Multi-task Learning Strategies for a Recurrent Neural Net in a Hybrid Tied-Posteriors Acoustic Model

Jan Stadermann, Wolfram Koska and Gerhard Rigoll

Institute for Human-Machine Communication
Technische Universität München
Arcisstrasse 21, 80290 Munich, Germany
Phone:+49-89-289-{28319, 28319, 28541},
Email: {stadermann, kos, rigoll}@mmk.ei.tum.de

Abstract

An important goal of an automatic classifier is to learn the best possible generalization from given training material. One possible improvement over a standard learning algorithm is to train several related tasks in parallel. We apply the multi-task learning scheme to a recurrent neural network estimating phoneme posterior probabilities and HMM state posterior probabilities, respectively. A comparison of networks with different additional tasks within a hybrid NN/HMM acoustic model is presented. The evaluation has been performed using the WSJ0 speaker independent test set with a closed vocabulary of 5000 words and shows a significant improvement compared to a standard hybrid acoustic model if gender classification is used as additional task.

1. Introduction

Hybrid NN/HMM acoustic models have been introduced some time ago [1] with several advantages over conventional HMMs with Gaussian mixtures: The NN is trained discriminatively and context information is easily incorporated. In [2] two recurrent neural networks (RNN) replace the multi-layer perceptron (MLP) as phoneme probability estimator. Recently tied-posteriors acoustic models are presented in [3] that extent the hybrid modeling paradigm to arbitrary HMM topologies such as context dependent models. The behavior of a MLP or a single RNN within a tied-posteriors model has been investigated in [4], where the RNN performed almost as well as a MLP, but with only 60% of the number of parameters. This work has been inspired by [5] where a RNN has been trained to estimate phoneme posterior probabilities and an enhanced feature vector simultaneously from noisy speech on a connected digit task. Here, we investigate a tied-posteriors acoustic model with a RNN using multi-task learning (MTL) with different higher level problems as additional tasks. Experiments have been carried out on the WSJ0 database. The next two sections briefly describes the neural network's topology and the benefits of MTL, section 4 gives a short overview about our tied-posteriors acoustic model, section 5 presents three sets of classes that have been used as additional tasks, finally section 6 shows the evaluation results and section 7 concludes this paper.

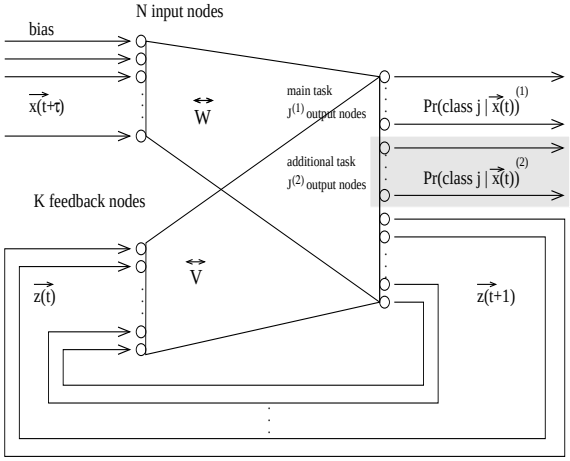

Figure 1: RNN architecture with two output blocks

2. Recurrent neural networks

The network's architecture that is a modification of the one introduced in [6] is illustrated in figure 1. It is a partial recurrent network with separate output and feedback nodes, but with full connection between input, output and feedback nodes. The input layer consists of one normalized feature vector with 12 standard MFCC features, log. energy and its first and second derivatives (39 components). The RNN possesses the general advantage that context information of past time steps is implicitly stored in the feedback nodes. By delaying the network decision by τ frames, we can incorporate information about future time frames, as well. So, applying input vector $\vec{x}(t+\tau)$ we obtain the output $\vec{y}(t)$ and the feedback vector $\vec{z}(t+1)$.

From: Interspeech 2005, Lisboa, ISCA

The main task consists of 47 or 139 output nodes with a softmax non-linearity, estimating phoneme probabilities or HMM state probabilities [7], respectively and in both cases 400 feedback nodes with a sigmoid non-linearity are used.

3. Multi-task learning

The basic principle of MTL in the framework of machine learning with neural networks is to use additional output nodes for additional problems. In our case, each output node estimates a conditioned class posterior probability given an input feature vector. The output vector $\vec{y}$ is extended by a block of output nodes for each additional task (see figure 1). For each block m the equation

$$\sum_{j=1}^{J^{(m)}} \Pr(j|\vec{x})^{(m)} = 1 \text{ for all } m \qquad (1)$$

holds by applying a softmax non-linearity separately on each block ($\Pr(j|\vec{x})^{(m)}$ denotes the class posterior probability of output block m given a feature vector $\vec{x}$). Training is performed equally for all tasks using back-propagation through time (BPTT) and the RPROP update strategy [8] with each block of output nodes having a separate set of target vectors. The function to be optimized is the cross entropy between network outputs and targets. In [9] properties of a MLP trained with MTL are outlined. Since standard back propagation (used for MLP training) and BPTT are very similar, the advantages mentioned in [9] are also valid for RNNs with one feedback layer (corresponding to a MLP's hidden layer). In brief, the important advantages of MTL are

- Statistical data amplification
  (using the same input data for multiple problems, using a limited amount of training data efficiently)
- Eavesdropping
  (a difficult classification can improve if another classification with the same data is successful)
- Representation bias
  (a better optimum of the training function might be found if two tasks agree on a similar extremum)

Apart from these advantages increasing the RNN's number supervised output nodes generally improves its performance: Section 6 illustrates the gain in terms of performance of RNNs trained on HMM states (139 supervised output nodes) instead of phonemes (47 targets).

4. Tied-posteriors acoustic models

The tied-posteriors approach presented in [3] integrates the neural net in the HMM framework. It is based on a

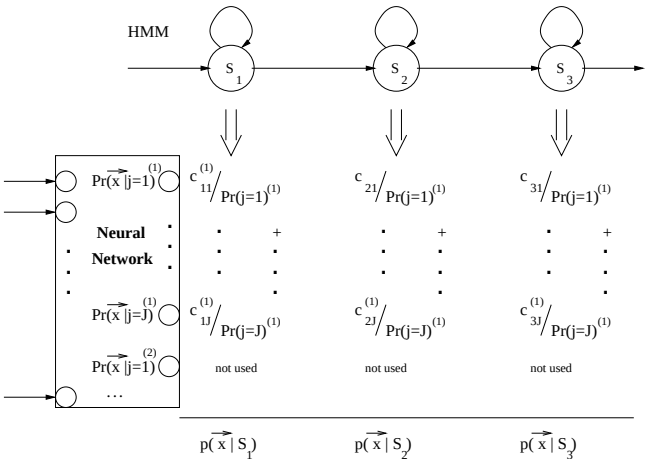

Figure 2: Tied-posteriors model with 3 HMM states

(continuous) tied-mixture system where the output probabilities of each state are denoted as

$$p(\vec{x}|S_i) = \sum_{j=1}^{J} c_{ij} \cdot p(\vec{x}|j)\,, \qquad (2)$$

where S_i is the HMM state and c_{ij} are the mixture coefficients. The idea is now to replace the probability density $p(\vec{x}|j)$ by the posterior probability $Pr(j|\vec{x})$ by using

$$p(\vec{x}|j) = \frac{Pr(j|\vec{x})^{(1)} p(\vec{x})}{Pr(j)^{(1)}} \qquad (3)$$

Since $p(\vec{x})$ is independent of the HMM state S_i it can be omitted and (2) becomes

$$p(\vec{x}|S_i) \propto \sum_{j=1}^{J^{(1)}} c_{ij}^{(1)} \cdot \frac{Pr(j|\vec{x})^{(1)}}{Pr(j)^{(1)}} \qquad (4)$$

Figure 2 shows the principle set-up of the tied-posteriors system using equation 4 for a 3 state monophone HMM. The posterior probability $Pr(j|\vec{x})^{(1)}$ is estimated by the RNN's first output block, the *a-priori* probability $Pr(j)^{(1)}$ can be estimated by counting the phoneme/state occurrences in the training set.
The number of output nodes $J^{(1)}$ of the first block is either equal to the number of phonemes or to the number of HMM states. Summarizing this model, it is important to note that only the output probabilities of the RNN's first output block are used for the HMM density computation. If all output blocks should be used, the HMM density computation could be extended to

$$p(\vec{x}|S_i) \propto \sum_{m=1}^{M} \sum_{j=1}^{J^{(m)}} c_{ij}^{(m)} \cdot \frac{Pr(j|\vec{x})^{(m)}}{Pr(j)^{(m)}} \qquad (5)$$

In section 6 the effect of using all output blocks for the HMM density computation as described in equation 5 is discussed.

5. Additional tasks

As outlined in section 4 the RNN's main task is HMM state classification or phoneme classification, respectively. For each set-up we have appended one of the following classification problems as additional task:

- Gender classification
  Since the gender classification has to be performed on each frame it consists of three classes (male, female, silence) thus requiring a second block with 3 output neurons.

- Broad phoneme classification
  The 45 phonemes are grouped into 9 phonetic classes (plosive, fricative, nasal, approximant, a-vocal, e-vocal, i-vocal, o-vocal, u-vocal) plus silence and short pause requiring 11 additional output nodes

- Grapheme classification
  A separate acoustic model has been trained using graphemes instead of phonemes [10]. The grapheme targets have been obtained by forced alignment of the training data. There are 28 additional classes (a,b,...,z,silence, short pause).

6. Experiments

Experiments are carried out on the speaker-independent test set *si 05* of the WSJ0 database with a closed vocabulary of 5000 words and applying a bigram and trigram language model. All RNNs have been trained on 7240 sentences of the *si 84* set (10% of the sentences are taken out for cross validation). The HMM state based segmentation and the phoneme level segmentation are obtained by forced Viterbi alignment. As soon as the RNN training is finished, the Baum-Welch algorithm is used to compute the mixture weights c_{ij}. For each phoneme (45 monophones plus *silence*) a 3-state HMM is created, for *short pause* a one-state HMM is used.
Table 1 shows the frame error rates

$$\mathrm{FER} = \frac{\#\text{correct classified frames}}{\#\text{frames}}$$

of the different RNNs using the cross validation data that is used as a stopping criterion during the training process. *Correct classified* is defined as agreement between the reference class index and the class index of the highest probability value. *FER 1* denotes the frame error rate on the main task (phoneme classification or HMM state classification) and *FER 2* denotes the frame error rate on the second task (either gender, broad phoneme or grapheme classification).

The FER of all MTL networks is worse than the one from a single-task trained RNN if phoneme probabilities are used as primary task. If HMM state probabilities are to be calculated the baseline FER is higher because each HMM state is to be assigned correctly (that are 139 classes in total). Here, the RNNs trained on HMM states and gender or graphemes, respectively perform slightly better than the baseline RNN. Table 2 presents the word error rates (WER) of the RNNs mentioned in table1 integrated in a hybrid tied-posteriors speech recognition system. It can be stated that the gender classification improves the phoneme based system a little bit and the HMM state based system significantly compared to the baseline single-task system. The added grapheme classification produces a slight degradation and the broad phoneme classification adds significantly more errors. In table 3 the same systems as in table 2 are evaluated, but this time using a trigram language model. Again the added gender classification improves the performance whereas the other tasks increase the number of errors.

Neural Network Tasks			
Main Task 1	**Add. Task 2**	**FER 1**	**FER 2**
Phonemes	-	25.29%	-
Phonemes	broad phonemes	25.91%	19.05%
Phonemes	gender	26.63%	6.4%
HMM states	-	35.04%	-
HMM states	broad phonemes	36.21%	18.67%
HMM states	gender	34.71%	4.8%
HMM states	graphemes	34.56%	33.59%

Table 1: Frame error rates (FER) of HMM state-based and phoneme-based RNNs with different additional tasks

Neural Network Tasks		
Main Task	**Add. Task**	**WER**
Phonemes	-	13.19%
Phonemes	broad phonemes	13.56%
Phonemes	gender	**12.96%**
HMM states	-	11.13%
HMM states	broad phonemes	12.54%
HMM states	gender	**10.20%**
HMM states	graphemes	11.30%

Table 2: Word error rates (WER) of HMM state-based and phoneme-based RNNs with different additional tasks, bigram language model

Finally table 4 compares the three systems with additional tasks to the single task baseline if all tasks are used for the HMM density calculation as described in equation 5. Only the system using broad phonemes improves a little bit the other systems get slightly worse. Since the decoding is slowed down as well using all tasks for the

Neural Network Tasks		
Main Task	**Add. Task**	**WER**
HMM states	-	8.52%
HMM states	broad phonemes	9.32%
HMM states	gender	**7.98%**
HMM states	graphemes	8.59%

Table 3: Word error rates (WER) of HMM state-based and phoneme-based RNNs with different additional tasks, trigram language model

Neural Network Tasks		
Main Task	**Add. Task**	**WER**
HMM states	-	11.13%
HMM states	broad phonemes	12.39%
HMM states	gender	**10.27%**
HMM states	graphemes	11.92%

Table 4: Word error rates (WER) of HMM state-based and phoneme-based RNNs with different additional tasks, bigram language model, all task are used for the HMM density calculation

density calculation does not result in any advantages.

How to explain these results? First, the additional task seems to have to be independent from the main task. This is true for the gender and (to a certain extent) grapheme classification but not for the broad phoneme classes that are created simply by grouping phonemes into larger units. Gender and grapheme symbols are related to the speech signal but are based on a different source than the phoneme segmentation. Second, it seems to be necessary that the additional task is "easier" to classify than the main task (*eavesdropping*, see section 3) which is true for the gender classification and broad phoneme classification.

7. Conclusion

The multi-task learning paradigm has been implemented for tied-posteriors NN/HMM acoustic models with recurrent neural networks as probability estimators. Additional tasks such as gender classification, broad phoneme classification or grapheme classification have been trained in parallel with phoneme and HMM state classification, respectively. The ASR system with the different models has been evaluated on the speaker independent WSJ0 test. A significant gain in terms of word errors has been achieved by training the RNN to classify HMM states and gender. Using the second task's output nodes for the HMM density calculation does not seem to be useful. A detailed look at the additional tasks reveals that they should be easier to classify and have to be independent from the main problem. Future research includes to find a better way to incorporate the result from the additional nodes in the HMM density calculation and to find better additional tasks with the necessary properties.

8. References

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

[2] M. Hochberg, G. Cook, S. Renals, A. Robinson, and R. Schechtman, "ABBOT Hybrid Connectionist-HMM Large-Vocabulary Recognition System," in *Spoken Language Systems Technology Workshop*, 1995, pp. 170–176.

[3] J¨org Rottland and Gerhard Rigoll, "Tied Posteriors: An Approach for Effective Introduction of Context Dependency in Hybrid NN/HMM LVCSR," in *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, June 2000.

[4] Jan Stadermann and Gerhard Rigoll, "Comparing NN Paradigms in Hybrid NN/HMM Speech Recognition using Tied Posteriors," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, St. Thomas, U.S. Virgin Islands, Nov. 2003.

[5] Shahla Parveen and Phil Green, "Multitask Learning in Connectionist Robust ASR using Recurrent Neural Networks," in *European Conference on Speech Communication and Technology*, Geneva, Switzerland, Sept. 2003, pp. 1813–1816.

[6] A. J. Robinson, "An Application of Recurrent Nets to Phone Probability Estimation," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 298–305, Mar. 1994.

[7] S. Santini and A. Del Bimbo, "Recurrent neural networks can be trained to be maximum A posteriori probability classifiers," *Neural Networks*, vol. 8, no. 1, pp. 25–29, 1995.

[8] Christian Igel and Michael H¨usken, "Improving the Rprop Learning Algorithm," in *Proceedings of the Second International Symposium on Neural Computation NC'2000*. 2000, pp. 115–121, ICSC Academic Press.

[9] Rich Caruana, "Multitask Learning," *Machine Learning*, , no. 28, pp. 41–75, 1997.

[10] Mirjam Killer, Sebastian St¨uker, and Tanja Schultz, "Grapheme Based Speech Recognition," in *European Conference on Speech Communication and Technology*, Geneva, Switzerland, Sept. 2003, pp. 3141–3144.