LARGE VOCABULARY SPEECH RECOGNITION WITH CONTEXT DEPENDENT MMI-CONNECTIONIST / HMM SYSTEMS USING THE WSJ DATABASE[1]

J. Rottland, Ch. Neukirchen, D. Willett, G. Rigoll
{rottland,chn,willett,rigoll}@fb9-ti.uni-duisburg.de
http://www.fb9-ti.uni-duisburg.de
Gerhard-Mercator-University Duisburg
Department of Computer Science
Bismarckstr. 90, D-47057 Duisburg, Germany

ABSTRACT

In this paper we present a context dependent hybrid MMI-connectionist / Hidden Markov Model (HMM) speech recognition system for the Wall Street Journal (WSJ) database. The hybrid system is build with a neural network, which is used as a vector quantizer (VQ) and an HMM with discrete probablility density functions, which has the advantage of a faster decoding. The neural network is trained on an algorithm, that tries to maximize the mutual information between the classes of the input features (e.g. phones, triphones, etc.) and the neural firing sequence of the network.
The system has been trained on the 1992 WSJ corpus (si-84). Tests were performed on the five- and twentythousand word, speaker independent (si_et) tasks. The error rates of a new context dependend neural network are 29% lower (relative) than the error rates of a standard (k-means) discrete system and the error rates are very close to the best continuous/semi-continuous HMM speech recognizers.

1. INTRODUCTION

There are several ways to build hybrid systems by combining neural networks with hidden Markov models. The most common approach is to use the network as a probability estimator for the HMMs. Our approach is different, because in our hybrid system architecture, a discrete baseline HMM speech recognition system is combined with a neural network used as vector quantizer that is trained by a new neural network training paradigm in order to maximize the mutual information between the classes of the input features presented during training and the corresponding output generated by the network [1]. Figure 1 shows an example of a single layer network.

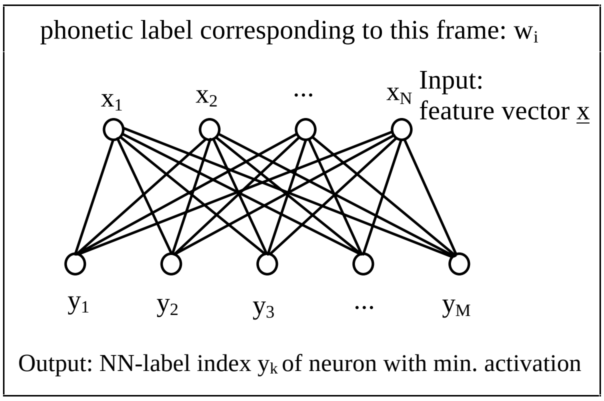
phonetic label corresponding to this frame: w_i
x_1
x_2
...
x_N
Input:
feature vector x
y_1
y_2
y_3
...
y_M
Output: NN-label index y_k of neuron with min. activation

Figure 1: Basic structure of the single layer MMI-NN

The training algorithm tries to maximize the mutual information I(Y,W) as given in equation (1).

$$\begin{aligned} I(Y,W) &= H(W) - H(W|Y) \\ &= -\sum_{W} p(w_i) \cdot \log_2 p(w_i) \qquad (1) \\ &\quad + \sum_{Y} \sum_{W} p(w_i, y_k) \cdot \log_2 p(w_i|y_k) \end{aligned}$$

Recently it has been shown, that it is possible to derive the exact proof that such an MMI training leads to neural codebooks that are optimal for the combination with discrete pattern classifiers [2]. The basic structure of such a system is still that of a discrete system, including its speed and efficiency, but due to the special training of the neural acoustic processor, the performance of this hybrid system is much better than that of any discrete HMM-based system. It has been demonstrated, that the resulting hybrid system obtains basically the same results as the best equivalent continuous parameter HMM systems on the RM database [3]. For the development of this new approach, the RM database was used purposely, because it is more compact and therefore more suitable for running risky and time consuming

[1] This work was supported by the *German National Science Foundation (DFG)*, project # Ri 658/3-1. Responsibility for the content of this paper is by the authors.

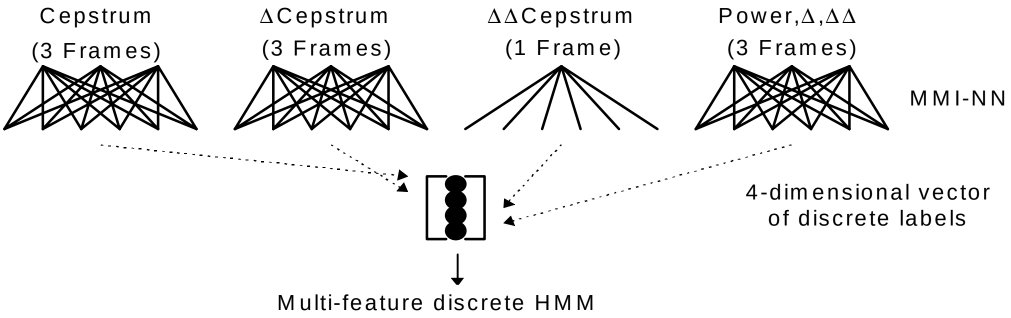

Figure 2: Context dependent connectionist vector quantizer with four features

experiments. Our goal is now to transfer these results and experiences to a larger and even more demanding database and to build a hybrid MMI-Connectionist / HMM speech recognition system for the WSJ database.

2. SYSTEM DESCRIPTION

The system presented here was build from the 1992 ARPA WSJ corpus (WSJ0) [4]. For the training of the MMI-NN and the HMMs the speaker independent (si-84) data was used. The features were 12 Mel frequency cepstral coefficients and log energy, plus the first and second order derivatives. The cepstral features were normalized for each sentence by subtraction of the cepstral mean calculated over the sentence. This results in 39 features per frame. Those 39 features were split in four streams (cepstrum, 1st and 2nd derivatives and power). For each stream a separate single layer neural network was trained. The size of the input layer was 12 for the first three streams and 3 for the fourth stream. The size of the output layer was 300 for all four networks. This led to four discrete labels for each frame. Those labels are the inputs of the multi-feature, discrete pdf HMMs. The topology of the HMMs were three state left-to-right models without skips. The total number of weights were 11700 for all four networks.

A second system was trained with some more advanced neural networks. In these networks the size of the input layer has been enlarged to three adjacent frames for all but the 2nd order derivative of the cepstrum. So, the size of input layer is three times larger as for the single frame network. The ΔΔCepstrum network is still single frame because the computation of the 2nd order derivative already uses the information of nine frames. The goal of enlarging the size of the input layer to three frames is to get a better context dependency for the acoustic feature vectors. The phonetic class of this three frame feature vector, which is needed for the training algorithm was taken from the center frame. Because of the three input frames, this network is named Multi-Frame (MF3) in this paper (Figure 2). The number of weights for all four networks increased to 27900 using the MF3 approach.

Time aligned transcriptions were needed for the training of the MMI-NN. Training can be performed on a standard workstation due to the relatively small size of the neural networks. Additionally, we have also implemented our information theory based training approach on a SPERT-II board purchased from the International Computer Science Institute (ICSI) in Berkeley. For a detailed description of the training algorithm see [1][12]. To achieve those transcriptions we used initial models from our RM-system to align the training data. Alignment was performed on monophones.

The pronunciations were taken from a lexicon provided by CMU [5]. Some extensions were made to the basic version of the lexicon, by scripts, also provided by CMU, which merged/introduced some phones in special contexts. This resulted in a phone set of 50 phones plus 2 phones for silence and an optional inter-word silence.

The triphone system led to 8591 triphone models with 25767 states which were then state-clustered to approx. 6500 states by a decision tree based clustering algorithm [6]. The cross word triphone system led to 21.259 models with more than 60.000 states, which were again state-clustered to approx. 6.500 states.

Recognition is done with a Viterbi decoder. The language models used during the tests were the original 1992 5k bigram and trigram language models with a perplexity of 110 and 62 for the 5k closed vocabulary test and the original 1992 20k language models.

Trigram recognition is done by a two pass decoding strategy. The first pass is done by a bigram recognition. The output of this pass is a word lattice

Improvement in word error rates					
	k-means	**MMINN**	Error reduction	**MF3-MMINN**	Error reduction
Monophones	29.1%	22.4%	**23.0%**	21.1%	**27.5%**
Triphones (Word internal)	13.4%	10.5%	**21.6%**	9.4%	**29.9%**
Triphones (Cross word)	-	9.2%	-	8.0%	**13.0%**

Table 1: Word error rates for the Nov'92 WSJ 5k closed vocabulary evaluation test with a bigram language model for the discrete / MMI-Connectionist approach

which is then rescored in the second pass with the trigram language model. In our test we are neglecting the fact, that this two pass strategy is not correct for cross word triphones, because there the acoustic score, which is derived from a special context in the first pass may be combined with a language model score of another context in the second pass.

3. EXPERIMENTS AND RESULTS

To verify the improvements of the MMI-Connectionist / HMM we trained an equivalent discrete HMM-system with a k-means vector quantizer comparable to the system presented in [7]. Table 1 gives the improvements in word error rates comparing the k-means system to the MMI-Connectionist system for monophones and for word-internal triphones. In the right column of Table 1 the results for the multi frame MMI network are given. The multi frame approach gives a further reduction in word error rates. For cross word triphones only hybrid results are available. Again the multi frame approach gives the best error rates.

The reduction of the word error rate in Table 1 is in the same order of magnitude as for our RM system, comparing a k-means system with a MMI-Connectionist system. In [8] the error reduction for the RM database was in average 18% for word-internal triphones. Here in the 5k WSJ test the improvement is even larger (21.6%) for the single frame network. This shows that the MMI-Connectionist approach also works well for larger vocabularies. It even seems to turn out that the MMI-Connectionist approach works better the larger the database is.

All results were produced on the 5k closed vocabulary test set and the 20k open vocabulary test set using the original bigram language models. To compare this system to standard continuous systems [9] some results for cross-word triphones and trigram language models are given in Table 2. Comparing the results in Table 2 with the official DARPA benchmark results in [9] shows that this discrete system already is one of the best systems on this task. All hybrid results in Table 2 are for the three frame network.

The results in Table 2 show that at a vocabulary size of 5k the difference between the best hybrid result and the best result in the official test in 1992 is very small with 0.4% absolute. For the 20k test the difference is larger, because there is no result available for the hybrid 20k cross word triphones. However the result for the 20k word internal triphones is as promising as the result for the 5k word internal triphones.

Test Set	System	bg LM	tg LM
si-84 5k nvp	hybrid (3Frames) wint triphones	9.4%	6.4%
si-84 5k nvp	hybrid (3Frames) xwrd triphones	8.0%	5.7%
si-84 5k nvp	best continuous in evaluation [9]	6.9%	5.3%
si-84 20k nvp	hybrid (3Frames) wint triphones	17.4%	14.7%
si-84 20k nvp	best continuous in evaluation [9]	15.2%	12.8%

Table 2: Comparison of word error rates for the three frame connectionist system with continuous systems for different model types (word internal/cross word triphones) and grammars (bigram bg, trigram tg) on the Nov'92 WSJ 5k NVP evaluation test

4. FUTURE WORK

The results in Table 2 show that the continuous systems [9][10] still have lower error rates than our MMI-Connectionist HMM system. To improve our system we plan to train our networks on a transcription aligned on the states of the models instead of the alignment on the models itself which we used so far. Another point which has to be checked is the influence of the dictionary to the recognition accuracy. As described above, we are using the CMU lexicon. In [11] the performance of the LIMSI speech recognizer using the CMU lexicon is compared with

the use of the LIMSI lexicon. There the CMU lexicon resulted in a word error rate which is more than 1% absolute higher for a 20k test. So, with the use of the LIMSI lexicon we expect that our word error rates will be even closer to those of the continuous systems.
At our institute, a demonstration version is available on a standard workstation, which runs in 2 times real-time for the 20k-WSJ speech recognition task. Using more advanced decoding techniques, we expect to achieve real-time performance in the near future.

5. CONCLUSION

In this paper we compared the performance of a “standard“ discrete HMM speech recognizer with a MMI-Connectionist / HMM speech recognizer. It shows that the MMI-Connectionist approach outperformed the k-means approach by far for the 5k WSJ evaluation test.
Furthermore, we showed that this system achieves the same or even better error reduction than our RM system when compared to a k-means system. So one can expect that this “discrete“ system is the best discrete speech recognizer and it has the potential to become as good, or even better, than other state-of-the-art recognizers, even on large databases like the WSJ. Additionally this system has the advantage of less computational complexity during recognition due to the discrete nature of the MMI-Connectionist system, resulting in a faster system, which is especially important for the very large vocabulary of the WSJ corpus.
By looking at the system now, it should be pointed out again, that this is the only discrete system ever build for the WSJ database and only the second hybrid system ever tested on this task. Even in this early stage the resulting error rate compares well to other systems [9]. Keeping in mind the fact that the MMI-Connectionist approach is not yet fully exploited and still perfectible, we hope that we will be able to build one of the most powerful systems for the WSJ database in the near future.

6. REFERENCES

[1] Rigoll G.: *Maximum Mutual Information Neural Networks for Hybrid Connectionist-HMM Speech Recognition Systems*. IEEE Trans. on Speech and Audio Processing, Special Issue on Neural Networks for Speech, Vol. 2, No. 1, January 1994

[2] Rigoll G., Neukirchen Ch.: *A new approach to hybrid HMM/ANN speech recognition using mutual information neural networks*, Advances in Neural Information Processing Systems 9, NIPS*96 Denver 1996

[3] Rigoll G., Neukirchen Ch., Rottland J.: *A new hybrid system based on MMI-Neural Networks for the RM speech recognition task*. Proc. IEEE Intern. Conference on Acoustics, Speech, and Signal Processing, Atlanta 1996

[4] Paul D. and Baker J.: *The Design for the Wall Street Journal-based CSR Corpus*, DARPA Speech and natural language Workshop, February 1992

[5] Carnegie Mellon Pronouncing Dictionary, CMUDICT ftp://ftp.cs.cmu.edu/project/fgdata/dict/cmudict.0.4.Z plus some extension scripts from Bob Weide (weide@cs.cmu.edu)

[6] Odell J., Woodland P., Young S.: *Tree-Based State Clustering for Large Vocabulary Speech Recognition*, International Symposium on Speech, Image Processing and Neural Networks, Hong Kong 1994

[7] Lee K.-F., Hon H.-W., Reddy R.: *An Overview of the SPHINX Speech Recognition System*, IEEE Transactions on ASSP, Vol. 38, No. 1, January 1990

[8] Rigoll G., Neukirchen Ch., Rottland J.: *Large Vocabulary speaker independent continuous speech recognition with a new hybrid system based on MMI-Neural Networks*, Proc. Eurospeech Madrid 1995

[9] Pallett D., Fiscus G., Fisher W., Garofolo J.:*Benchmark Tests for the DARPA Spoken Language Program*, Human Language Technology, Plainsboro NJ, 1993

[10] Woodland P.C., Odell J.J., Valtchev V. & Young S.J.: *Large Vocabulary Continuous Speech Recognition Using HTK*. Proc. IEEE Intern. Conference on Acoustics, Speech, and Signal Processing, Adelaide 1994

[11] Lamel L. and Adda G.: *On Designing Pronunciation Lexicons for Large Vocabulary, Continuous Speech Recognition*, Intern. Conference on Spoken Language Processing, Philadelphia 1996

[12] Neukirchen Ch, Rigoll G.:*Advanced Training Methods and New Network Topologies for Hybrid MMI-Connectionist/HMM Speech Recognition Systems*. Proc IEEE Intern. Conference on Acoustics, Speech, and Signal Processing, Munich 1997