INITIAL SPEECH RECOGNITION RESULTS USING THE MULTINET ARCHITECTURE

E.B. Pizzolato & T.J. Reynolds

Department of Computer Science
University of Essex
Colchester, CO4 3SQ, UK
Email : {ebpizz , reynt} @essex.ac.uk

ABSTRACT

Multinet is a connectionist architecture designed for certain difficult multi-class pattern classification tasks. These are characterised by very large input feature spaces, rendering a monolithic classifier impractical. The architecture consists of a layer with at least one primary 'detector' for each class, followed by a combining net which estimates the posterior probabilities for all classes. Typically primary detectors only input a subset of the input features. Thus the architecture decomposes classification in two ways: by class and by factoring of the input space dimensions. Multinet incorporates the ideas of Modular Neural Networks and Ensembles. In this paper we investigate the use of Multinet on standard speech recognition tasks and present results for phoneme recognition on TIMIT and word recognition on RM. We show Multinet's performance is comparable with standard HMM and hybrid HMM-NN systems that we run on the same tasks. The value and potential of the Multinet approach is shown by detailing successive improvements to the Multinet system which are easily obtained because of the modularity of the architecture.

1. INTRODUCTION

Current commercial ASR systems use Hidden Markov Models (HMMs). They are a powerful framework for speech recognition but depend on strong assumptions about the speech process [1]. Hybrid HMM-NN systems make weaker assumptions and combine the temporal modelling of HMMs with the use of Neural Networks as probability estimators [2]. For example, Renals et al. [2] used an MLP probability estimator with 234 input nodes, 1000 hidden nodes and 69 outputs. Training monolithic NNs of this size can require several days. This can severely hamper the research and development of such systems, and has implications for the practicality of the approach in, for example, speaker adaptation.

Several architectures have been proposed to reduce network training times (as well as to improve classification). Neural net ensembles [3,4], for example, have been successfully applied to a number of classification problems, including speech recognition [5,6]. They are a combination of different classifiers applied to the same task. Based on the principle of divide-and-conquer, Modular Neural Networks (MNNs) [7,8,9] on the other hand, aim at combining the results of different classifiers applied to different sub-tasks. As well as speeding up training these architectures can also (1) simplify the modelling process, (2) produce solutions that would have been impractical or infeasible with a monolithic net and (3) produce better results by exploiting different properties of the training data. Sub-tasks can often be solved in a simpler fashion with less resources. Therefore, (1) and (2) are direct consequences of the application of either MNNs or ensembles. The exploitation of the different properties of the training data is a particular characteristic of the ensemble solution. Classifiers are applied to different parts of the *same* problem and a solution is derived from the combination of the individual responses of each classifier.

The Multinet architecture [10,11] is a layer-structured neural net architecture which can function as an ensemble or MNN. The major novelty of this architecture is that the classification task is redefined as a result of the modularisation: separate networks can use different input feature vectors as appropriate to their classification sub-task. The main focus of this paper is to present results for the Multinet architecture functioning as an MNN. We compare phoneme and word recognition results on the TIMIT and RM speech corpora against those produced by a standard HMM and a standard HMM-NN hybrid system.

2. THE MULTINET ARCHITECTURE

A direct modularisation of the standard HMM-NN hybrid system would consist of a layer of small networks, each approximating the posterior probability of its class, on the same input space. However we would like to go further and specialise each net to its individual task, including supplying it with tailored input data. This generalisation of the modularisation concept requires us to add another layer to the simple architecture. This is because the individual detectors are no longer estimators in the same space. Their outputs must be combined in order to produce posterior probabilities on the total input space. The resulting architecture is shown in figure 1.

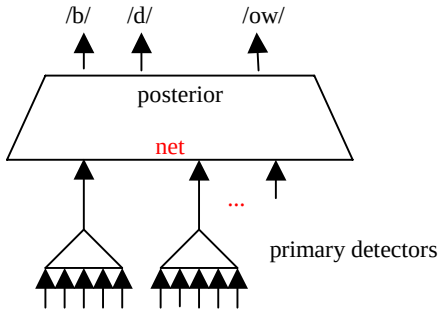

Figure 1: The Multinet Architecture.

2.1 The Primary Detectors

Primary detectors may be trained one-at-a-time to discriminate their class from all other classes. Unlike the monolithic net, training a primary detector is a simple task as each detector just needs to learn the distribution of its class against the background. Its resources are dedicated and can be precisely tuned to need. We avoid awkward problems of cross-talk and resource hogging which can occur in a multiple-output classifier. For certain detectors we may also reduce the number of inputs, as sounds that are essentially stationary can be identified with less feature vectors. As a result of these simplifications, the total architecture requires fewer parameters and trains faster. Further simplification of the modelling process can be achieved by excluding some of the out-class data when training particular primary detectors.

2.2 The Posterior Net

The posterior net has as many outputs as phone classes. It is connected to all the outputs of the primary detectors and trained in a separate exercise with the primary detectors held fixed. Its task is to estimate the posterior probabilities of all phones on the complete input feature space and can be viewed as a Supra-Bayesian [12] net which treats the opinions of other classifiers as data. The task of the posterior net is closely related to the way the primary detectors were trained. If all primary detectors share the same input space, then they approximate the actual posterior probability for each phoneme and the posterior net has no role. However, if the primary detectors are trained on different subsets of the input data, or with different pre-processing techniques then the role of the posterior net becomes more important, because the individual probability estimations are not the actual posterior probabilities. An interesting case is the training of the primary detectors for each vowel. Vowels are almost steady sounds and do not need extra information about time development. However they are distinguished from the diphthongs exactly because of this information. In other words, they can be distinguished in the higher-dimensional space of multiple frames and with delta information, but in the space of reduced dimensions, diphthongs project confusingly on top of them. By taking the diphthongs out of the training set of the vowels, we create a representation for the vowels in this space of reduced dimensions, but we have also changed the balance of the original space. The posterior net corrects the balance by appropriate positive and negative scaling of the primary detectors. So, for example, the positive indication of a diphthong inhibits the response of the posterior net for a vowel.

The posterior net also corrects any distortions in the priors which may have been introduced by biased selections of data for training primary detectors. In [10] we have shown that the posterior net can successfully combine the estimations of primary detectors applied to different analysis time-scales and on biased data. Provided the primary detectors do not actually *discard* any information from the total input feature space, then, given sufficient resources, the posterior net can deliver good posterior probability estimations.

3. PHONEME RECOGNITION

We have evaluated the Multinet architecture on both TIMIT and RM speech corpora. We collapse the 61 TIMIT phone labels into 42. These differ only slightly from the standard 39 phone set [13] as /zh/ is not collapsed into /sh/ and /aa/ is also separated from /ao/. Our front-end analysis extracts 16 mel-frequency cepstral coefficients (MFCCs), including energy, plus 16 ΔMFCCs from each 16 ms of speech, at an 8ms frame rate.

The baseline Multinet system has 42 primary detectors, one for each phoneme. The detectors are all MLPs with 20 hidden nodes, except for the diphthong detectors where 50 hidden nodes are used. The choices of input data were based on a rough judgement of the different requirements of the phone classes. In order to minimise the inputs and hence the size of individual detectors, we avoid, as much as possible, the use of many input frames and the delta information. The posterior net is an MLP with 42 inputs and outputs and 100 hidden nodes.

Two other systems are also evaluated for comparison:

(1) A standard context-independent continuous-density HMM system, with three-state left to right models and one skip transition. The plosives and affricatives, however, are modelled with two states. For each state, 8 mixtures of gaussians are computed in order to model the emission likelihoods for the 32 coefficients (16 MFCCs + 16 ΔMFCCs).

(2) A hybrid HMM-NN system where the NN architecture is an MLP with 288 inputs (9 frames of 32 coefficients each), 1000 hidden nodes and 42 outputs.

In order to observe the behaviour of the recognisers for each class, we performed phoneme recognition against the TIMIT mark-up. Table 1 shows these results along with results for the Sphinx system [14] on the same task.

SYSTEM	% correct
Sphinx HMM	58.7
HMM(baseline)	59.7
HMM-NN(baseline)	68.2
MULTINET(baseline)	61.7
MULTINET (improved)	62.8

Table 1: Phoneme recognition on TIMIT mark-up.

Although the overall phoneme recognition performance is comparable to the HMM system, we showed in [10] that the Multinet underperforms on nasals, semi-vowels and diphthongs. It compensates by being better at plosives and silence. This is an indication that some of our original decisions on the number of input frames to present to certain primary detectors, were not appropriate. Fortunately, one can retrain some of the primary detectors very quickly and re-apply them even to the same posterior net. Changing the number of input frames for the nasals and including deltas for the semi-vowels improved the phoneme recognition. No doubt further improvements would be possible.

4. EXPERIMENTS ON RM

Although phoneme recognition is a good measure of the quality of a system, a full evaluation comes with word recognition results. In order to perform a fair comparison with the two other methods (HMMs and HMM-NN) we built our own decoder so that the three systems would have the same characteristics. The decoder performs a standard Viterbi beam search over the space of likelihoods (HMMs) or scaled likelihoods (hybrid and Multinet systems) and takes into account a few phonological rules (for instance, geminate deletion). It also presents some flexibility at the lexical level where alternative or optional paths are allowed.

Our context-independent continuous-density HMM system has the same structure as applied to phoneme recognition on TIMIT. It is of special importance as it produces the alignment used to train both the standard HMM-NN hybrid system and the Multinet architecture. It is also our reference system and for this reason was compared to a well-known context-independent system: the Sphinx system [15]. However the comparison is complex: Sphinx is a discrete density system which employs 7 states per phoneme. It is also presented in several versions. Table 2 shows the results of word recognition applied on the Sphinx88 test-set of the RM corpus, using a word pair grammar, for the HMM systems. Sphinx3C is an attempt to improve the discrete density model by applying more than one codebook to the modelling process and therefore is closer to the continuous-density model we have employed. However, we haven't devoted much effort to improving our HMM baseline system as our major goal is to provide equal conditions of comparison for the three different methods.

SYSTEM	WORD ACCURACY
Sphinx (baseline)	58.1%
HMM(baseline)	**70.7%**
Sphinx 1C	76.1%
Sphinx 3C	81.1%

Table 2: Comparison of HMM baseline system and the Sphinx versions. 1C and 3C stand for one and three codebooks respectively.

The HMM-NN hybrid system also has the same structure as used on the phoneme recognition experiment and delivers a word accuracy of 68.8%. It is important to notice, however, that the learning process was interrupted after the 8th epoch due to the overall long training time. We believe that a better recognition performance could have been achieved if the training had gone further.

Four versions of the Multinet system are analysed. Table 3 presents the number of input frames for each phoneme in each broad class across the versions. The number of hidden nodes is set to 20 when no delta information is supplied to the classifier and 40 or 50 otherwise. The diphthongs are modelled the same way across all versions as their nature demands an input representation that takes time development into account. Also silence is modelled the same way across the versions with 7 frames and no delta information.

Apart from Version 3, the other two versions exclude the diphthongs when training each vowel detector. Therefore a posterior net with 100 hidden nodes is employed in order to provide a final posterior probability estimation for each phoneme.

CLASS	Baseline	V1	V2	V3
Plosive	7	7+Δ	7+Δ	7+Δ
Fricative	7	7	7+Δ	7+Δ
Vowel	7	7	7	7+Δ
Nasal	7	7	7+Δ	7+Δ
Diphthong	9+Δ	9+Δ	9+Δ	9+Δ
Semi-Vowel	7	7+Δ	7+Δ	7+Δ

Table 3: Primary detector inputs by broad class

Version 1 is intended to show the benefit of using additional delta representation on plosives and semi-vowels. Version 2 is intended to show the role of the posterior net as the primary detectors for vowels are not trained against the diphthongs. Its result should be very close to the one produced by Version 3. Finally, Version 3 is included to show the benefit of employing delta information on every phoneme. As all detectors share almost the same input space in this version, no posterior net is employed. Table 4 presents the progress of the word accuracy rates across the Multinet versions and the results of the HMM and HMM-NN methods. No assessment is carried out on individual phoneme resources such as number of hidden nodes. Rough estimations indicated that the numbers previously stated (20 hidden nodes for input space without delta information and 40-50 otherwise) are a reasonable trade-off between training time and classification capability. Also, the resources of the posterior net are limited to 100 hidden nodes and further analysis is also necessary to assess the influence of this parameter on the overall word accuracy.

SYSTEM	WORD ACCURACY
MULTINET Baseline	55.3%
MULTINET V1	58.8%
MULTINET V2	63.0%
MULTINET V3	65.6%
STANDARD HMM-NN	68.8%
STANDARD HMM	70.7%

Table 4: Word accuracies for the systems

Version 1 presents an improvement over the baseline system, however, the word accuracy of 58.8% is only slightly better than the baseline 55.3% rate. Including the delta information on all other phonemes apart from the vowels, improves the word accuracy in version 2 to 63.0%. This result is close to the word accuracy of the final version (65.6%) but not yet close to our baseline HMM and HMM-NN results. The difference in word accuracy between the final version of the Multinet and Version 2 may be due to insufficient free parameters in the posterior net to model each class. Further investigations will be carried out in order to determine the appropriate number of hidden nodes of the posterior net. The difference between the final version of the Multinet system and the two other baseline methods may also be explained by the lack of resources provided to each individual primary detector.

5. CONCLUSION

Multinet is a new hybrid HMM-NN architecture that can function as a Modular Neural Network (MNN) or Ensemble. We have been investigating Multinet as a MNN and have illustrated how this architecture can help us to attack performance deficiencies in a piecemeal fashion. We have yet to really exploit the possibilities inherent in the individual design of primary detectors. This is an extra flexibility that the architecture provides, but which necessarily requires more development effort. This comes with the power to decide the resources as well as pre-processing on an individual phoneme basis. This contrasts with the standard HMM-NN hybrid approach, where only the total resources are set a priori and there is no control over the allocation of resources to individual classes. The allocation is implicitly performed during training and is not necessarily optimal. In particular infrequent classes may well be starved of net resources.

A monolithic phone-classifier MLP can take up to 336 hours to train (24 hours per iteration). Any re-design requires complete retraining. A complete Multinet architecture takes comparable or less time to train on one computer, according to the detectors used. Retraining one detector, however, takes less than 9 hours. We are also able to train individual detectors on separate computers. With enough computers we could train all the primary detectors in 9 hours by exploiting this training parallelism. Training the posterior net adds about 10 hours to this time. At the moment we can use 3 PCs for training, and thus are able to divide our training time by 3.

6. ACKNOWLEDGMENT

One author (E. B. Pizzolato) gratefully acknowledges the support of the Brazilian Agency CAPES (Grant No. 1125/94-8), and the Department of Computer Science, Universidade Federal de Sao Carlos, Brazil.

7. REFERENCES

1. Morgan N. & Boulard H. “Continuous Speech Recognition – An Introduction to the Hybrid HMM/Connectionist Approach”. *IEEE Signal Processing Magazine*, pp. 25-42, May 1995.

2. Renals S., Morgan N., Bourlard H. et al. “Connectionist Probability Estimators in HMM Speech Recognition”. *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 161-173, 1994.

3. Hansen L.K. & Salomon P. “Neural Network Ensembles”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 993-1001, 1990.

4. Krogh A. & Vedelsby J. “Neural Network Ensembles, Cross Validation and Active Learning” *in Advances in Neural Information Processing Systems*, vol. 7, MIT Press, 1995.

5. Jordan M.I. & Jacobs R.A. “Hierarchical Mixtures of Experts and the EM Algorithm”. *Neural Computation*, vol. 6, pp. 181-214, 1994.

6. Fritsch J. “ACID/HNN – Clustering Hierarchies of Neural Networks for Context-Dependent Connectionist Acoustic Modelling”. *Proceedings of ICASSP*, 1998.

7. Waibel A., Sawai H. and Shikano K. “Modularity and Scaling in Large Phonemic Neural Networks”, *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 12, pp. 1888-1898, 1989.

8. Ananda R., Mehotra K., Mohan C.K. & Ranka S. “Efficient Classification for Multiclass Problem using Modular Neural Networks. *IEEE Transactions on Neural Networks*, vol. 6, pp. 117-124, 1995.

9. Auda G. & Kamel M. “Modular Neural Network Classifiers: a Comparative Study”, *Proceedings of ICASSP*, 1998.

10. Reynolds T.J. & Pizzolato E.B “Phoneme Classification with Multinets”, *Proceedings of the International Conference on Signal Processing*, Beijing, China, 1998.

11. Reynolds T.J., Pizzolato E.B. & Antoniou C. “Multinet: a New Connectionist Architecture for Speech Recognition”, *Proceedings of the International Conference on Artificial Neural Networks*, Skovde, Sweden, 1998.

12. Jacobs, R.A. “Methods for Combining Experts Probability Assessments”, *Neural Computation*, vol. 7, pp 867-888, 1995.

13. Robinson A.J. & Fallside F. “Phoneme Recognition from the TIMIT Database using Recurrent Error Propagation Networks”, *Technical Report CUED/F-INFENG/TR-42*, Cambridge, UK, march 1990.

14. Lee K.F & Hon H.W. “Speaker-Independent Phone Recognition using Hidden Markov Models”. *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 1641-1648, 1989.

15. Lee K.F “Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System”, *PhD Thesis*, Carnigie Mellon University, 1988.