. The hybrid HME/HMM system uses the same preprocessing as the MOG system but concatenates 7 adjacent MFCC's to obtain its final feature vector instead of applying an LDA transform. A more detailed description of the recognition engine (decoder,lexicon,LM) can be found in [2].
The following table summarizes results obtained by applying some of the stream combination methods described earlier. The first two rows, termed MOG and HME list the word error rates for the stand-alone systems. The next three rows contain results for combined systems.

System	SWB WER	CH WER	SWB+CH WER
MOG	34.1	47.3	40.2
HME	37.3	49.3	42.8
SLI	33.0	45.7	38.9
MEI	33.1	45.5	38.8
CMI	33.9	47.0	40.0

The MOG system used in the above experiments differs from our current best system (used in the 1997 Hub-5E evaluation) in several respects. The evaluation system uses an expanded set of phones, refined acoustic/phonetic models, MLLR adaptation and multiple interpolated language models. For comparison, we tested this evaluation system on the test set used in this paper, where it achieves a word error rate of 31.7%. Despite the much better performance of this system, static linear interpolation with the HME system still gave a relative improvement of 1.5% on the 1997 evaluation set.

We now describe the three combined systems that were tested:

- **Static Linear Interpolation (SLI):** Constant interpolation weights of 0.6 (MOG) and 0.4 (HME)
- **Mixtures of Experts Interpolation (MEI):** Using a single two-layer gating MLP with 64 hidden units and 32 input features (the same as for the MOG system). Extracting appropriate features for a gating network seems very important for this technique and further work in this direction is necessary. Also, the application of separate gating networks for different classes of models might be worth investigating.
- **Confidence Measure Interpolation (CMI):** Here, we were using word confidence measures (estimates of the posterior word correct probabilities) from the MOG system to weight the two systems.

7. CONCLUSIONS

We present a competitive speech recognition architecture for the Switchboard telephone speech corpus. Using methods for adaptive combination of two different acoustic modeling paradigms, the system achieves a word error rate of 38.8% on a subset of the 1997 SWB+CH development test set, improving the recognition rate by 3.5% relative to the mixture of Gaussians system. The context-dependent HME/HMM system evaluated as a stand-alone system achieves a word error rate of 42.8% on the same test set.

REFERENCES

[1] Finke M., Rogina I.: *Wide Context Acoustic Modeling in Read vs. Spontaneous Speech*, ICASSP 97, Munich, Germany.

[2] Finke M., Fritsch J., Geutner P., Ries K., Zeppenfeld T., Waibel A.: *The JanusRTk Switchboard/Callhome 1997 Evaluation System*, Proceedings of LVCSR Hub-5E workshop, May. 13-15, Baltimore, Maryland.

[3] Fritsch J.: *Modular Neural Networks for Speech Recognition*, Tech.Rep. CMU-CS-96-203, Carnegie Mellon University, Pittsburgh PA, August 1996.

[4] Fritsch J., Finke M., Waibel A.: *Context-Dependent Hybrid HME/HMM Speech Recognition Using Polyphone Clustering Decision Trees*, ICASSP 97, Munich, Germany.

[5] Hampshire II, J. B., Waibel A. H. *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.

[6] Jacobs, R. A. *Methods for Combining Experts' Probability Assessments*, Neural Computation 7, 867-888, MIT Press, 1995.

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

[8] Proceedings of LVCSR Hub 5 workshop, Apr. 29 - May 1, MITAGS, Linthicum Heights, Maryland.

[9] Rogina, I., Waibel, A.: *Learning State-Dependent Stream Weights for Multi-Codebook HMM Speech Recognition Systems*, ICASSP 94, Adelaide, Australia.

[10] Zeppenfeld T., Finke M., Ries K., Westphal M., Waibel A.: *Recognition of Conversational Telephone Speech Using the Janus Speech Engine*, ICASSP 97, Munich, Germany.

[11] Zhao, Y., Schwartz, R., Sroka, J. & Makhoul, J. *Hierarchical Mixtures of Experts Methodology Applied to Continuous Speech Recognition*, ICASSP 95, Detroit, USA.