# Hierarchical Neural Networks (HNN) for Chinese Continuous Speech Recognition

**Ying Jia, Limin Du, Ziqiang Hou**
Lab of Interactive Information System
Institute of Acoustics, Chinese Academy of Science
Beijing, 100080, P. R. China
Email: jiaying@farad.ioa.ac.cn, xss@dsp.ac.cn

**Abstract**

To integrate the hierarchy structure of discrimination between all HMM states for Chinese Initials and Finals, we constructed in this paper Hierarchical Neural Networks (HNN), which differ from Jordan's HME in such extensions as more complex parameterization for gate and/or expert and dimension-reduced expert network. With these extensions, we can reuse those pre-trained simple node networks in a hierarchy structure (HNN), and fine-tune them jointly by Generalized Expectation Maximization (GEM) algorithm. The proposed HNNs were used within hybrid HMM-ANN models to perform the estimation of posterior probabilities for HMM states. Instead of using a large monolithic neural network, the HNN system can be trained in a short time compared with MLP estimator and result in a speed-up in decoding time over the conventional systems. We have applied the proposed hybrid HMM-HNN method to the recognition task of Chinese Continuous Speech., achieve a promising word error rate of 26.4%.

## 1. Introduction

It has been shown both theoretically and practically that Hybrid HMM-ANN 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 flexible with taking into account the correction between the successive feature vectors, and context-dependency modeling, and are compact, offering faster decoding speeds than standard systems. Unfortunately training of such Big Dumb Neural Networks requires orders of magnitude more computation than the more common HMM training paradigm. This fact is most obvious in Large Vocabulary Continuous Speech Recognition tasks.

The large monolithic neural network in hybrid HMM-ANN models tends to discriminate each state from all other competitive states within a same output layer at frame level. In fact, it has been observed that a state was often been confused with a small set of other HMM states. For instance, 'p' is often been confused with 'd', and 'g'. These observations inspire us to exploit the structure information of discrimination by integrating the "divide and conquer" philosophy.

Jordan and Jacobs introduced the Hierarchical Mixture of Experts (HME) to solve the regression problems using the divide-and-conquer strategy. It has been observed that HMEs for the LVCSR tasks are inherent with a mismatch of the pre-defined hierarchy structure and the simple node networks (Generalized Linear Models).

In this paper, we constructed a Hierarchical Neural Network (HNN) with such extension of HME as more complex parameterizations for gates/experts and dimension-reduced expert network. Benefited from these extensions, we can easily integrate the knowledge of perceptual confusions among Chinese Initials and Finals developed by Jialu Zhang into the HNN construction.

## 2. Hierarchical Structure of Discrimination Between Chinese Initials and Finals

A cluster analysis of perceptual features of Chinese speech sounds carried out by Jialu Zhang etc. in 1982 has revealed that:

1) Chinese Initials with the same manner of articulation are frequently confused with each other, and different perceptual features, such as voiced and Nasality etc., are of different importance on the perception of Initials as shown in figure 2;

2) Chinese Finals with similar endings have great psychological similarity, and are easily confused with each other, as shown in figure 3;

3) The combination rules of Chinese syllables are of some effects on the perception of Chinese Initials and Finals.

The above conclusions have also been conformed by the analysis of confusion matrices gained from the HMM-based Chinese Speech Recognition systems [4].

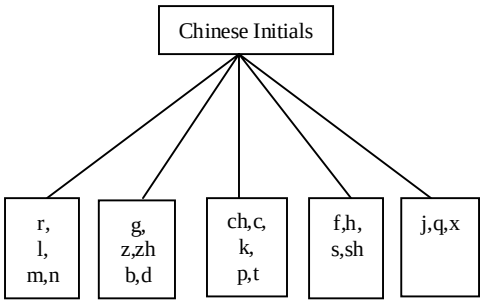


Figure 2. The hierarchical clustering tree of Chinese initials

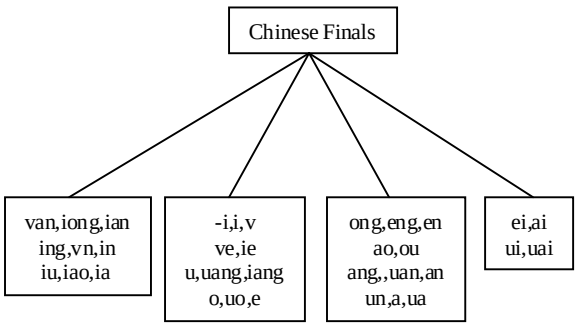


Figure 3. The hierarchical clustering tree of Chinese finals

## 3. Integrating the Hierarchical Structure into Posterior Estimation

With the structure information about perception of Chinese Initials and Finals, we can divide the complete set of Chinese Initials and Finals into some easily confused groups of units, which are hierarchically organized into a Confusion Tree. Further more, we can divide the task of discrimination among the whole set of Chinese Initial and Finals into some subtasks that correspond to confusion groups, and the subtasks can be done using relatively small neural networks. These ideas can be applied to the posterior estimation formulated as following:

Let $Q$ denote the set of all HMM states $q_i$. Consider we split $Q$ into $N$ disjoint and non-empty subsets $Q_n$. A particular state $q_i$ will now be a member of $Q$ and exactly one of the subsets $Q_n$. Therefore, we can rewrite the posterior probability of states $q_i$ as a joint probability of state and appropriate subsets $Q_n$ and factor it according to

$$p\left(q_i \middle| X_{t-c}^{t+c}\right) = p\left(q_i, Q_n \middle| X_{t-c}^{t+c}\right) = p\left(Q_n \middle| X_{t-c}^{t+c}\right) p\left(q_i \middle| Q_n, X_{t-c}^{t+c}\right)$$

Thus, the global task of discriminating between all the states in $Q$ has been converted into two subtasks. One is discriminating between subsets $Q_n$ and another is independently discriminating between the states $q_i$ contained within each of the subsets $Q_n$. Recursively repeating this process yields a hierarchical tree-organized structure.

From section 2, we know that for Chinese Initials and Finals the constituent of $Q$ and $Q_n$ are stable and the discrimination within each subset $Q_n$ can be done by a relative small neural networks.

## 3. Hierarchical Neural Networks

Hierarchical Neural Networks (HNN) was proposed to exploit the hierarchy structure of Jordan's HME [1]. With such extensions of HME as more complex parameterizations for gates and/or expert [5] and dimension-reduced expert network [6], it's very convenient for us to integrate knowledge about class

distribution in feature space into the HNN construction. Inheriting the hierarchy of HME, HNNs still follow the principle of divide-and-conquer. Fig. 1 shows a binary branching HNN with 2 level, every node in the HNN is a multi-layered feed-forward network with single hidden layer.

A one-depth HNN performs the following computation:

$$P(l \mid x,s) = \sum_{i=1}^{C} P(l \mid c_i, x, s) P(c_i \mid x, s)$$

where $i = 1, \cdots, L$ indicates phoneme class, $c_i$ represents a local region in the input space, and $C$ is the number of regions. $P(c_i \mid x, s)$ can be viewed as a gating network, while $P(l \mid c_i, x, s)$ can be viewed as a local expert classifier (expert network) in the region $c_i$. In a two-level HNN, each region $c_i$ is divided in turn into $C$ sub-regions. The term $P(l \mid c_i, x, s)$ is then computed in a similar manner to equation (1), and

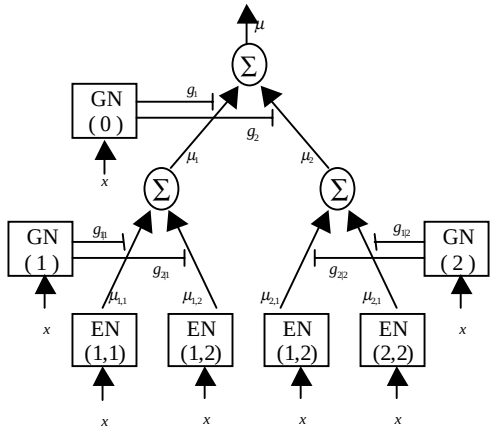

Fig. 1, a 2-branching HME with 2 depth

so on. If in some of these sub-regions there are no data available, we back off to the parent network.

The HNN can still be trained efficiently using the Generalized Expectation Maximization (GEM) algorithms with on-line updates, which is an iterative approach to maximum likelihood estimation. Each iteration of an EM algorithm is composed of two steps: an Expectation step (E-Step) and a Maximization Step (M-Step). The E-Step of the EM algorithm reduces to computing the expected values of the indicator variable that gives the set of posterior probabilities $\{h\}$. The M-Step reduces to a set of independently weighted maximum likelihood problems for the experts and gates. Thus the weight for GN (1), at time $t$, is the joint posterior probability of this node $h_1^{(t)}$, and the weight for EN(1,1) is the joint posterior probability of this node $h_{11}^{(t)}$. The target outputs for the gating networks are the conditional posterior probabilities of the node in question, and the targets for compact experts are the dimension-reduced versions $y_{i,j}$ of the global target output $y$.

Since each expert network and gating network is a MLP with single hidden layer, We can use Gradient Ascent algorithms to update the parameters for each node network independently. Here we realize the derivation of a set of separate likelihood functions (the cross-entropy between the posterior probabilities and the prior probabilities weighted by its path probability) for each gating or expert with respect to their parameters as on-line method. The likelihood to be maximized for EN (i,j) is

$$l(\theta_{i,j}) = \sum_{t}^{t+\Delta t} h_{ij}^{(t)} \ln P_{ij}\left(Y^{(t)}\right)$$

where $P_{ij}$ was assumed to be multinomial density. For the top-level gating network, the likelihood is

$$l(v) = \sum_{t}^{t+\Delta t} \sum_{l} h_l^{(t)} \ln g_l^{(t)}$$

The likelihood function for the second level gating network (i=1,2) is

$$l(v_i) = \sum_{t}^{t+\Delta t} h_i^{(t)} \sum_{m} h_{m|i}^{(t)} \ln g_{m|i}^{(t)}$$

In the above equations, parameters were updated after having accumulated $\Delta t$ samples.

Once the maximum likelihood problems of the M-Step have been completed, the E-step is repeated, computing a new set of posteriors $\{h\}$ for the next $\Delta t$ observations, which become the new weights for the M-Step.

## 4. Experiments On Chinese Continuous Speech

Performance of the proposed method was examined by simulations on a Chinese continuous speech database uttered by a single male speaker. All speech signals were sampled at a rate of 16kHz and pre-emphasized with a digital filter $1-0.98z^{-1}$. The recognition feature include the first 12 cepstral coefficients calculated over frames of length 10 ms. Each frame consists of 160 samples of speech signal and there is a 5 ms overlap between adjacent frames.

In our initial application of the HNN method to the recognition of Chinese continuous speech, we used phonetic context-independent HMEs to estimate the likelihood at each state of 2-state HMMs for initials and 4-state HMMs for finals. We implemented a 3-level HME, with the input space divided into 2 regions, and one region for Chinese Initials in further divided into 5 sub-regions, the other for Chinese Finals into 13 sub-regions, as shown in fig. 2 and fig.3. All gating and local expert networks in the HNN have identical structures — a MLP with single hidden layer.

To speed up the training of HNN, we first trained all of the MLPs included in the HNN separately and then combine them together to be fine-tuned using GEM algorithm. In fact, we just tuned the gate weights and fixed the expert weights.

We tested the HNN implementation on a test set of 250 sentences. The word error rates reported here were based on the same test set for a number of different systems. Table 1 shows the word error rate for i) the baseline HMM system; ii) the hybrid HMM/HME system with modified priors.

Table 1 : Comparison of HNN vs. MLP

| | Hybrid HMM-MLP system | Hybrid HMM-HNN system |
|---|---|---|
| Word error rate on test | 33.6% | 26.4% |
| Word error rate on tran | 28.4% | 23.6% |
| Training time | 12 days and nights | 6 days and nights |

From table 1, we can see that the Hybrid HMM/HNN system only used about half training time than hybrid HMM-MLP system but resulted in good performance in word recognition accuracy.

## 5. Conclusion

To integrate the hierarchy structure of discrimination between all HMM states for Chinese Initials and Finals, we constructed in this paper Hierarchical Neural Networks (HNN), which differ from Jordan's HME in such extensions as more complex parameterization for gate and/or expert and dimension-reduced expert network. With these extensions, we can reuse of the trained small networks and fine-tune them jointly in HNNs. Compared with large monolithic neural networks, the HNN system can be trained in a short time and use a faster decoding time

## References


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

[2] Michael I. Jordan, Robert A. Jacobs, *Hierarchical Mixtures of experts and the EM Algorithm*, Neural Computation6, 181-214(1994).

[3] Zhang Jialu, Lu Shinan,and Qi Shiqian, *A cluster analysis of the perceptual features of Chinese speech sounds*, Journal of Chinese Linguistics, Vol. 10, 1982. PP190-206.

[4] Chen Tao, *A Real-time Speaker-Dependent Syllable Recognition System of Complete Vocabulary of Mandarin*, Institute of Acoustics, CAS, 1991.

[5] Ying Jia, Limin Du, Ziqiang Hou, *Hierarchical Neural Networks and its Generalized Expectation Maximization Training Algorithm*, (to be submitted to) Chinese Journal of Electronics.

[6] Ying Jia, Limin Du, Ziqiang Hou, *Hierarchical Neural Networks with compact experts network*, (to be submitted to) Chinese Journal of Electronics.