A NEURAL NETWORK BASED, SPEAKER INDEPENDENT, LARGE VOCABULARY, CONTINUOUS SPEECH RECOGNITION SYSTEM: THE *WERNICKE* PROJECT

A. J. Robinson * L. Almeida † J.-M. Boite ‡ H. Bourlard ‡ F. Fallside * M. Hochberg * D. Kershaw *
P. Kohn § Y. Konig § N. Morgan § J. P. Neto † S. Renals * M. Saerens ‡ C. Wooters §

‡*Lernout and Hauspie SpeechProducts, Belgium*
∗ *Cambridge University Engineering Department, UK*
†*Instituto de Engenharia de Sistemas e Computadores (INESC), Portugal*
§*International Computer Science Institute (ICSI), USA*
(Author list is alphabetical with the exception of the typist.)

ABSTRACT

This paper describes the research underway for the ESPRIT WERNICKE project. The project brings together a number of different groups from Europe and the US and focuses on extending the state-of-the-art for hybrid hidden Markov model/connectionist approaches to large vocabulary, continuous speech recognition. This paper describes the specific goals of the research and presents the work performed to date. Results are reported for the resource management talker-independent recognition task. The paper concludes with a discussion of the projected future work.

Keywords: *Recognition, Neural Nets, HMM.*

1. BACKGROUND

WERNICKE is an ESPRIT funded Basic Research project which started in October 1992. The project aims are to exploit the hybrid structures consisting of hidden Markov model-artificial neural network (HMM-ANN) combinations to improve the state-of-the-art in large vocabulary speech recognition. The project brings together partners with existing skills and baseline systems in the area: Lernout and Hauspie SpeechProducts (LHS, Belgium) and International Computer Science Institute (ICSI, US) for hybrid HMM-ANN structures using feedforward networks; Cambridge University Engineering Department (CUED, UK) using recurrent neural networks for hybrids; Instituto de Engenharia de Sistemas e Computadores (INESC, Portugal) for multi-layer perceptron (MLP) training and speaker adaptation. In addition, ICSI provides the computing environment necessary for the computationally expensive research.

1.1. Hybrid HMM-ANN Speech Recognition

The hybrid HMM/connectionist approach (first proposed by Bourlard and Wellekens [1]) combines the temporal modelling structure of the HMMs with the pattern classification capabilities of artificial neural networks. As in HMMs, a Markov process is used to model the basic temporal nature of the speech signal. This provides the structure for specification of a language model and incorporates constraints on the duration of the modelled words. The connectionist structure is used to model the local (in time) acoustic signal conditioned on the Markov process. This makes use of the result that connectionist networks satisfying certain regularity conditions provide class probability estimates for given input patterns [2, 1]. The advantages of the hybrid approach are numerous:

- discriminative training is straightforward for connectionist architectures
- phone models are combined resulting in efficient usage of parameters
- local acoustic correlation is explicitly modelled.
- correlations (even high order) between different features can be exploited without severe distributional assumptions
- connectionist models are highly parallel structures which lead to efficient hardware implementation.

There are two basic connectionist architectures currently employed in the WERNICKE hybrid systems. The first is the investigation of the MLP as a phone probability estimator [3, 4]. This structure employs the MLP as a static pattern classifier where temporal acoustic context is modelled via a multiple frame input layer. The second architecture under evaluation as a phone probability estimator is a recurrent neural network (RNN). This CUED developed system models acoustic context via a fully recurrent set of hidden *state* nodes [5]. Both systems achieve performance results comparable with other state-of-the-art speech recognition systems [6].

1.2. Project Objectives

The WERNICKE project is addressing the problem of large vocabulary speech recognition with HMM-ANN systems. This is a difficult problem requiring advances in speech science, developments in statistical modelling, and the integration of connectionist models within the statistical framework. Since the integration of feedforward and recurrent MLPs have already been shown to be quite successful, the main goal of this Esprit project is to develop and test these approaches on large scale problems with systematic comparisons and assessment with state-of-the-art systems. In this framework, this project aims at:

- extending these approaches to larger tasks and evaluating techniques to handle the resulting problems of scale
- investigating further theoretical and experimental issues related to performance improvement (e.g., context modelling, speaker adaptation, etc.)
- incorporating proven speech processing techniques developed for other systems
- porting of all the (training and test) algorithms being used to fast neural network hardware (RAP) developed at ICSI and, eventually, to a PC-based system
- defining a standard recognizer, Y0 (pronounced "why nought"), that will be used to compare the different hybrids.

1.3. Computing

The HMM-ANN approach is very computationally expensive. The networks currently require about 10^{13} floating point operations to train and future estimates of the required compute power is one or two orders of magnitude greater. Training these systems on standard workstations is very impractical and nearly impossible. To address this issue, the project partners have each acquired the Ring Array Processor (RAP) developed at ICSI [7]. The RAP provides 0.5 GigaFLOPS of processing power and has been designed specifically for connectionist processing.

2. PROGRESS

The major effort of the first six months of the project has been the development of baseline systems. These state-of-the-art systems provide the partners with a reference point from which to evaluate the effectiveness of their research. Each partner has implemented a hybrid HMM/MLP system which runs on the RAP. There are slight variations between the baseline systems at the different sites (see below), but all have been developed with common software.

Each partner has developed its own training system that has been evaluated on the baseline recognizer to insure that they all lead to (approximately) the same performance. Evaluations on the different sites' systems should lead to a *standardized* baseline system by the completion of the first year of the project. Preliminary results for the different baseline systems evaluated on the ARPA resource management (RM) task [8] are shown in Table 1. The MLP has a seven frame input, 1000 hidden units (250,000 parameters) and the input is augmented with difference coefficients. The RNN has 220 state units (70,000 parameters). The Feb89 test set was used to set the word transition probability for the MLP recogniser and was used as a cross validation set for the RNN recogniser. The front end options are:

MEL+ a 20 channel mel-scaled filter bank with voicing features

PLP 12th order perceptual linear prediction

MFCC 12th order mel-frequency cepstral coefficients

Where the last two figures are the frame rate and width of the Hamming window in milliseconds. The remaining portion of this section describes the baseline-system development research performed to date.

Net	Pre-Processing	Error Rate %			
		feb89	oct89	feb91	sep92
RNN	MEL+ 16/32	4.8%	6.1%	5.4%	10.7%
RNN	MFCC 10/20	6.1%	7.6%	7.4%	12.1%
RNN	MFCC 16/32	5.9%	6.3%	6.1%	11.5%
MLP	PLP 10/20	5.1%	5.7%	5.8%	12.2%
MLP	PLP 16/32	5.6%	6.7%	6.1%	12.9%
MLP	MFCC 10/20	5.7%	7.1%	7.6%	12.0%
MLP	MFCC 16/32	6.6%	7.8%	8.5%	15.0%

Table 1. HMM-ANN baseline system summary and performance.

2.1. Common phone set and pronunciation dictionary

For evaluations on the RM task, the original RNN system used a 49 phone set and the pronunciations used in the SPHINX system [9]. For the same task, the feedforward system used the phone set and most likely pronunciations developed at SRI and included on the RM CDROM database. Comparison and analysis of the different phone sets and pronunciations used in both systems found a reduction in the word error rate of 10% for the SRI-developed phone/pronunciation sets. All partners now use these sets for their baseline systems.

2.2. Common Viterbi decoder

As part of this collaboration, the partners have developed a decoding system referred to as Y0. The current version of Y0 accepts vectors of local distances which are generated as the negative log of the output of an MLP or recurrent network running on the RAP machine. Given these vectors, Y0 runs the dynamic programming algorithm producing a recognized string of words. Y0 consists of approximately 5000 lines of C++ code.

This collaboration has resulted in significant improvements to our recognizer. Some features of Y0 include - the ability to use multiple pronunciation word models, optional silence states at the ends of words, improved pruning strategies, and the ability to do a forced Viterbi alignment using local distances from a connectionist probability estimator. This decoder works with all the systems at the different sites.

2.3. Probability smoothing

When MLPs are trained according to LMS or entropy criteria, it can be shown that large values of probabilities will be better estimated than small values. As a consequence, we are beginning to investigate smoothing techniques combining probabilities with those from other estimators with better properties for the small values (e.g., a single Gaussian). It has been shown on a database other than RM that this could lead to some significant improvement of the recognition performance at the word level.

2.4. Connectionist approach comparison

The common recognizer Y0 gives us the ability to compare several connectionist approaches for probability estimation when the other "variables" are the same. Our goal is to start with the same features, and the same initialization for the targets, and ideally the same number of free parameters for the different paradigms and compare the performance on the same test set.

More specifically our current focus is to compare the RNN with the MLP approach on the RM test set.

The comparison of the feedforward and recurrent model is rendered difficult because the two schemes currently operate on different parameterisations of the acoustic data computed on different time scales. Results to date show the feedforward network performs best with a preprocessor based on a 10ms frame rate Perceptual Linear Predictor (PLP) analysis [10], whilst the RNN preprocessor performs best for a 16ms Mel scaled spectral representation. Work is underway to quantify and unite these approaches so that a common pre-processing can be used for the evaluation of the two methods. To date we have evaluated both models on the commonly used mel scaled cepstral coefficient representation at both frame rates (see table 1).

3. FUTURE WORK

The work completed in the first year of the project provides all the members with a state-of-the-art baseline system. From this vantage point, it is now possible to extend the systems to new areas of research. Several important questions still need to be addressed in the hybrid HMM-ANN framework. These include the issue of context dependent phone models, the development of better training procedures, the investigation of fast speaker adaptation in hybrid systems, the issue of the number of parameters and generalisation, and a better theoretical understanding of the hidden units activations and weights in ANNs.

3.1. Context dependent phone models

State-of-the-art HMM-based recognizers now use context-dependent phonetic units to improve their performance. Thanks to the advantages of the hybrid HMM-ANN approaches, it has been shown recently that context-independent, single-state phonemic HMM-ANN approaches performed nearly as well as context-dependent, multi-state HMM system. However, it is still not clear what additional improvement one can expect from context-dependent hybrid systems. Recently, a solution to the problem of applying MLP techniques to context-dependent HMMs was presented in [11]. This approach will be tested in the framework of this project.

3.2. Fast speaker adaptation

Fast speaker adaptation (i.e. adaptation involving much less training data and time than those used in the initial speaker-independent training) has shown to be an effective way to improve recognition performance in classical HMM-based recognizers. To our knowledge, however, this issue has not yet been addressed in the context of hybrid HMM-ANN recognizers.

In a first approach we are using a single additional layer, at the output of the speaker-independent MLP (SI-MLP). This layer is trained to map the probability estimates supplied by the SI-MLP to estimates that are more appropriate to the current speaker. A second approach will consist of actually re-training the SI-MLP on the current speaker's data, to obtain a speaker-specific MLP. The issues of *off-line* adaptation (i.e., adaptation using some fixed text pronounced by the speaker at enrollment time) and *on-line* adaptation (i.e. adaptation while useful recognition is also taking place) will be addressed.

3.3. Regularised training

We currently avoid overtraining the MLPs used to estimate state output probabilities using a cross-validation technique. This involves holding out a portion of the training data as a validation set. The performance of the network on this set is used to determine when training should be halted. Stopping early in this manner is an *ad hoc* form of regularisation, in which there is a force towards an initial (random) weight matrix. We are investigating the use of an explicit parameterised prior on the weight matrix, and the use of an objective Bayesian procedure to set the regularisation parameters (hyperparameters). The regularisation terms we have been experimenting with for sets of weights have been Gaussians and mixtures of Gaussians. By using different hyperparameters for different sets of weights we are able to allow the data to specify which weights should be most strongly determined by the data, and which are less relevant to the problem.

3.4. Extension to a larger database

The limitations with the RM task has led the project to undertake the Wall Street Journal (WSJ) task. This change in databases is desirable for a number of reasons:

- research into multiple pronunciation models will require a substantially greater amount of training data than is available with the RM task
- the WSJ task – with substantially more variation in acoustic context – will show whether the implicitly modelled acoustic context of the HMM-ANN approach scales to a larger database.
- evaluation on the WSJ task allows for direct comparison of the hybrid approaches with other state-of-the-art systems.

3.5. Faster training algorithms

The size of the speech database appears to be a significant factor in the performance of speech recognition systems. Connectionist systems are currently at a disadvantage in that they take about an order of magnitude more computation to train than pure HMM based systems. Thus faster training procedures would allow connectionist systems to be applied to larger databases and receive better acceptance in the speech recognition community.

3.6. Split nets

In view of the WSJ task, work has started on splitting very large databases among separate (and smaller) networks that can be trained independently and quickly. The networks are recombined (properly) during recognition to generate the required probabilities. After some theoretical work, initial experiments on RM have shown that it was possible to split the training data across two smaller nets without any significant loss in recognition performance both at the frame and word levels.

3.7. Extension of architectures for improved context modelling

Comparisons of different preprocessing methodologies, input frame widths, and delays of targets for the RNN and MLP systems indicate that there may be an optimal context window which can be implicitly modelled with the context-independent

systems. The connectionist framework allows for the automatic learning of the appropriate context via incorporation of gamma filters [12] into the input layer. Experiments to both improve the overall recognition performance and provide insight into the nature of the context being modelled are planned for the next year.

3.8. Multiple Pronunciations

Work is underway to implement multiple pronunciations per word and first results based on TIMIT transcriptions show a small decrease in the error rate. We intend to expand the application domain of our technique and we are considering the ARPA Wall Street Journal and SAM databases for future development.

3.9. Alternative Approaches

In the framework of this project, alternative hybrid approaches will be investigated and tested on the same baseline system and databases. Among these we have:

Predictive networks There is a relationship between predictive networks (as proposed by [13] and nonlinear autoregressive modelling. While they are also very attractive from the theoretical point of view, they have their own weaknesses (e.g no discrimination, and very noisy estimate of probabilities) which make it difficult to actually get significant improvements out of them. A (theoretical) solution could be to merge both predictive and discriminant approaches.

Output feedback In the initial theory of the hybrid HMM/MLP approach [1] it was suggested to have contextual inputs but also feedback from the output units to the input layer (to model correlation at both the acoustic vector and HMM state level). This initial architecture can be implemented and leads to the investigation of approaches mixing RNN and acoustic input context.

3.10. Other work

Additionally, we are continuing to do work on acoustically robust features and on accent modelling.

4. CONCLUSION

The WERNICKE project is a substantial effort in research into the problem of large vocabulary speech recognition. The partners bring to the project a strong background both in the field of speech processing and connectionist methods. The HMM-ANN is an approach with a great deal of potential for the large vocabulary recognition problem. Work to date has shown that the approach is competitive with HMM-based systems, although there has been substantially less research in the HMM-ANN field. We expect the work of the WERNICKE project over the next few years to show a significant improvement in performance relative to the context-dependent, multiple mixture HMM approach.

5. ACKNOWLEDGEMENTS

WERNICKE is ESPRIT project (6487).

Two of the authors, AJR and SR are supported by SERC research fellowships. ICSI work is partially funded by an SRI subcontract from ARPA contract MDA904-90-C-5253.

REFERENCES

[1] Hervé Bourlard and Christian J. Wellekens. Links between Markov models and multilayer perceptrons. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(12):1167–1178, December 1990.

[2] Eric B. Baum and Frank Wilczek. Supervised learning of probability distributions by neural networks. In Dana Z. Anderson, editor, *Neural Information Processing Systems*. American Institute of Physics, 1988.

[3] N. Morgan, H. Bourlard, S. Renals, M. Cohen, and H. Franco. Hybrid neural network/hidden Markov model systems for continuous speech recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 1993. In press.

[4] Hervé Bourlard and Nelson Morgan. *Continuous Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, 1993. In Press.

[5] Tony Robinson. A real-time recurrent error propagation network word recognition system. In *Proc. ICASSP*, volume I, pages 617–620, 1992.

[6] Proceedings of the DARPA artificial neural network speech technology meeting, January 1993.

[7] N. Morgan, J. Beck, P. Kohn, J. Bilmes, E. Allman, and J. Beer. The Ring Array Processor (RAP): A multiprocessing peripheral for connectionist applications. *Journal of Parallel and Distributed Computing*, 14:248–259, 1992.

[8] Patti Price, William M. Fisher, Jared Bernstein, and David S. Pallett. The DARPA 1000-word Resource Management database for continuous speech recognition. In *Proc. ICASSP*, pages 651–654, 1988.

[9] Kai-Fu Lee. *Automatic Speech Recognition: The Development of the SPHINX System*. Kluwer Academic Publishers, Boston, 1989.

[10] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87:1738–1752, 1990.

[11] Hervé Bourlard, Nelson Morgan, Chuck Wooters, and Steve Renals. CDNN: A context dependent neural network for continuous speech recognition. In *Proc. ICASSP*, volume II, pages 349–352, 1992.

[12] J. C. Principe, B. de Vries, and P. G. de Olivieira. The Gamma filter – A new class of adaptive IIR filters with restricted feedback. *IEEE Transactions on Signal Processing*, 41(2):649–656, February 1993.

[13] Esther Levin. Hidden control neural architecture modeling of nonlinear time varying systems and its applications. *IEEE Transactions on Neural Networks*, 4(1):109–116, 1993.