A NN/HMM HYBRID FOR CONTINUOUS SPEECH RECOGNITION WITH A DISCRIMINANT NONLINEAR FEATURE EXTRACTION

Gerhard Rigoll, Daniel Willett

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

ABSTRACT

This paper deals with a hybrid NN/HMM architecture for continuous speech recognition. We present a novel approach to set up a neural linear or nonlinear feature transformation that is used as a preprocessor on top of the HMM system's RBF-network to produce discriminative feature vectors that are well suited for being modeled by mixtures of Gaussian distributions. In order to omit the computational cost of discriminative training of a context-dependent system, we propose to train a discriminant neural feature transformation on a system of low complexity and reuse this transformation in the context-dependent system to output improved feature vectors. The resulting hybrid system is an extension of a state-of-the-art continuous HMM system, and in fact, it is the first hybrid system that really is capable of outperforming these standard systems with respect to the recognition accuracy, without the need for discriminative training of the entire system. In experiments carried out on the Resource Management 1000-word continuous speech recognition task we achieved a relative error reduction of about 10% with a recognition system that, even before, was among the best ever observed on this task.

1. INTRODUCTION

Standard state-of-the-art speech recognition systems utilize Hidden Markov Models (HMMs) to model the acoustic behavior of basic speech units like phones or words. Most commonly the probabilistic distribution functions (pdfs) are modeled as mixtures of Gaussian distributions. These mixture distributions can be regarded as output nodes of a Radial-Basis-Function (RBF) network that is embedded in the HMM system [4]. Contrary to neural training procedures the parameters of the HMM system, including the RBF network, are usually estimated to maximize the likelihood of the training observations. In order to combine the time-warping abilities of HMMs and the more discriminative power of neural networks, several hybrid approaches arose during the past five years, that combine HMM systems and neural networks. The best known approach is the one proposed in [2] and [3]. It replaces the HMMs' RBF-net with a Multi-Layer-Perceptron (MLP) which is trained to output each HMM state's posterior probability. Another hybrid approach was presented by our group in [6, 7]. By combining a discrete HMM speech recognition system and a neural quantizer and maximizing the mutual information between the VQ-labels and the assigned phoneme-classes, this approach outperforms standard discrete recognition systems. We showed that this approach is capable of building up very accurate systems with an extremely fast likelihood computation, that only consists of a quantization and a table lookup. Unfortunately though, the mentioned hybrid systems yet failed to substantially outperform very good continuous systems with respect to the recognition accuracy. And in addition to that, it is well known, that weak discrimination in likelihood based approaches can be improved by discriminative training objectives, like MCE or MMI. This discriminative training, however, is computationally extremely expensive. Especially when a context-dependent acoustic modeling is being used with thousands of HMMs and pdfs.

2. HYBRID CONTINUOUS HMM/MLP APPROACH

Therefore we followed a different approach, namely the extension of a state-of-the-art continuous system, that achieves an extremely good recognition performance, with a neural net that is trained with MMI-methods related to those in [9]. This architecture was reported on in [10]. The parameters of the HMM system, i. e. the RBF-part of this architecture, are trained efficiently with the EM-algorithm maximizing the likelihood of the acoustic observations. In a discriminative training procedure, only the parameters of the additional neural component are adjusted. This way, the additional network component transforms the speech features in order to increase discrimination at the output nodes of the RBF-net. This discriminative training is performed on a system of low complexity (monophones) to keep the computational costs reasonably small.

2.1. ARCHITECTURE

The basic architecture of the hybrid system is illustrated in Figure 1. The neural net functions as a feature transformation that takes several additional adjacent feature vectors into account to produce an improved discriminant feature vector that is fed into the HMM system.
This architecture allows (at least) three ways of interpretation; 1. as a hybrid system that combines neural nets and continuous HMMs, 2. as an LDA-like transformation that incorporates the HMM parameters into the calculation of the transformation matrix and 3. as feature extraction method, that allows the extraction of features according to the underlying HMM system with an incorporation of adjacent frames. The considered types of neural networks are linear transformations, MLPs and recurrent MLPs. A detailed description of the possible topologies is given in Section 3.
With this architecture, additional past and future feature vectors can be taken into account in the probability estimation process without increasing the dimensionality of the Gaussian mixture components. Instead of increasing the HMM system's number of parameters the neural net is utilized to produce more discriminant feature vectors with respect to the trained HMM system. Of course, adding some kind of neural net increases the number of parameters too, but the increase is much more moderate than it would be when increasing each Gaussian's dimensionality. And, as the major reason for training a neural transformation instead of the HMM system's parameters, we want to reuse the trained net from a low complexity system for providing features for a system of higher complexity.

2.2. TRAINING OBJECTIVE

In our experiments the neural net is trained to improve the frame-based discrimination among the pdfs, i. e. to maximize the mutual information between the acoustical features and the system's output at the output nodes of the RBF-net. Certainly, other approaches for discriminative training could be applied as well, like

those presented in [9, 5, 8].
The MMI criterion is usually formulated in the following way:

$$\hat{\lambda}_{MMI} = \underset{\lambda}{\operatorname{argmax}}\, I_\lambda(X, W)$$

$$= \underset{\lambda}{\operatorname{argmax}}(H_\lambda(X) - H_\lambda(X|W)) = \underset{\lambda}{\operatorname{argmax}} \frac{p_\lambda(X|W)}{p_\lambda(X)} \quad (1)$$

This means that following the MMI criterion the system's free parameters λ have to be estimated to maximize the quotient of the observation's likelihood $p_\lambda(X|W)$ for the known transcription W and its overall likelihood $p_\lambda(X)$. With $X = (x(1), x(2), ...x(T))$ denoting the training observations and $W = (w(1), w(2), ...w(T))$ denoting the HMM states - assigned to the observation vectors in a Viterbi-alignment - the frame-based MMI criterion becomes

$$\hat{\lambda}_{MMI} \approx \underset{\lambda}{\operatorname{argmax}} \sum_{i=1}^{T} I_\lambda(x(i), w(i))$$

$$= \underset{\lambda}{\operatorname{argmax}} \prod_{i=1}^{T} \frac{p_\lambda(x(i)|w(i))}{p_\lambda(x(i))}$$

$$\approx \underset{\lambda}{\operatorname{argmax}} \prod_{i=1}^{T} \frac{p_\lambda(x(i)|w(i))}{\sum_{k=1}^{S} p_\lambda(x(i)|w_k)P(w_k)} \quad (2)$$

where S is the total number of HMM states, $(w_1, ...w_S)$ denotes the HMM states and $P(w_k)$ denotes each states' prior-probability that is estimated on the alignment of the training data or by an analysis of the language model.
Eqn. (2) can be used to re-estimate the Gaussians of a continuous HMM system directly. In [9] we reported the slight improvements in recognition accuracy that we achieved with this parameter estimation. However, it turned out, that only the incorporation of additional past and future features in the probability calculation pipeline can provide more discriminative emission probabilities and a major advance in recognition accuracy. The proposed neural net offers an ideal way to incorporate additional feature frames without increasing each Gaussian's dimensionality.

2.3. TRAINING PROCEDURE

For the parameter estimation according to Eqn. (2) we chose a gradient descent procedure. At first, for matter of simplicity, we will consider a linear network that takes P past feature vectors and F future feature vectors as additional input. With N denoting the number of extracted features per frame, the linear net denoted as a $(P + F + 1) \times N$ matrix NET, each component $x'(t)[c]$ ($c = 1...N$) of the network output $x'(t)$ computes to

$$x'(t)[c] = \sum_{i=0}^{P+F} \sum_{j=1}^{N} x(t-P+i)[j] \cdot NET[i*N+j][c] \; \forall c = 1...N \quad (3)$$

so that the derivative with respect to a component of NET easily computes to

$$\frac{\partial x'(t)[c]}{\partial NET[i*N+j][\hat{c}]} = \delta_{c,\hat{c}} x(t-P+i)[j] \quad (4)$$

In a continuous HMM system with diagonal covariance matrices the pdf of each HMM state w is modeled by a mixture of Gaussian components like

$$p_\lambda(x|w) = \sum_{j=1}^{C_w} d_{wj} \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\sigma}_\mathbf{j}|}} e^{-\frac{1}{2}\sum_{l=1}^{N} \frac{(m_j[l]-x[l])^2}{\sigma_j[l]}} \quad (5)$$

A pdf's derivative with respect to a component $x'[c]$ of the net's output becomes

$$\frac{\partial p_\lambda(x'|w)}{\partial x'[c]} =$$

$$\sum_{j=1}^{C_w} d_{wj} \frac{(x[c] - m_j[c])}{\sigma_j[c]} \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\sigma}_\mathbf{j}|}} e^{-\frac{1}{2}\sum_{l=1}^{N} \frac{(m_j[l]-x'[l])^2}{\sigma_j[l]}} \quad (6)$$

With $x(t)$ in Eqn. (2) now replaced by the net output $x'(t)$ the partial derivative of Eqn. (2) with respect to a probabilistic distribution function $p(x'(i)|w_k)$ computes to

$$\frac{\partial I_\lambda(x'(i), w(i))}{\partial p_\lambda(x'(i)|w_k)} = \frac{\delta_{w(i),w_k}}{p_\lambda(x(i)|w_k)} - \frac{P(w_k)}{\sum_{l=1}^{S} p_\lambda(x(i)|w_l)P(w_l)} \quad (7)$$

Thus, using the chain rule the derivative of the net's parameters with respect to the frame-based MMI criterion can be computed as displayed in Eqn. (8)

$$\frac{\partial I_\lambda(X, W)}{\partial NET[l][c]} = \quad (8)$$

$$\sum_{i=1}^{T} \left(\sum_{k=1}^{S} \left(\frac{\partial I_\lambda(x(i)|w(i)))}{\partial p_\lambda(x'(i)|w_k)} \frac{\partial p_\lambda(x'(i)|w_k)}{\partial x'(i)[c]} \frac{\partial x'(i)[c]}{\partial NET[l][c]} \right) \right)$$

2.4. ADVANTAGES OF THE PROPOSED APPROACH

When using a linear network, the proposed approach strongly resembles the well known Linear Discriminant Analysis (LDA) [1] in architecture and training objective. The main difference is the way the transformation is set up. In the proposed approach the transformation is computed by taking directly the HMM parameters into account whereas the LDA only tries to separate the features according to some class assignment. With the incorporation of a trained continuous HMM system the net's parameters are estimated to produce feature vectors that not only have a good separability in general, but also have a distribution that can be modeled with mixtures of Gaussians very well. Our experiments given at the end of this paper prove this advantage. Furthermore, contrary to LDA, that produces feature vectors that don't have much in common with the original vectors, the proposed approach only slightly modifies the input vectors. Thus, a well trained continuous system can be extended by the MMI-net approach, in order to improve its recognition performance without the need for completely rebuilding it. In addition to that, the approach offers a fairly easy extension to nonlinear networks (MLP) and recurrent networks (recurrent MLP). This will be outlined in the following section. And, maybe as the major advantage, the approach allows keeping up the division of the input features into streams of features that are strongly uncorrelated and which are modeled with separate pdfs. The case of multiple streams is discussed in detail in Section 4. Besides, the MMI approach offers the possibility of a unified training of the HMM system and the feature extraction network or an iterative procedure of training each part alternately.

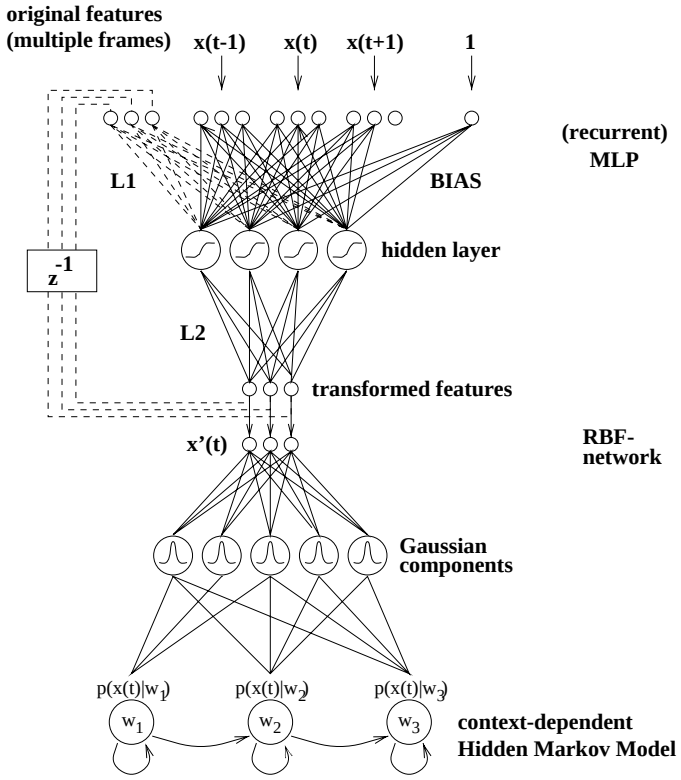

Figure 1. Hybrid system with a nonlinear (recurrent) feature transformation

3. NETWORK TOPOLOGIES

Section 2 explained how to train a linear transformation with respect to the frame-based MMI criterion. However, to exploit all the advantages of the proposed hybrid approach the network should be able to perform a nonlinear mapping, in order to produce features whose distribution is (closer to) a mixture of Gaussians although the original distribution is not.

3.1. MLP

When using a fully connected MLP as displayed in Figure 1 with one hidden layer of H nodes, that perform the nonlinear function f, the activation of one of the output nodes $x'(t)[c]$ becomes

$$x'(t)[c] = \sum_{h=1}^{H} L2[h][c] \cdot \tag{9}$$

$$f\left(BIAS_h + \sum_{i=0}^{P+F}\sum_{j=1}^{N} x(t-P+i)[j] \cdot L1[i*N+j][h]\right)$$

which is easily differentiable with respect to the nonlinear network's parameters. In our experiments we chose f to be defined as the hyperbolic tangents $f(x) := tanh(x) = (2(1+e^{-x})^{-1} - 1)$ so that the partial derivative with respect to i. e. a weight $L1[\hat{i} \cdot N + \hat{j}][h]$ of the first layer computes to

$$\frac{\partial x'(t)[c]}{\partial L1[\hat{i} \cdot N + \hat{j}][h]} = x(t-P+\hat{i})[\hat{j}] \cdot L2[h][c]$$

$$\cdot cosh\left(BIAS_h + \sum_{i=0}^{P+F}\sum_{j=1}^{N} x(t-P+i)[j] \cdot L1[i*N+j][h]\right)^{-2} \tag{10}$$

and the gradient can be assembled according to Eqn. (8).

3.2. RECURRENT MLP

With the incorporation of several additional past feature vectors as explained in Section 2, more discriminant feature vectors can be generated. However, this method is not capable of modeling longer term relations, as it can be achieved by extending the network with some recurrent connections. For the sake of simplicity, in our experiments we simply extended the MLP as indicated with the dashed lines in Figure 1 by propagating the output $x(t)$ back to the input of the network (with a delay of one discrete time step). This type of recurrent neural net is often referred to as a 'Jordan'-network. Certainly, the extension of the network with additional hidden nodes in order to model the feedback more independently would be possible as well.

4. MULTI STREAM SYSTEMS

In HMM-based recognition systems the extracted features are often divided into streams that are modeled independently. This is useful the less correlated the divided features are. In this case the overall likelihood of an observation computes to

$$p_\lambda(x|w) = \prod_{s=1}^{M} p_{s\lambda}(x|w)^{w_s} \tag{11}$$

where each of the stream pdfs $p_{s\lambda}(x|w)$ only uses a subset of the features in x. The stream weights w_s are usually set to unity.
A multi stream system can be improved by a neural extraction for each stream and an independent training of these neural networks. However, it has to be considered that the subdivided features usually are not totally independent and by considering multiple input frames as illustrated in Figure 1 this dependence often increases. It is a common practice, for instance, to model the features' first and second order delta coefficients in independent streams. So, for sure the streams lose independence when considering multiple frames, as these coefficients are calculated using the additional frames. Nevertheless, we found it to give best results to maintain this subdivision into streams, but to consider the stronger correlation by training each stream's net dependent on the other nets' outputs. A training criterion follows straight from Eqn. (11) inserted in Eqn. (2).

$$\hat{\lambda}_{MMI} = \arg\max_\lambda \prod_{i=1}^{T} \frac{p_\lambda(x(i)|w(i))}{p_\lambda(x(i))}$$

$$= \arg\max_\lambda \prod_{i=1}^{T}\prod_{s=1}^{M} \left(\frac{p_{s\lambda}(x(i)|w(i))}{p_{s\lambda}(x(i))}\right)^{w_s} \tag{12}$$

The derivative of this equation with respect to the pdf $p_{\hat{s}\lambda}(x|w)$ of a specific stream $\hat{s}$ depends on the other streams' pdfs. With the w_s set to unity it is

$$\frac{\partial I_\lambda(x'(i), w(i))}{\partial p_{\hat{s}\lambda}(x'(i)|w_k)} = \left(\prod_{s\neq\hat{s}} \frac{p_{s\lambda}(x(i)|w(i))}{p_{s\lambda}(x(i))}\right)$$

$$\left(\frac{\delta_{w(i),w_k}}{p_{\hat{s}\lambda}(x(i)|w_k)} - \frac{P(w_k)}{\sum_{l=1}^{S} p_{\hat{s}\lambda}(x(i)|w_l)P(w_l)}\right) \tag{13}$$

Neglecting the correlation among the streams the training of each stream's net can be done independently. However, the more the incorporation of additional features increases the streams' correlation, the more important it gets to train the nets in a unified training procedure according to Eqn. (13).

	base-line	LDA	linear	MLP	Jordan Net
monoph 1 strm	24%	21%	21%	21%	21%
monoph 4 strms	11.8%	11.0%	10.9%	10.8%	10,9%
triph 4 strms	5.2%	5.3%	4.8%	4.7%	4.7%

Table 1. Word error rates achieved in the experiments on the RM database

5. EXPERIMENTS AND RESULTS

We applied the proposed approach to improve a context-independent (monophones) and a context-dependent (triphones) continuous speech recognition system for the 1000-word Resource Management (RM) task. The systems used linear HMMs of three emitting states each. The tying of Gaussian mixture components was performed with an adaptive procedure according to [11]. The HMM states of the word-internal triphone system were clustered in a tree-based phonetic clustering procedure. Decoding was performed with a Viterbi-decoder and the standard wordpair-grammar of perplexity 60. The gradient descent training was performed with the RPROP algorithm on the monophone multi-stream system. For training the weights of the recurrent connections we chose real-time recurrent learning. The average error rates were computed using the test-sets Feb89, Oct89, Feb91 and Sep92.

The table above shows the recognition results with single stream systems in its first row. These systems simply use a 12-value Cepstrum feature vector without the incorporation of delta coefficients. The first column displays the results on the baseline continuous systems. Columns 3-5 show the results for these systems extended into hybrid systems with the same HMM parameters but additional neural networks on top of the HMMs. For comparison, Column 2 shows the results if the baseline HMM system is retrained with acoustic features derived in a LDA with the same input and output dimensions. The systems with an input transformation use one additional past and one additional future feature vector as input. The proposed approach achieves the same performance as the LDA with the same input and output dimensions, but it is not capable of outperforming it.

The second row lists the recognition results with four stream monophone systems that use the first and second order delta coefficients in additional streams plus log energy and this values' delta coefficients in a forth stream. The MLP system trained according to Eqn. (11) with 36 hidden nodes slightly outperforms the other approaches. The incorporation of recurrent network connections does not improve the system's performance.

The third row lists the recognition results of four stream systems with a context-dependent acoustic modeling (triphones) that makes use of the neural networks and the LDA taken from the monophone four stream system of row two. The Estimation of the HMM parameters of these systems was simply performed to maximize the observation likelihood using the EM-algorithm. On the one hand, this was done to avoid the computational complexity that MMI training objectives cause on context-dependent systems. On the other hand, this demonstrates that the feature vectors produced by the trained networks have a good discrimination for continuous systems in general. Again, the MLP system outperforms the other approaches and achieves a very remarkable word error rate. With a recognition rate of 4.7% as average of all four test-sets the system is one of the best ever reported, although it does not make use of cross-word acoustic modeling and is not trained discriminatively.

6. CONCLUSION

The paper has presented a novel approach to discriminant feature extraction. A MLP network has successfully been used to compute a feature transformation that outputs extremely suitable features for continuous HMM systems. The experimental results have proven that the proposed approach is an appropriate method for including several feature frames in the probability estimation process without increasing the dimensionality of the Gaussian mixture components in the HMM system. Furthermore did the results on the triphone speech recognition system prove that the approach provides discriminant features, not only for the system that the mapping is computed on, but for HMM systems with a continuous modeling in general. The application of recurrent networks did not improve the recognition accuracy. The longer range relations seem to be very weak and they seem to be covered well by using the neighboring feature vectors and first and second order delta coefficients. The proposed unified training procedure for multiple nets in multi-stream systems allows keeping up the subdivision of features of weak correlations and gave us best profits in recognition accuracy.

ACKNOWLEDGMENTS

This work was partly sponsored by the DFG (German Research Foundation) under contract number Ri 658/6-1.

REFERENCES

[1] X. Aubert, R. Haeb-Umbach, H. Ney, "Continuous mixture densities and linear discriminant analysis for improved context-dependent acoustic models", Proc. ICASSP, 1993, pp. 648–651.

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

[3] M. M. Hochberg, G. D. Cook, S. J. Renals, A. J. Robinson, A. S. Schechtman, "The 1994 ABBOT Hybrid Connectionist-HMM Large-Vocabulary Recognition System", Proc. ARPA Spoken Language Systems Technology Workshop, 1995.

[4] H. Ney, "Speech Recognition in a Neural Network Framework: Discriminative Training of Gaussian Models and Mixture Densities as Radial Basis Functions", Proc. ICASSP, 1991, pp. 573–576.

[5] Y. Normandin, R. Lacouture, R. Cardin: "MMIE Training for Large Vocabulary Continuous Speech Recognition" Proc. ICSLP, 1994, pp. 1367–1370.

[6] G. Rigoll, "Maximum Mutual Information Neural Networks for Hybrid Connectionist-HMM Speech Recognition", IEEE-Trans. Speech Audio Processing, Vol. 2, No. 1, Jan. 1994, pp. 175–184.

[7] G. Rigoll, C. Neukirchen, "A new approach to hybrid HMM/ANN speech recognition using mutual information neural networks", Advances in Neural Information Processing Systems (NIPS-96), Denver, 1996, pp. 772–778.

[8] V. Valtchev, J.J. Odell, P.C. Woodland, S.J. Young "Lattice-Based Discriminative Training for Large Vocabulary Speech Recognition" Proc. ICASSP, 1996, pp. 605–608.

[9] D. Willett, C. Neukirchen, R. Rottland, "Dictionary-Based Discriminative HMM Parameter Estimation for Continuous Speech Recognition Systems", Proc. ICASSP, 1997, pp. 1515–1518.

[10] D. Willett, G. Rigoll, "Hybrid HMM/NN Speech Recognition with a discriminant neural feature extraction" Advances in Neural Information Processing Systems (NIPS-97), Denver, 1997.

[11] D. Willett, G. Rigoll, "A New Approach to Generalized Mixture Tying for Continuous HMM-Based Speech Recognition", Proc. EUROSPEECH, 1997, pp. 1175–1178.