ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 243-248

Speech Recognition With A New Hybrid Architecture Combining Neural Networks And Continuous HMM

Daniel Willett, Gerhard Rigoll

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

Abstract. In this paper, we focus on a novel NN/HMM architecture for continuous speech recognition. The architecture incorporates a neural feature extraction to gain more discriminative feature vectors for the underlying HMM system. The feature extraction can be chosen either linear or non-linear and can incorporate recurrent connections. With this hybrid system, that is an extension of a state-of-the-art continuous HMM system, we managed to significantly outperform these standard systems. Experimental results show a relative error reduction of about 10% on a remarkably good recognition system based on continuous HMMs for the Resource Management 1000-word continuous speech recognition task.

1. INTRODUCTION

State-of-the-art speech recognition systems utilise Hidden Markov Models (HMMs) to model the acoustic behaviour of basic speech units like phones or words. Most commonly, the probabilistic distribution functions (pdfs) are modelled as mixtures of Gaussian distributions. Contrary to neural training procedures and contrary to the natural objective of discriminating the correct from the incorrect transcriptions, the parameters of the HMM system, including the Gaussian mixture distributions, are usually estimated to maximise 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. Although the ordinary Gaussian modelling can be regarded as a RBF-network, too [1], the term "hybrid" is most commonly used for any neural extension of this framework.

The best known hybrid approach is the one described in [2]. It replaces the RBF-net with a Multi-Layer-Perceptron (MLP). The approach is based on the interpretation of the MLP outputs as posterior HMM-state probabilities according to [3]. Recently, other types of networks like Recurrent Networks [2] and Hierarchical Mixtures of Experts [4] have been applied successfully in this kind of hybrid framework.

In [5] Bengio et al. demonstrated how to carry out a global optimisation of a NN/HMM combination, in which the NN is utilised as an additional component on top of a HMM system. It turned out that Maximum Likelihood estimation is inadequate for this type of hybrid, but that a MMI estimation of such a network can provide remarkable improvements in recognition performance.

Another approach that should be mentioned in the context of hybrid systems, is known as Linear Discriminant Analysis (LDA) [6]. LDA was mainly introduced to tackle the Curse Of Dimensionality that occurs when a large number of features are extracted from the speech signal and the training data is limited. With a linear transformation the LDA reduces the dimensionality of the HMM input while it minimises the information loss of the reduced feature vector with respect to a phone or HMM-state alignment.

Recently [7], our group presented a novel hybrid speech recognition approach that combines a discrete HMM speech recognition system and a neural quantiser. By maximising 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 quantisation and a table lookup. This resulted in a hybrid system with recognition performance equivalent to the best continuous systems, but with a much faster decoding. Nevertheless, it has turned out that this hybrid approach is not really capable of substantially outperforming very good continuous systems with respect to the recognition accuracy.

This observation is similar to experiences with the Posteriori-Network approach that was mentioned above. For the decoding procedure, this architecture offers a very efficient pruning technique (Phone Deactivation Pruning [2]) that is much more efficient than pruning on likelihoods, but in most tasks, this approach did not substantially outperform standard continuous HMM systems.

2. THE NOVEL 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 [7]. The major difference in this approach is the fact that the acoustic processor is not replaced by a neural network, but that the Gaussian probability density component is retained and combined with a neural component in an appropriate manner. Contrary to [5] we propose to perform the MMI estimation of the NN on a system of low complexity and reuse it as feature transformation in a more complex system, where simple ML estimation is performed to keep the computational costs reasonably small.

2.1. ARCHITECTURE

The basic architecture of this hybrid system is illustrated in Figure 1. The neural net takes several adjacent feature vectors into account to produce an improved feature vector that is fed into the HMM system. This way, the additional neural component can be regarded as being part of the feature extraction, utilised to output discriminative feature vectors, whose distribution can be modelled well by mixtures of Gaussians. The architecture allows several ways of interpretation; 1. as a hybrid system that combines neural networks and (continuous) HMMs, 2. as a 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 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 trained to produce more discriminant feature vectors with respect to the trained HMM system. Of course, adding some kind of neural component increases the number of parameters, too, but the increase is much more moderate than it would be when increasing each Gaussian's dimensionality.

2.2. TRAINING OBJECTIVE AND ALGORITHM

Contrary to LDA, the training of the proposed feature transformation is performed in a (frame-based) MMI procedure that increases discrimination at the output level of a ML-trained HMM system. This way, all the HMM constraints are taken into consideration. The MMI criterion is usually formulated in the following way:

$$\hat{\lambda}_{MMI} = \underset{\lambda}{\operatorname{argmax}}(H_\lambda(X) - H_\lambda(X|W)) = \underset{\lambda}{\operatorname{argmax}} \frac{p_\lambda(X|W)}{p_\lambda(X)} \qquad (1)$$

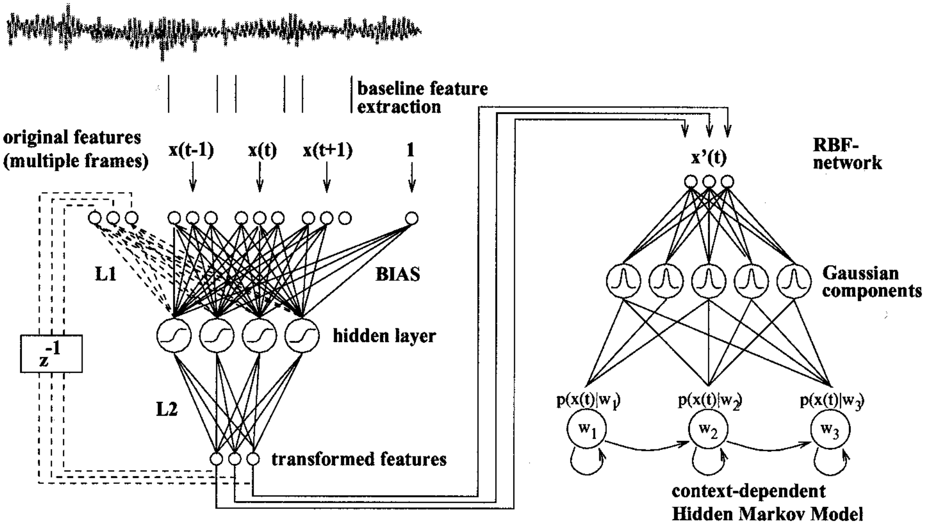

Figure 1: Architecture of the hybrid NN/HMM system

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

$$\hat{\lambda}_{MMI} \approx \underset{\lambda}{\mathrm{argmax}} \sum_{i=1}^{T} I_\lambda(x(i), w(i)) \approx \underset{\lambda}{\mathrm{argmax}} \prod_{i=1}^{T} \frac{p_\lambda(x(i)|w(i))}{p_\lambda(x(i))} \qquad (2)$$

where $p_\lambda(x(i))$ can be estimated as $\sum_{k=1}^{S} p_\lambda(x(i)|w_k)p(w_k)$ with S denoting the total number of HMM states and $(w_1, \dots w_S)$ the HMM states and $p(w_k)$ 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. However, it turned out, that only the incorporation of additional features in the probability calculation pipeline can provide more discriminative emission probabilities and a major advance in recognition accuracy. Thus, we experienced it to be more convenient to train an additional neural net in order to maximise Eqn. (2). Besides, this approach offers the possibility of improving a recognition system by applying a trained feature extraction network taken from a different system. Section 4 will report our positive experiences with this 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 the linear net denoted as a $(P+F+1) \times N$ matrix NET, each component $x'(t)[c]$ of the network output $x'(t)$ computes to

$$x'(t)[c] = \sum_{i=0}^{P+F} \sum_{j=1}^{N} x(t-P+i)[j] \cdot \mathrm{NET}[iN+j][c] \qquad \forall c \in \{1 \dots N\} \qquad (3)$$

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

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

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

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

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

$$\frac{\partial p_\lambda(x'|w)}{\partial x'[c]} = \sum_{j=1}^{C_w} d_{wj} \frac{(x[c]-m_j[c])}{\sigma_j[c]} \frac{1}{\sqrt{(2\pi)^n|\sigma_j|}} e^{-\frac{1}{2}\sum\limits_{l=1}^{N}\frac{(m_j[l]-x'[l])^2}{\sigma_j[l]}} \tag{6}$$

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

$$\frac{\partial I_\lambda(x'(i), w(i))}{\partial p_\lambda(x'(i)|w_k)} = \frac{\delta_{w(i),w_k}}{p_\lambda(x(i)|w_k)} - \frac{p(w_k)}{\sum\limits_{l=1}^{S} p_\lambda(x(i)|w_l)p(w_l)} \tag{7}$$

Thus, using the chain rule the derivative of the net's parameters with respect to the frame-based MMI criterion can be easily composed from Eqns. (4), (6) and (7) and a gradient descent procedure can be used to determine the optimal parameter estimates.
For reasons of simplicity this section 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. A detailed description of how to derive the gradient when using feed-forward networks (MLP) and recurrent networks (Jordan-Network) is given in [8].

3. MULTI STREAM SYSTEMS

In HMM-based recognition systems the extracted features are often divided into streams that are modelled independently. This is useful the less correlated the divided features are. In this case, the overall likelihood of an observation is computed as the product of all the individual likelihoods weighted by stream weights according to

$$p_\lambda(x|w) = \prod_{s=1}^{M} p_{s\lambda}(x|w)^{w_s} \tag{8}$$

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 are usually not totally independent and that by considering multiple input frames as illustrated in Figure 1 the independence often decreases. It is a common practice, for instance, to model the features' first and second order delta coefficients in

system	baseline system	LDA	linear MMI-net	MLP (H=36)	Jordan-Network
monophones one stream	24%	21%	21%	21%	-
monophones four streams	11.8%	11.0%	10.9%	10.8%	10.9%
triphones four streams	5.2%	5.3%	4.8%	4.7%	4.7%

Table 1: Word error rates achieved in the experiments

independent streams, so that 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. (8) inserted in Eqn. (2). With the weights w_s set to unity, the derivative with respect to the pdf $p_{\hat{s}\lambda}(x|w)$ of a specific stream $\hat{s}$ becomes

$$\frac{\partial I_\lambda(x'(i), w(i))}{\partial p_{\hat{s}\lambda}(x'(i)|w_k)} = \left(\prod_{s \neq \hat{s}} \frac{p_{s\lambda}(x(i)|w(i))}{p_{s\lambda}(x(i))}\right)\left(\frac{\delta_{w(i),w_k}}{p_{\hat{s}\lambda}(x(i)|w_k)} - \frac{p(w_k)}{\sum_{l=1}^{S} p_{\hat{s}\lambda}(x(i)|w_l)p(w_l)}\right) \tag{9}$$

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. (9).

4. EXPERIMENTS AND RESULTS

The experiments were run on 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 [9]. 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. Training of the MLP was performed with the RPROP algorithm. For training the weights of the recurrent connections we chose real-time recurrent learning. The average error rates were computed on the test-sets Feb89, Oct89, Feb91 and Sep92. The systems that incorporate an input transformation use one additional past and one additional future feature vector as input. All transformations, including the LDA, that we applied for reasons of comparison, have the same input and output dimensionality. Table 1 shows the recognition results with single stream systems in its first row. These systems simply use a 12-dimensional Cepstrum feature vector without the incorporation of delta coefficients. The proposed approach achieves the same performance as the LDA, but it is not capable of outperforming it.

The second row lists the recognition results with four stream systems that use the first and second order delta coefficients in additional streams plus log energy and delta coefficients in a forth stream. The MLP system trained according to Eqn. (8) slightly outperforms the other approaches. The incorporation of recurrent network connections does

not improve the system's performance.
The third row of the table lists the recognition results with four stream systems with a context-dependent acoustic modelling (triphones). The applied LDA and the MMI-networks were taken from the monophone four stream system. On the one hand, this was done to avoid the computational complexity that the MMI training objective causes 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.

5. 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 modelling 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 neighbouring 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.

References

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

[2] M. M. Hochberg et al., "The 1994 ABBOT Hybrid Connectionist-HMM Large-Vocabulary Recognition System", *Proc. ARPA Spoken Language Systems Technology Workshop,* 1995.

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

[4] Y. Zhao et al., "Hierarchical Mixtures of Experts applied to Continuous Speech Recognition" *Proc. IEEE-ICASSP,* 1995, pp. 3443–3446.

[5] Y. Bengio et al., "Global Optimization of a Neural Network - Hidden Markov Model Hybrid" *IEEE-Transactions on NN,* Vol. 3, No. 2, 1992, pp. 252–259.

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

[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, Dec. 1996, pp. 175–184.

[8] D. Willett, G. Rigoll, "Hybrid NN/HMM-Based Speech Recognition with a Discriminant Neural Feature Extraction" *Advances in Neural Information Processing Systems (NIPS-97),* Denver, Dec. 1997.

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