TIED POSTERIORS: AN APPROACH FOR EFFECTIVE INTRODUCTION OF CONTEXT DEPENDENCY IN HYBRID NN/HMM LVCSR

Jörg Rottland, Gerhard Rigoll

Department of Computer Science
Faculty of Electrical Engineering
Gerhard-Mercator-University Duisburg, Germany
e-mail: {rottland,rigoll}@fb9-ti.uni-duisburg.de

ABSTRACT

This papers presents a method to improve the recognition rate of hybrid connectionist/HMM speech recognition systems. At the same time this approach allows the easy introduction of context dependent models in the hybrid framework. The approach is based on a standard hybrid connectionist/HMM recognizer, in which the neural nets are trained to estimate the a posteriori probabilities for all phones in each input frame. In the approach presented here, the probabilities of the neural nets are used to replace the codebook of a tied-mixture HMM system. Therefore the resulting system is called *tied posterior*. The advantages of this structure are that an arbitrary HMM-topology can be used, and that all context dependency and all clustering techniques used in tied-mixture systems can be applied to this hybrid speech recognition system. The approach has been evaluated on the Wall Street Journal (WSJ) database, with the result, that it outperforms the standard hybrid approach on this task.

1. INTRODUCTION

In recent years it could be shown, that the combination of neural networks with Hidden Markov Model (HMM) techniques is a powerful alternative approach to multi-mixture continuous HMMs. Due to the combination of the different methods the resulting system is called hybrid. In most hybrid speech recognizers the neural nets are used to estimate the a posteriori probabilities for the phones in each feature. The neural nets used for the estimation of the emission probabilities are usually either an MLP [1] or a recurrent neural net [2]. Those hybrid systems in which the neural network is used as a probability estimator and which are well known are referred to as *standard hybrid* speech recognizers in the following sections.

Although these systems have shown good recognition results, they still have two disadvantages compared to a multi-mixture continuous HMM system:

The first disadvantage is that the topology of the HMMs used in most of these standard hybrid systems is limited to a single HMM state for each model. This limitation leads to a very simple structure compared to other HMM systems, in which usually 3 states are used for each phone model. One state represents the beginning of the phone, one the static part of the phone and the third state is used to model the transition into the next phone. Due to the simple structure used in hybrid speech recognizers the modeling capabilities of the HMM are limited, specially concerning duration modeling and detailed modeling of the phone changes.

Another disadvantage of standard hybrid systems is based on the fact, that in the standard hybrid approach the neural net is trained in a manner, that each phonetic model corresponds to exactly one output of the neural net. I.e. the size of the output layer of the probability estimator equals the number of phonetic units used in the phonetic transcription. For simple systems without context dependency this structure leads to a small output layer of the neural net, usually between 40 and 70 neurons. When using context dependent models this structure becomes a problem, because then the huge number of models leads to a huge size of the output layer (e.g. 10000 triphones) of the neural net. This large output layer increases the number of parameters to be estimated in the neural net. The neural net used in [1] already has several hundred thousand parameters, so extending the output layer by a factor of more than 100 will result in more than 10 million parameters and thus this leads to sparse data problems for the training of these large neural networks. Additionally the time needed to train neural networks of this size will be too long, so that a training becomes impossible. Therefore the straightforward approach of building and clustering context dependent models, which is usually used in other HMM systems cannot be used with those hybrid systems, because it would lead to the problems described above. Therefore the use of context dependent models needs special algorithms and structures in these hybrid systems e.g. [3]. In the following section we will present an extension of the standard hybrid systems to over-

come the problems of the limited structure and the context dependent modeling described above.

2. TIED POSTERIORS

The approach presented in this paper is based on tied-mixture HMM technology. In these tied-mixture systems the emission probabilities b of each state S are computed by a weighted sum of a fixed number J of pdfs g_j, called the codebook [4].

$$b_i(\mathbf{x}) = p(\mathbf{x}|S_i) = \sum_{j=1}^{J} c_{ij} \cdot g_j(\mathbf{x}) = \sum_{j=1}^{J} c_{ij} \cdot p(\mathbf{x}|j) \quad (1)$$

The weighting factors c_{ij} are usually estimated using the maximum likelihood (ML) algorithm. In Eq. 1 the sum is computed by conditional probability $p(\mathbf{x}|j)$ multiplied with the factor c_{ij}. The conditional probability $p(\mathbf{x}|j)$ usually is a Gaussian pdf. When using a neural network as a probability estimator to replace those Gaussian pdfs the result will be some kind of hybrid speech recognizer [5]. This replacement is basically what has to be done to transform a tied-mixture system into a hybrid tied-posterior speech recognizer as proposed in this paper. In order to replace the conditional probability used in the HMMs with the posterior probability $P(j|\mathbf{x})$, which is the output of the neural network, a scaled likelihood as in [1] can be used. This probability can be expressed using the posterior probabilities $P(j|\mathbf{x})$ and the a priori class probabilities $P(j)$, which can be estimated using the training data:

$$\frac{p(\mathbf{x}|j)}{p(\mathbf{x})} = \frac{P(j|\mathbf{x})}{P(j)} \quad (2)$$

In the standard hybrid systems in [1] and [2] it has been shown, that neural networks can be used as probability estimators for posterior probabilities. Using those neural network posteriors together with Eq. 1 and Eq. 2 leads to the tied posterior approach, in which the emission probabilities can be computed as:

$$b_i(\mathbf{x}) = p(\mathbf{x}|S_i) = \sum_{j=1}^{J} c_{ij} \cdot \frac{P(j|\mathbf{x})}{P(j)} \quad (3)$$

In Eq. 3 the emission probabilities of the HMMs are calculated by a weighted sum of the outputs of neural networks scaled by the a priori class probability. The factors c_{ij} have to be estimated in the same way as for the tied-mixture case using the Baum-Welch algorithm.

When instead of the ML-estimates the c_{ij} are chosen to be

$$c_{ij} = 0 \quad \text{if} \quad i \neq j \quad \text{and} \quad c_{ij} = 1 \quad \text{if} \quad i = j \quad (4)$$

the hybrid tied posterior approach is transformed into a standard hybrid speech recognition system equivalent to the system considered in [1], in which only one of the network outputs is used for each HMM. That shows, that the standard hybrid system is a special case of the proposed tied-posterior speech recognizer.

When using the weights c_{ij} one has the advantage, that now the estimation of the posterior probabilities becomes independent of the HMM structure. This means, that in contrast with the standard hybrid approach where the HMM only consists of a single state, now it is possible to use HMMs with more than one state. The difference between those states is then described by different state dependent vectors of the weights c_{ij}. This opens the opportunity to build 3-state phone HMMs in a hybrid speech recognizer, as often used in other HMM systems for large vocabulary speech recognition.

The second advantage of this approach is that now it is possible to model context dependent HMMs without changing the neural net used for the probability estimation. This extension is a straightforward adaption of the context dependency used in tied mixture systems: First all monophone weights are copied to newly created context dependent HMMs which are based on that monophone. Then these context dependent posterior weights can be retrained using the Baum-Welch algorithm, or due to the large number of weights normally needed the weights can be clustered, which can be done either with a data-driven or a tree-based clustering procedure.

The neural network used as a probability estimator in this approach is the same network as would be used in standard hybrid approaches. This means that these hybrid systems can be transformed into a tied-posterior speech recognizer without a retraining of the neural weights. So this extension to the hybrid approach is a solution to the two major problems hybrid speech recognizers were suffering before.

The values of the weights c_{ij} used in Eq. 3 after the ML-estimation of those weights is similar to the values of Eq. 4, this means there is one weight, which is close to 1, few weights are above zero or a floor value, and the most values are set to zero or a floor value. Especially in the center state of the monophones the concentration of the weights to a very small number could be observed, whereas in the other states the spectrum of weights was wider. This result is basically what could be expected and shows, how the phone transitions in the first and last state are modeled. Furthermore this result shows, how this approach tries to handle misclassification of the neural network.

Further advantages of the tied posteriors are: The approach provides the most efficient way to combine the advantages of posterior-based hybrid speech recognition technology with the advantages of context dependent modeling. This context dependent modeling can be done with estab-

lished tools, such as tree-based clustering, and can be extended to all levels (e.g. cross-word triphones). Additionally this approach combines discriminative training techniques for the posteriors, with further optimization using Maximum Likelihood methods for the HMM parameters. Compared to a traditional tied-mixture system, the tied posterior approach can exploit the well-known multi-frame technique much more efficiently.

In the following sections the quality of this approach will be tested in some experiments.

3. EXPERIMENTS

The approach presented above has been evaluated using the *Wall Street Journal* (WSJ) database [6] of 1992 and 1993 (WSJ0 and WSJ1). The neural network used to estimate the posterior probabilities in the following experiments is a three layer MLP. The features computed from the speech data are 12-dimensional cepstral vectors plus the signal energy. Those features are computed every 10ms with a hamming window of 25ms. Additionally the first and second order derivatives of those 13 values have been used, which totals to 39 features in the feature vector of each speech frame. For the input layer of the MLP the features of 7 adjacent frames have been used, which results in a input layer size of 273 (7*39). The size of the hidden layer has been chosen to be 1000 and the size of the output layer is determined by the number of phones in the dictionary. The LIMSI-dictionary, which has been used to phonetically transcribe the training data includes 45 phones. Two additional phones were used to describe the intra- and inter-word silence. This dictionary with the given extensions results in a total number of 47 phones and thus the size of the output layer is 47. The overall number of parameters used in this MLP is approx. 320-thousand.

The MLP has been trained using the quicknet tools [7], which have been developed to train the neural networks for the hybrid approach in [1] on special hardware. This hardware is a vector processor, which was specially designed to reduce the long training time of the neural networks. The advantages of the network training on this hardware are described in [8]. The error function used with these tools is the cross entropy.

The data used to train the neural network and the HMMs were the WSJ0 speaker independent training data, which are 7240 sentences from 84 different speakers, resulting in 5.5 million speech frames using the feature extraction described above. With these training data several systems have been build. The first system is a baseline system, which is a standard hybrid speech recognizer as proposed in [1] using the neural network described above. This system serves two purposes, first of all it is used as a baseline system to verify the success of the proposed approach. The second purpose of this approach is to verify the quality of the neural network described above, which can be compared with the results of other hybrid systems published in the official results of the WSJ evaluation in [9].

The next system is based on the tied posterior approach described above with a single state HMM system as in the baseline system. This system will show the gain in accuracy by using this approach on the same HMM topology. The total number of HMM-weights in this system is 47 in each HMM, so that the total number of parameters in the HMMs is approx. 2200, which is very small compared to the number of parameters in the neural network.

The third system will be a three state HMM system, based on the tied posterior approach to investigate the improvements achieved by the changed topology, which is only possible due to the tied posterior approach. The number of HMM-Parameters now is three times larger than in the system before, thus it is approx. 6600 HMM-weights.

The last system uses state-clustered word-internal triphones which will demonstrate, that context dependent models can be introduced very easily and straightforwardly into this tied posterior approach. The clustering is performed using a tree based clustering with a phonetic question tree. The clustering results in approx. 4-thousand states, which gives 200-thousand individual weights. This amount of weights is already close to the number of parameters in the neural network, so that both parts of this system now have a comparable degree of freedom.

4. RESULTS

The following tests have been executed on the 5000 words speaker independent evaluation tests of the WSJ0 and WSJ1. The WSJ0 test used is the si_et_05 task, which is a speaker independent test with 12 speakers and approx. 40 sentences for each speaker. The WSJ1 test used here is the Hub2 test, which is a similar test with 10 speakers with approx. 20 sentences for each speaker. As language models the supplied bigram model of the WSJ0 has been used.

System	WSJ0 si_et_05	WSJ1 Hub2
Baseline	15.8	20.1
TP 1-state	15.6	19.7
TP 3-states	10.7	14.5
TP 3-states & triphones	9.4	12.5

Table 1: Word error rates on 5k-test sets using a bigram language model

The results in Tab. 1 show, that the error rate of the baseline approach, which is a standard hybrid speech recognizer, can be improved by using the proposed tied posterior (TP)

approach. The use of 3-state models in this approach further reduces the error rate as shown in the next experiment. The last row of Tab. 1 shows the results of the context dependent modeling using state clustered word-internal triphones, which again reduces the error rate due to the better modeling capabilities of the context dependent HMMs.

Comparing the results of the baseline system with the results of other hybrid systems in the publication of the official evaluation [9] shows, that the performance of our baseline hybrid system is lower than the performance of the other systems. This performance shows, that the neural network used in our experiments can be improved. Nevertheless, when using the tied posterior approach with this *low performance* network and a 3-state HMM the results of the presented system are already close to other hybrid systems. With the use of context dependent models the error rate is reduced again, as expected.

5. FUTURE WORK

As shown in the previous section, the performance of our baseline system is not as good as it could be. This reduced performance is due to a different network topology, with less hidden neurons, a different feature extraction and a different dictionary with less phones compared to the neural network in the system described in [1]. With a better neural network not only the performance of the baseline system will improve, the error rates for the tied posterior approach will improve as well because the same network is used in this approach.

So one major aspect to further reduce the error rates will be to improve the quality of the neural probability estimator. Especially the size of the output layer should be increased, e.g. by using another dictionary or by remapping some phones in order to improve the system.

A further investigation of the context dependent models could lead to a better system as well. Here other context dependencies have to be evaluated, e.g. cross word triphones and other clustering procedures and parameters have to be tested.

6. CONCLUSION

In this paper we presented an approach to improve the recognition accuracy of hybrid speech recognizers that are based on a neural network used to estimate the emission probabilities of an HMM. Furthermore we could demonstrate that this approach can be easily used to introduce context dependent models in this hybrid framework. This is achieved by interpreting the probability outputs of the neural network as the codebook of a tied-mixture system. In our experiments on the WSJ database this *tied posterior* approach outperformed the baseline hybrid speech recognition system when using the same neural network in the tied posterior approach and the baseline hybrid system. The result achieved with this system on the WSJ1 test set is one of the best results ever reported using word-internal triphones. We believe, that this approach has several advantages, providing it with the potential of becoming one of the most powerful architectures for future speech recognition systems.

7. ACKNOWLEDGMENTS

The lexicon used for the transcription of the WSJ database was provided by LIMSI. The tools used to train the neural networks were provided by ICSI.

8. REFERENCES

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

[2] A. J. Robinson. An Application of Recurrent Nets to Phone Probability Estimation. *IEEE Transactions on Neural Networks*, 5(2):298–305, March 1994.

[3] J. Fritsch, M. Finke, and A. Waibel. Context–dependent hybrid HME / speech recognition using polyphone clustering decision trees. In *Proc. ICASSP*, pages 1759–1762, 1997.

[4] X. D. Huang and M. A. Jack. Semi-continuous hidden Markov models for speech signals. *Computer Speech and Language*, 3(3):239–251, May 1989.

[5] H. P. Hutter. Comparison of a new hybrid connectionist-SCHMM approach with other hybrid approaches for speech recognition. In *Proc. ICASSP*, pages 3311–3314, 1995.

[6] D. B. Paul and J. M. Baker. The Design for the Wall Street Journal-based CSR Corpus. In *Proc. of the DARPA Speech and Natural Language Workshop*, pages 357–362, Pacific Grove, CA, 1992. Morgan Kaufmann.

[7] D. Johnson. ftp://ftp.icsi.berkeley.edu/pub/real/davidj/quicknet-v0_96.tar.gz.

[8] J. Wawrzynek, K. Asanović, B. Kingsbury, D. Johnson, J. Beck, and N. Morgan. Spert-II: A Vector Microprocessor System. *Computer*, 29(3):79–86, March 1996.

[9] D. S. Pallett, J. G. Fiscus, W. M. Fisher, J. S. Garofolo, B. A. Lund, and M. A. Przybocki. 1993 Benchmark Tests for the ARPA Spoken Language Program. In *Proc. of the Human Language Technology Workshop*, pages 49–74, Plainsboro, New Jersey, March 1994.