PERFORMANCE OF HYBRID MMI-CONNECTIONIST / HMM SYSTEMS ON THE WSJ SPEECH DATABASE[1]

J. Rottland, Ch. Neukirchen, D. Willett

Gerhard-Mercator-University Duisburg
Department of Computer Science, Bismarckstr. 90, BC
D-47057 Duisburg, Germany

{rottland,chn,willett}@fb9-ti.uni-duisburg.de
http://www.fb9-ti.uni-duisburg.de

ABSTRACT

In this paper, a hybrid MMI-connectionist / hidden Markov model (HMM) speech recognition system for the Wall Street Journal (WSJ) database is presented. The HMM part of this system uses discrete probability density functions (pdf). The neural network (NN) is used to replace a classical vector quantizer (VQ) like a k-means or LBG algorithm, which are typically used in discrete HMM systems. The NN is trained on an algorithm, that tries to achieve maximum mutual information (MMI) between the generated output labels and the underlying phonetic description.
The system has been trained and tested with the five thousand word speaker independent WSJ task. The error rates of the MMI-Connectionist approach are 21% lower than the error rates of a k-means system. The system achieves error rates which have been achieved before only by the best continuous/semi-continuous HMM speech recognizers, with the advantage of a faster recognition algorithm.

1. INTRODUCTION

MMI-connectionist/HMM speech recognition represents a new approach for hybrid speech recognition systems and has been tested extensively for the Resource Management (RM) database [1]. In this hybrid system architecture, a discrete baseline HMM speech recognition system is combined with a neural network used as vector quantizer and 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 neural firing sequence of the network [2]. 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 [13]. 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 an equivalent continuous parameter HMM systems on the RM database [1]. We now believe that this is the most powerful discrete system ever built and the only discrete system capable of competing with the best continuous parameter HMM systems. New advanced neural network training methods [3] have shown further improvements, so that it can be expected that this hybrid approach will soon represent one of the most powerful acoustic processors for continuous speech recognition.
For the development of this new approach, the RM database was used purposely, because it is more compact and therefore it is more suitable for running risky and time consuming 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. It is expected that such a system - once equipped with all special features built into the RM-based version - can compete with the most advanced systems designed for the WSJ database. We also believe that such a system would be one of the first

[1] This work was supported by the Deutsche Forschungsgemeinschaft, project # Ri 658/3-1. Responsibility for the content of this paper is by the authors.

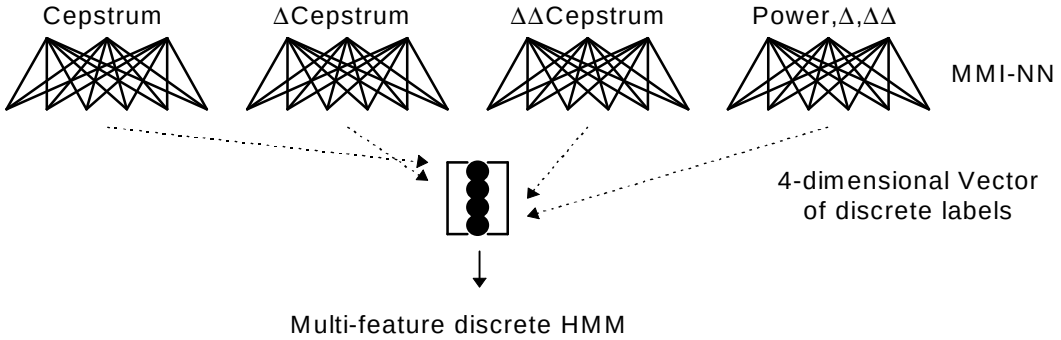

Figure. 1 Connectionist vector quantizer with four features

discrete systems ever built and tested for the WSJ database.

2. SYSTEM-DESCRIPTION

Our system was build from the 1992 ARPA WSJ corpus (WSJ0). For the training of the MMI-NN and the HMMs the speaker independent (SI-84) data has been used. The features used 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 resulted in four parameter 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 is 12 for the cepstrum networks and 3 for the power network. The size of the output layer is 300 for all four networks. This led to four discrete labels for each frame (Figure 1). Those labels are the inputs of the multi-feature, discrete pdf HMMs. For a better context dependency the size of the input layer of each neural network can be enlarged to three, five or more adjacent frames.

The topology of the HMMs were three state left-to-right discrete models without skips. Only one set of models has been trained for all gender (gender independent).

The pronunciations were taken from a lexicon provided by CMU [4]. This resulted in a phone set of 50 phones plus 2 phones for silence and an optional inter-word silence.

Time aligned transcriptions were needed for the training of the MMI-NN. (For a detailed description of the training algorithm see [2].) To obtain those transcriptions we used initial models from our RM-system to align the training data. After the MMI-training of the neural networks, the training data was quantized using this network. First an initial monophone system was build. That system was used to initialize context dependent systems. This context dependent systems were a triphone system with word internal triphones and a triphone system with cross word triphones. Due to the fact that the triphone systems led to a large number of states, clustering techniques had to be applied, to suppress sparse data problems. Another problem of triphones and especially of cross word triphones is the large amount of unseen triphones in the test data (vocabulary) and the large number of triphones with little training data. Therefore a clustering procedure with a phonetic decision tree was used, similar to [5].

The triphone system led to 8.591 triphone models with 25.767 states, which were then state-clustered to

Improvement in word error rates					
(k-means) Monophones	(MMINN) Monophones	Error reduction	(k-means) Triphones	(MMINN) Triphones	Error reduction
29.1%	22.4%	**23.0%**	13.4%	10.5%	**21.6%**

Table 1: Word error rates for the Nov'92 WSJ 5k closed vocabulary evaluation test with word-internal Triphones and a bigram Language Model for the discrete / MMI-Connectionist approach

comparison of word error rates		
word-internal triphones bigram language model	cross-word triphones bigram language model	cross-word triphones trigram language model
10.5	8.9	6.6

Table 2: %Word error rates of the connectionist system for different model types and grammars on the Nov'92 WSJ 5k evaluation test

approx. 6.500 states by the decision tree based clustering algorithm. The cross word triphone system led to 21.259 models with over 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.
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 which is then rescored in the second pass with a trigram language model.

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 [6]. 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. All results were produced on the 5k closed vocabulary test set with the original bigram language model.
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 [7] 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%). This shows that the MMI-Connectionist approach also performs well for larger vocabularies. It even seems to turn out that the MMI-Connectionist approach works better the larger the database is.
Table 2 shows the improvements, achieved by using cross word triphones and a trigram language model. This demonstrates, that enhancements, which are used in "standard" continuous pdf HMM speech recognizers can be applied to this hybrid system as well.

4. FUTURE WORK

Comparing the results of the MMI-Connectionist HMM system in Table 2 with equivalent continuous pdf HMM systems [8][9] shows that there is still a gap between our discrete system and the best continuous pdf HMM systems. This gap can be closed by using more of the features used in our RM-system in [1], which are in detail: Introduction of a larger input layer by using multiple adjacent frames as an input vector. A second feature would be an alignment on states rather than on phones for a better training of the MMI neural networks. The use of our new multilayer MMI training [3] is a third way to improve the recognition rates. For the RM system all the improvements together had the effect that the connectionist speech recognition system performed as well as a continuous pdf HMM system [1]. This was an improvement in word error rate of further 22% compared to the single frame connectionist system in [3], which is a system comparable to the system presented in this paper.
Up to now all the neural networks were trained on workstations, without any special hardware, in an acceptable time. But with the introduction of the above described features, the training time for the WSJ database will rise into a region, where it is no longer practicable to train these networks on workstations. Therefore we will use a vector microprocessor system to increase the speed of the training. The hardware we are going to use is SPERT-II board, which has proved its capability for fast estimation of neural weights [10]. Implementation of the MMI training algorithm on the SPERT board is expected to be finished soon.
Because of the promising results of our approach on the 5k task the system will be adapted to the 20k WSJ task shortly.

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. As stated above, the introduction of some advanced features to the neural network led to a further improvement of 22% for the RM database [1]. A comparable improvement on the WSJ system presented in this paper would lead to a system, which would become as good as most continuous pdf systems on the WSJ database. So one can expect that this “discrete“ system 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 decoding.

By looking at the system now, in its early implementation stage, it should be pointed out again, that this is probably the only discrete system ever build for the WSJ database and only the second hybrid system [12] ever tested on this task. Even in this early stage the resulting error rate compares well to other systems [9]. Keeping in mind the numerous improvements which are not yet implemented, the special fact that the MMI-Connectionist approach seems to work even better for the WSJ database than for the RM database, and the general fact that the entire 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.

REFERENCES

[1] 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

[2] 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

[3] Neukirchen Ch., Rigoll G.: Training of MMI neural networks as vector quantizers. Internal Report TI-01, March 1996, http://www.fb9-ti.uni-duisburg.de

[4] ftp://ftp.cs.cmu.edu/project/fgdata/dict/cmudict.0.4.Z

[5] 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

[6] 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

[7] 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

[8] 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

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

[10] Wawrzynek J., Asanovic K., Kingsbury B., Beck J., Johnson D., and Morgan N.: SPERT-II: A Vector Microprocessor System, IEEE Computer March 1996

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

[12] Robinson T.,Hochberg M.,Renals S.:IPA: Improved Phone Modelling with Recurent Neural Networks. Proc. IEEE Intern. Conference on Acoustics, Speech and Signal Processing, Adelaide 1994

[13] 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, Denver 1996