A NEW HYBRID SYSTEM BASED ON MMI-NEURAL NETWORKS FOR THE RM SPEECH RECOGNITION TASK

Gerhard Rigoll, Christoph Neukirchen, Jörg Rottland

Department of Computer Science
Faculty of Electrical Engineering
Gerhard-Mercator-University Duisburg
D-47057 Duisburg, GERMANY
e-mail: {rigoll,chn,rottland}@fb9-ti.uni-duisburg.de

ABSTRACT

We present a hybrid speech recognition system for speaker independent continuous speech recognition. The system combines a novel information theory based neural network (NN) paradigm and discrete Hidden Markov models (HMMs) including State-of-the-Art techniques like state clustered triphones. The novel NN type is trained by an algorithm based on principles of self-organization that achieves maximum mutual information between the generated output labels and the basic phonetic classes. The structure of the hybrid system is quite similar to a classical VQ-HMM system but the vector quantizer (VQ) is replaced by the NN. To evaluate the system we use the speaker independent part of the resource management (RM) database. We recently obtained an important improvement by introducing a novel kind of context dependent basic classes used by the acoustic processor. The average RM recognition result with a word-pair grammar is now 95,2% what is significantly better than a classical VQ-system, slightly better than a different hybrid system with a recurrent network as probability estimator, and very close to the best continuous probability density function (pdf) HMM speech recognizers.

1. INTRODUCTION

Today the most successful approach to the problem of speaker independent large vocabulary recognition of continuous speech seems to be Hidden Markov modelling (HMM). The traditional HMM systems try to model the distributions of the acoustic feature vectors by very complex sets of continuous parameter pdfs (e.g. mixtures of gaussian densities [1]). This involves estimating a large number of parameters in training and a CPU time consuming pdf calculation during recognition. The usage of discrete pdfs accelerates recognition speed but degrades recognition performance due to the information loss caused by the (most commonly used k-means) vector quantizer (VQ).

In parallel, hybrid ANN/HMM speech recognition systems combining both HMM technology and artificial Neural Networks (NN) have been developed to achieve State-of-the-Art performance [2]. These classical hybrid systems use a NN to estimate local probabilities without any assumptions to a family of pdfs. This leads to a very effective way of parameter sharing through all models and a fast Viterby decoder can be constructed [3]. However there are several drawbacks: e.g. when using large speech databases, the network weights of the hybrid system can only be trained in reasonable time with a special purpose computer hardware.

We propose a different hybrid approach that uses State-of-the-Art HMM techniques (e.g. state clustered triphone modelling) and a novel NN paradigm. The NN maps the continuous valued feature vectors to discrete label indices so it replaces the traditional VQ process. The network is trained in a self organizing approach that does not try to force constrained outputs. After training, the NN avoids the inherent information loss by preserving as much information about the underlying phonetic descriptions as possible (i.e. maximum mutual information (MMI)). Finaly the hybrid system uses triphone HMMs with discrete pdfs, so arbitrary distributions can be modelled and a fast decoder can be constructed.

2. MMI-NNs AS LABELERS

2.1. From minimum distortion to MMI-VQs

In speech recognition systems with discrete pdf HMMs typically codebook-VQs are used to map the acoustic feature vectors on discrete labels (e.g. in [4]). Codebook design is usually performed via unsupervised cluster algorithms (e.g. LBG, k-means) to achieve minimum distortion. However minimum distortion codebooks do not necessarily lead to an optimum speech recognition system that produces minimum word recognition error.

One major drawback of unsupervised codebook generation is the independence of HMM training. Recognition performance can be increased by adding information about the underlying class descriptions (e.g. phones, subphonetic units) to the VQ. Thus supervised cluster algorithms can be applied leading to slightly better recognition rates in many cases.

Another popular way to add phonetic information to the VQ is to replace the codebook by a NN-VQ. Multilayer perceptrons (MLP) and euclidean distance classifiers have been trained as VQ using the backpropagation algorithm or the LVQ paradigm, respectively. In the MLP case recognition results have improved [5], LVQ results were somewhat ambiguous [6]. Both training approaches tend to minimize the static frame-by-frame classification error rate of the NNs. But a better static classification of the VQ does not neccesarily lead to a better speech recognition system that has to classify a dynamic stream of acoustic features.

An optimum VQ can be constructed using an objective function that considers the information in the feature stream instead of a frame-by-frame criterion. It is shown in [7] that such kind of optimum VQ maximizes the mutual information (MMI) between the stream of basic classes used by the HMMs (e.g. phones, states) and the generated label stream. As proven in [7] discrete HMMs that are estimated by the ML criterion and MMI-VQs fit extremely well since the VQ preserves as much phonetically important information contained in the acoustic feature stream as possible.

2.2. The MMI-NN paradigm for labelers

Considering the facts given above it appears very promising to replace the VQ by a neural network that is trained on the MMI objective function. A MMI-NN has been proposed in [8] for vector quantization in a relatively simple phoneme recognition task. The main goal of the research project presented in this paper is to transfer the improvements reported in [8] to a more difficult speech recognition task using the RM database.

The MMI-NN used here is a very simple one-layered neural network similar to an LVQ topology (see Fig. 1). There is full connectivity between all input nodes and all output nodes of the network. Each connection is associated with a weight w_{lk}. Network input is the current acoustic feature vector $\vec{x}$. The activation f_k of the k-th output node is given by the euclidean distance between input activation and the corresponding NN-weights:

$$f_k = \sum_l (w_{lk} - x_l)^2$$

The output node with the lowest activation determines the NN-output label y_k (see Fig. 1).

This kind of NN topology has the advantage of being flexible in output layer size C, i.e. number of different output labels can be chosen arbitrarily. So there is no limitation on very small codebook sizes, as in MLP-based hybrid systems where output layer size is limited by the number of used phones or states (see [5]). On the other hand the number of phonetic classes can be even larger (e.g. using states of triphones) than the output layer size.

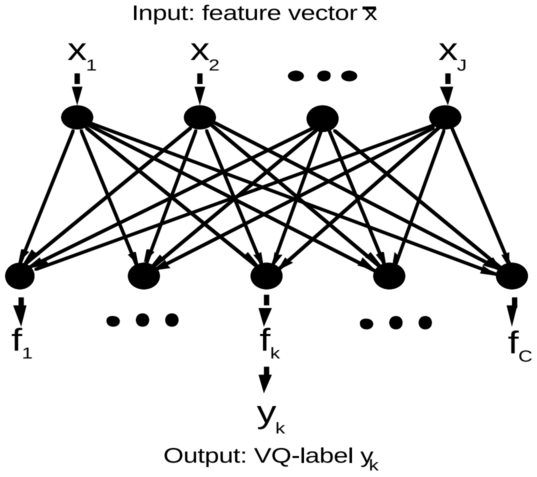

Fig. 1: Topology of the one-layer MMI-NN

To train the MMI-NN we use an iterative learning approach. The training data has to be aligned on the used phonetic class boundaries using a Viterbi aligner. In a first step the network weights are set to fixed initial values. Suitable initial values may be obtained from k-means prototypes or from the weights of a trained SOM. Then the mutual information value $I(W,Y)$ of the initialized NN is calculated using all phonetical labeled training data vectors. Then, all NN weights are changed by a fixed offset value Δ step by step and the new value of $I(W,Y)$ is determined after each step. If there is an increase of the mutual information the weight change will be accepted, otherwise the change will be discarded. In this learning algorithm the order of changing the NN weights is given by the frequency of firing output nodes, i.e. the weights of the most frequently firing output neuron will be changed at first, and so on. If there is no increase of $I(W,Y)$ after a few iterations then the learning scheme is repeated using a smaller fixed offset value Δ. A more detailed description of the MMI training algorithm can be found in [8].
This information theory based learning algorithm is different to other NN paradigms in some specific ways: During iterative training of the MMI-NN there is no forced network output when presenting the input vectors (unlike online backpropagation e.g. for MLPs). Instead of this, all NN-output labels and all phonetic labels of the training data are considered simultaneously. In each MMI-training iteration the network weights are changed to increase the amount of mutual information by principles of self organization.

3. HYBRID SYSTEM DESCRIPTION

3.1. Construction of the MMI-NNs

For NN training, the training database has to be phonetically aligned. This can be e.g. achieved with a system described in [1]. The MMI-NN is trained in reasonable time on a standard workstation by applying the

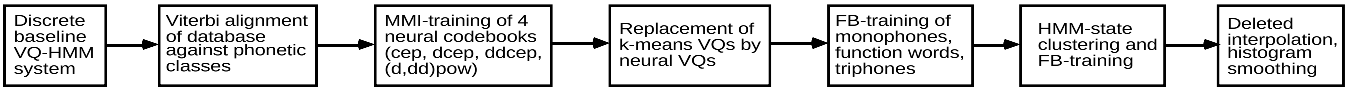
Discrete baseline VQ-HMM system
Viterbi alignment of database against phonetic classes
MMI-training of 4 neural codebooks (cep, dcep, ddcep, (d,dd)pow)
Replacement of k-means VQs by neural VQs
FB-training of monophones, function words, triphones
HMM-state clustering and FB-training
Deleted interpolation, histogram smoothing

Fig. 2: Construction of the context-dependent hybrid system

self-organizing learning algorithm described above. Several adjacent frames of acoustic features are integrated to form a spliced feature vector, that increases the input layer size. This multiple frame approach leads to an extension of acoustic context dependency of the NN as stated in [2]. To decrease the size of the MMI-NN in order to reduce training time the network is splitted up into several smaller nets in a similar way as multiple codebooks [4]; i.e. each block of acoustic features and their derivatives are assumed to be independent and are labeled by separate networks.

3.2. Training of context dependent HMMs

In the hybrid system discrete pdf triphone HMMs are used to model the distribution of the labels generated by the MMI-NNs. All used HMMs have a strict left-to-right topology and three emitting states are used. The output probability of each state is calculated by the product of each neural codebook probability. Since the structure of the hybrid system is equivalent to a traditional HMM system, all well known State-of-the-Art modelling techniques can be applied. Thus HMM construction follows the strategy described in [4]:

At first 49 monophone models are trained via the forward backward algorithm. Then the most frequent 33 function words are modelled by seperate HMMs. To capture coarticulation effects 2309 within-word triphone models are built. Some interpolation and smoothing techniques have to be used to handle estimation problems due to sparse training data; i.e. data driven HMM-state clustering, deleted interpolation between monophones and triphones, smoothing of discrete pdfs by gaussian distributions.

The complete development of the MMI-NN hybrid system is summarized in Fig. 2.

4. EXPERIMENTS AND RESULTS

For system evaluation we use the speaker independent DARPA Resource Management (RM) 997 word task. All used MMI-NNs and HMMs are trained with 3390 sentences spoken by 109 different speakers. Every 10ms 39 acoustic features are extracted: 12 MFCCs, log-energy and their first and second derivatives. We use the context of 3 adjacent frames (30ms) to form a 117 element spliced vector. The network is subdivided into 4 independent parts: Three MMI-NNs (input layer size: 36) for the MFCCs and their derivatives and one MMI-NN (input layer size: 9) for log-energy plus derivatives. The output layer sizes of all MMI-NNs are chosen to 200.

As starting point for MMI-NN training the prototypes of a k-means VQ (codebook size: 200) are used to initialize the network weights. Each NN is trained by 15 iterations of the information theory based learning algorithm. The phonetic class label stream consists of the states of triphones of the RM sentences obtained by Viterbi alignment with a standard HMM system.

To evaluate the recognition performance of the new hybrid system the standard DARPA test sentences Feb'89, Oct'89, Feb'91 and Sep'92 are used. Recognition of the test sentences is performed via Viterbi decoding including a beam search technique and the official word pair grammar (test perplexity: 60).

Fig. 3 shows the evolution of RM recognition rates during hybrid system development. For comparison we also report recognition rates of other systems: The bottom line represents a traditional k-means VQ system using discrete monophone HMMs, recognition rate is 86%. The same system with state-clustered triphone HMMs (middle line) achieves 93,2%, thus context dependent modelling improves the discrete system by ca. 7% (absolute). The top line represents a State-of-the-Art continuous pdf system (HTK), it is more than 2% (absolute) better than the traditional discrete system.

A hybrid MMI-NN system with monophone basic NN classes and monophone HMMs achieves a 88,8% average recognition rate, that is an improvement by nearly 3% (absolute) compared to the VQ system. When using function word HMMs or state-clustered triphone HMMs the hybrid approach results in 90,3% and 94.5%, respectively. When using 3 adjacent frames to form a spliced feature vector what means to capture temporal context, the recognition rate of the hybrid system further improves up to 94,9%. This rate can be further increased to 95,2%, by introducing a novel context dependent NN that uses states of triphones as basic phonetic classes. So the MMI-NN hybrid system performes significantly better than any other classical discrete HMM system.

The recognition results of the hybrid system for all used RM test sets are given individually in the boldface column of Tab. 1 (rate given in parenthesis is accuracy). Again the table shows clearly that the hybrid system outperformes a classical discrete HMM recognizer, which only differs in the kind of VQ, in all four test sets. We also compare the MMI-NN hybrid with the ANN/HMM hybrid system presented in [3], that uses a recurrent network as probability estimator. As shown in Tab. 1 the MMI-NN system performs nearly equal or even better than this hybrid system in all tests, although it has to be mentioned that this system has been recently improved by introducing context dependency. Finally we give the

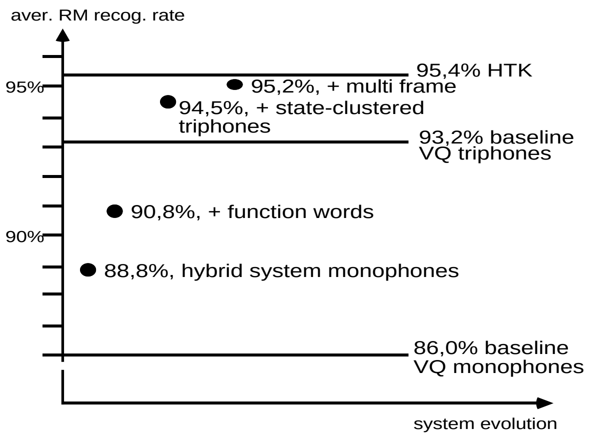

Fig. 3: Evolution of average RM recognition rates

recognition results of a State-of-the-Art continuous pdf system (HTK) for all tests. It can be seen that there is still a small gap between the rates of HTK and the MMI-NN hybrid system. But future system improvements will hopefully lead to a further increase of hybrid recogniton rates.

It should be noted, that the reported recognition results of all compared systems were obtained without cross-word modelling, without gender modelling and without a multiple pronounciation lexicon.

5. CONCLUSIONS

The results of our experiments have clearly shown that a classical discrete HMM system can be improved by replacement of the VQ by a NN that is trained to achieve maximum mutual information between output labels and basic phonetic classes. The obtained recognition rates are very close to the best continuous pdf systems and the results of the MMI-NN hybrid might be the highest ever reported for a system based on discrete HMMs. Future system improvement can be done in both the HMM part and the NN part. Any classical HMM system improvements (as cross-word triphones, corrective training, decision-tree based state-clustering, signal processing, better smoothing techniques) can be easily carried out with the hybrid system. Furthermore we will focus our experiments on NN development, such as joint NN optimization [8], cross-validation, different NN topologies and activation functions as well as faster NN training algorithms.

6. REFERENCES

[1] P.C. Woodland, S.J. Young. "The HTK tied-state continuous speech recognizer," *Proc. Eurospeech*, 1993, pp. 2207-2210.

[2] N. Morgan, H. Bourlard. "Neural Networks for Statistical Recognition of Continuous Speech," *Proc. IEEE*, Vol. 83, No. 5, May 1995, pp. 742-770.

[3] A.J. Robinson. "An Application of Recurrent Nets to Phone Probability Estimation," *IEEE Trans. Neural Networks*, Vol. 5, No. 2, Mar. 1994, pp. 298-305.

[4] K.F. Lee, H.W. Hon, R. Reddy. "An Overview of the SPHINX Speech Recognition System," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 38, No. 1, Jan. 1990, pp 35-45.

[5] P. Le Cerf, W. Ma, D. Van Compernolle. "Multilayer Perceptrons as Labelers for Hidden Markov Models," *IEEE Trans. Speech Audio Processing*, Vol 2, No. 1, Jan. 1994, pp. 185-193.

[6] G. Yu et al. "Discriminant analysis and supervised vector quantization for continuous speech recognition," *Proc. IEEE-ICASSP*, 1990, pp. 685-688.

[7] M. Ostendorf, J.R. Rohlicek. "Joint quantizer design and parameter estimation for discrete Hidden Markov Models," *Proc. IEEE-ICASSP*, 1990, pp. 705-708.

[8] 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.

RM SI word recognition rate with word pair grammar: Correct (Accuracy)				
Test Set	**Hybrid MMI-NN system**	Baseline k-means VQ system	Hybrid system with rec. NN pdf estim. [3]	Continuous pdf system (HTK) [1]
Feb.'89	**96,5 % (95,7 %)**	94,3 % (93,6 %)	95,7 % (95,0 %)	96,0 % (95,5 %)
Oct.'89	**95,4 % (94,6 %)**	93,5 % (92,0 %)	94,8 % (94,2 %)	95,4 % (94,9 %)
Feb.'91	**95,9 % (95,3 %)**	94,4 % (93,5 %)	95,4 % (94,4 %)	96,6 % (96,0 %)
Sep.'92	**93,1 % (91,2 %)**	90,7 % (88,9 %)	91,5 % (90,0 %)	93,6 % (92,6 %)
average	**95,2 % (94,2 %)**	93,2 % (92,0 %)	94,3 % (93,4 %)	95,4 % (94,7 %)

Tab. 1: Comparison of SI RM recognition rates between MMI-NN hybrid system and other systems