A Comparative Study of OCON and MLP Architectures for Phoneme Recognition.

S.J.Haskey and S.Datta

Electronic and Electrical Engineering Department
Loughborough University, Loughborough, LE11 3TU, UK.

ABSTRACT

In this paper a comparative study between One-Class-One-Network (OCON) and Multi-Layered Perceptron (MLP) neural networks for vowel phoneme recognition is presented. The OCON architecture, first proposed by I.C.Jou et al 1991, is similar in design to a conventional feed-forward MLP, only each class had its own dedicated sub-network containing a single output node. Conventional MLPs usually consist of fully-connected nodes which not only result in a large number of weighted connections but also create the problem of cross-class interference. Using vowel phoneme data from the DARPA TIMIT corpus of read speech, MLP and OCON architectures were trained and the relative effects of recognition and convergence rates during both intra and inter-class adaptation tested. The OCON showed an increase in the convergence rate of 273% and an improvement of adapted recognition rates against the MLP of over 12%. However, due to the isolated nature of each OCON class, it was unable to utilise inter-class information. This resulted in a recognition rate reduction of over 6% for unadapted phonemes during adaptation of remaining vowels, compared with the MLP results.

1. THE OCON

A large fully-connected network can potentially contain many hundreds of neurons, each connected via weights to many others. This can make the training and adapting of such a network a long and difficult task. In addition, fully connected networks are prone to cross-class interference. Cross-class interference occurs when adapting towards a single class in a multi-class network, inevitably altering shared weights. As the network gets larger the interference increases, drastically degrading the convergence rate of the shared weights due to the influence of conflicting signals. This can lead to, after adaptation towards a single class, the impaired classification for the remaining classes within the network. To eliminate these problems, I.C.Jou et al [2] proposed a new neural network architecture called the One-Net-One-Class. The same principle was later taken on by S.Y.Kung[3][4], who named the architecture the 'One-Class-One-Net' or the 'OCON' for short. The OCON is similar in design to that of a conventional MLP (see Figure 1a) only each class has its own dedicated subnet containing a single output neuron (see Figure 1b). Each OCON subnet is specialised for distinguishing its own class from other patterns, resulting in fewer nodes being required in the hidden layers for each class. I.C.Jou first used the OCON architecture in 1991 for optical character recognition (OCR). Later S.Y.Kung [4] also applied the OCON architecture to OCR, achieving a training accuracy of 99.5% compared with 94% from a conventional MLP. Such architectures have also been used for texture classification, Electrocardiograph (ECG) analysis and the classification of mandarin speech syllables and isolated English words with a hybrid Time Delay Neural Networks (TDNN) and OCON structure [5].

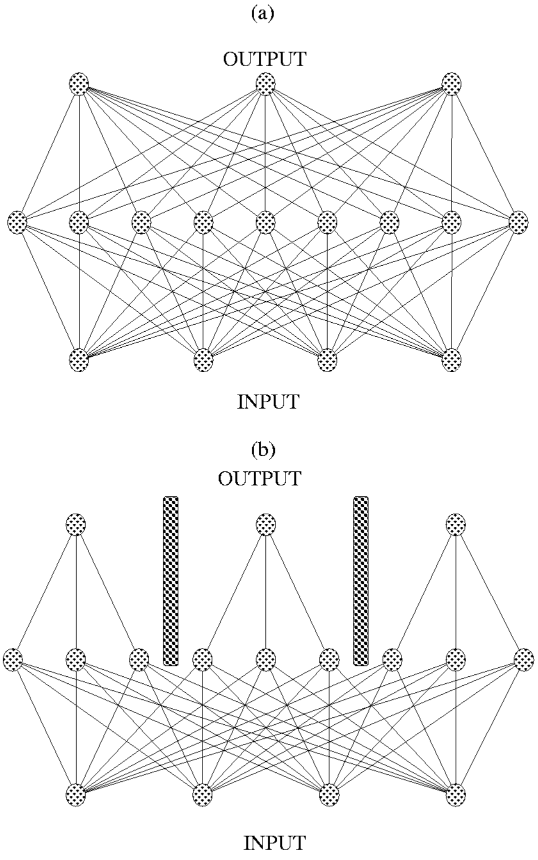

Figure 1: (a) A fully-connected MLP architecture.
(b) An OCON Neural Network Architecture

2. THE SPEECH DATA

All the speech data used during the comparative study was obtained from the DARPA TIMIT corpus of read speech [1]. 12 vowel phonemes spoken by male speakers from the TIMIT dialect region 7, the western geographical area of the U.S, were used for training and testing the ANN architectures. Vowel phonemes were specifically chosen since they are the most spectrally well defined of all phonemes making them more easily and reliably recognised and ideal for a comparative study. In addition, to avoid large deviations between phonemes during the comparative study, phonemes from speakers with the same gender and dialect were selected. Male speakers from dialect region 7 were selected because of the availability and good representation of training and testing data available from this group. However, of the 13 vowel phonemes available, using the ARPABET representation [6], vowel /UW/ was not used due to the limited number of utterances leaving the 12 vowel phonemes, /IY/, /IH/, /EY/, /EH/, /AE/, /ER/, /AX/, /AH/, /UH/, /OW/, /AO/, /AA/. During the experimentation it was not only of interest to test the effect of recognition rates and convergence on the adapted vowels but also the effect the adaptation had on the remaining unadapted vowels. Unfortunately, testing the effects of inter-class adaptation on 12 vowel phonemes is a very labour intensive procedure and so the phoneme groups were reduced further. They were split into 3 distinct groups with respect to the tongue-hump position in the oral cavity during their production, 'front', 'middle', and 'back'. They were grouped in this way since phonemes from the same tongue-hump group show some acoustic similarities [7]. The front vowel phonemes were /IY/, /IH/, /EY/, /EH/ and /AE/, the middle vowel phonemes were /ER/, /AX/ and /AH/, and the back vowel phonemes were /UW/, /UH/, /OW/, /AO/ and /AA/. Using 'Speech Tools' [8] the relevant phoneme data was extracted from the recorded 16kHz speech files within the TIMIT corpus. Each phoneme file was pre-emphasised, to compensate for the -6db/octave roll-off of voiced speech and windowed using 8 over-lapping hamming windows, each representing 16ms of speech. The speech data in each window was used to generate 12 linear predictive coefficients (LPCs) which were normalised by dividing by the first. The first coefficient could therefore be eliminated since it was always equal to one. This left 11 LPCs for each window resulting in a total of 8x11=88 coefficients representing each vowel phoneme. Linear prediction with its simple coding and well documented behaviour was specifically chosen as the most appropriate form of speech pre-processing since all experimentation was primarily concerned with the performance of the ANN architectures.

3. ANN ARCHITECTURES

To test the performance of the OCON architecture on the vowel phoneme speech data, a comparative study with the more conventional MLP was set. The OCON and MLP architectures were represented by three networks each, corresponding to the 'front', 'middle' and 'back' tongue-hump groups of the speech data. For each phoneme group the MLP and OCON networks (see Figure 2) were modelled using the Stuttgart Neural Network Simulator (SNNS) [9]. All the networks contained the same number of input nodes, 88, dictated by the number of input coefficients representing each speech utterance. The total number of output nodes for each network was dependent on the phoneme group, five phoneme classes for the front and back and three phoneme classes for the middle. The six networks, with every node using the sigmoidal activation function, were modelled with fully connected adjoining layers, except for the hidden and output layers of the OCON architecture.

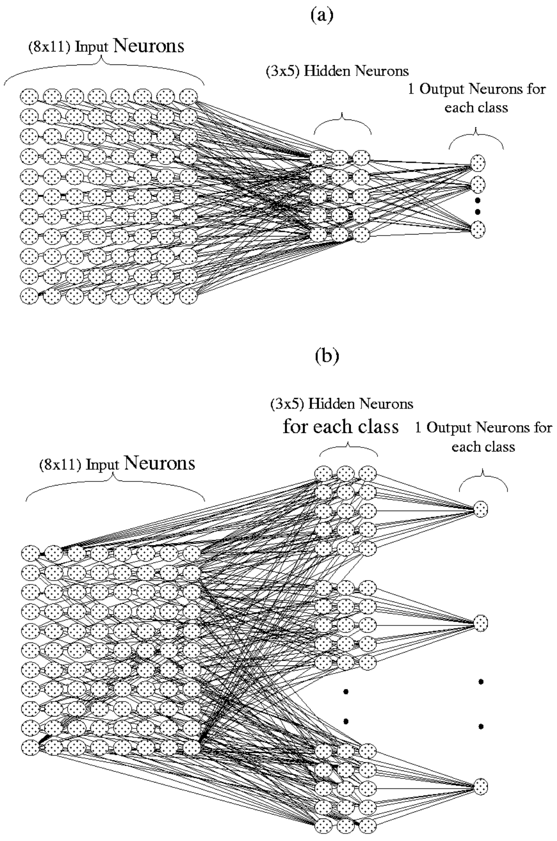
(a)
(8x11) Input Neurons
(3x5) Hidden Neurons
1 Output Neurons for each class
(b)
(3x5) Hidden Neurons for each class
1 Output Neurons for each class
(8x11) Input Neurons

Figure 2: (a) Fully connected MLP architecture.
(b) Fully connected OCON architecture.

Each network was trained with male TIMIT training set from dialect region 7. The weight and bias values within the networks were initially randomised and the standard back propagation algorithm used to train the networks, producing the six 'base-classifiers' necessary for the experimentation. The male TIMIT 'test set' for dialect region 7 consisted of 15 male speakers. Since there was only interest in intra-speaker effects and not inter-speaker effects, all the speech data from every test speaker was amalgamated and categorised with respect to its phoneme content. The networks were then ready for adaptation and testing, but before that could occur, a single common back-propagation learning-rate for both the MLP and OCON networks had to be found. This was achieved by training one of the MLP and OCON networks with various learning rates. A learning rate of 0.5 was selected since it offered both networks fast convergence without any instabilities.

Each of the six base-classifiers was adapted and tested using the 'test set.' Each network was adapted towards one of its relevant phoneme classes for a total of 100 cycles, during which 7 result snapshots were taken at 1, 3, 5, 10, 20, 50 and 100 cycles. Due to the non-linearity of network adaptation, the number of cycles between each result snapshot increased to produce a graph that offered a clear picture of the network's behaviour. The results taken at each snapshot were the recognition rates of both the adapted phonemes and the remaining unadapted phonemes within the same network. After adapting for 100 cycles towards each phoneme class, the weights and bias' within each network were reset to their initial base-classifier values ready for the next adaptation procedure involving another phoneme class.

4. RESULTS

Comparative results for the MLP and OCON architectures were obtained for adaptation towards each of vowel phoneme class and the effect on the remaining unadapted vowel phoneme classes within the same networks. 2 graphs were produced containing the averaged data from all the vowel phonemes for the adapted and unadapted phonemes recognition rates (see Figure 3(a)(b)). As well as recognition rates, another area of interest was each network's convergence rate. The convergence rate for each of the 2 averaged data graphs was calculated by differentiating the recognition-rate data (calculating the distance between adjacent rates). However calculating the convergence rate in this way was viewed as being unrealistic since the closer the recognition rates reach the perfect goal of 100%, the greater the significance of recognition improvement. To reflect this the convergence rate y was calculated using equation :

$$y = \left(\frac{100 - \chi_n}{100 - \chi_{n+1}} \right) - 1 \qquad (1)$$

where χ_n and χ_{n+1} are two adjacent recognition rates. The term -1 in equation 1 was used for normalise the graphs so that positive values indicated positive convergence and negative values negative convergence. The 2 convergence rates graphs were generated were for all the adapted vowel phonemes (see Figure 4(a)), and all the unadapted vowel phonemes (see Figure 4(b)). Figure 3(a) shows that the OCON networks show a clear improvement for the recognition rates of adapted vowel phonemes over the conventional MLP networks. On average, for all vowel phonemes, the experimentation shows a 12.3% increase in recognition rates for the OCON networks [10][11]. This result echoes the improvements shown in other data classification systems utilising OCON architectures [2][3][4][5]. Furthermore, the OCON architecture not only increases the adaptation rate but also reduces the processing time necessary for each adaptation cycle due to the reduction in network weights. This is shown in figure 4(a) with the increased rate of convergence for each OCON network, offering a 273% increase against the MLP for adapted phonemes. However, the OCON architectures as they stand, deal badly with inter-class adaptation. Although the rates of convergence for both networks are roughly the same, figure 4(b), figure 3 (b) shows that the OCON networks offer worse recognition rates for unadapted vowel phonemes over the conventional MLP networks. From figures 3 (b) we find that the average drop in recognition rates for the OCON networks, compared with the MLP networks, is 6.3%.

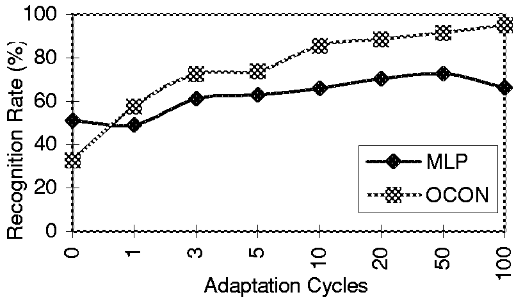

Figure 3(a): Average Recognition Rates for All Adapted Vowel Phonemes for an MLP and OCON Network

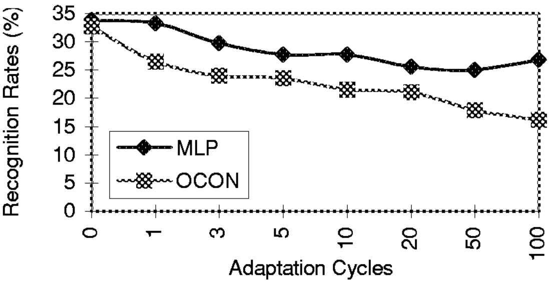

Figure 3(b): Average Recognition Rates for All Unadapted Vowel Phonemes for an MLP and OCON Network

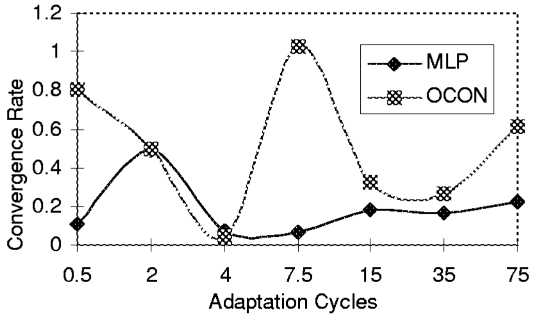

Figure 4(a): Average Convergence Rates for Adapted Vowel Phonemes for an MLP and OCON Network

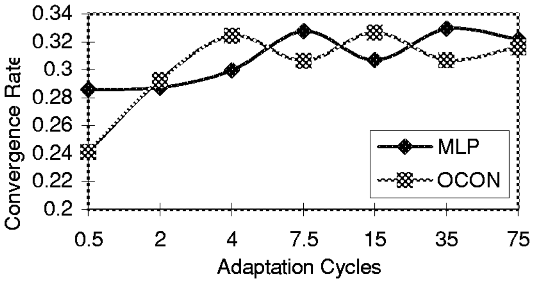

Figure 4(b): Average Convergence Rates for Unadapted Vowel Phonemes for an MLP and OCON Network

5. CONCLUSION

As expected the OCON behaves better than the MLP when adapting and testing the same phoneme. This is primarily due to the individual networks in each OCON network being dedicated to each class. Not only are there fewer connections and hence weighted axes to train, but each network only has to deal with information concerning a single class. As a result the OCON not only reduces the processing time for each adaptation cycle, but also rapidly increases the convergence rate. However, the OCON architecture as it stands, deals badly with inter-class adaptation. When adapting to a class, the OCON shows a lower recognition rate for the remaining phonemes in the network compared to that of the MLP. This indicates that there must exist some common speaker information within all the classes in a network which isn't being exploited in the isolated networks of the OCON. Although in many applications cross-class interference can be a problem, MLPs compared to OCONs appear to use it to their advantage for inter-class adaptation. As a result an ideal network would be a hybrid OCON architecture containing isolated networks for improved single class adaptation but with some inter-class bonding to profit from any common speaker information. However it would be important that any hybrid OCON network should concentrate adaptation only on common speaker information as adaptation towards common class information could result in harmful cross-class interference.

6. REFERENCES

[1] DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM, Oct 1990.

[2] Jou, I., Tsay, Y., and Tsay, S. "Parallel Distributed Processing with Multiple One-Output Back-Propagation Neural Networks," *Proc, Int Symp on Circuits and Systems, Singapore,* pp.1408-1411, 1991.

[3] Kung, S.Y., and Taur, J.S. "Decision-Based Neural Networks with Signal/Image Classification Applications," *IEEE Transactions on Neural Networks, Vol.6, No.1,* pp. 170-181, January 1995.

[4] Kung, S.Y. "Digital Neural Networks," Prentice Hall, Englewood Cliffs, NJ.

[5] Hwang, J.N., and Hang Li. "Interactive query learning for isolated speech recognition." In Kung, S.Y., Fallside, F., Sorensen, J.A., and Kamm, C.A. "Neural Networks for Signal Processing," I, pp.513-522, *Proceeding of the 1991 IEEE Workshop*, Princeton, NJ, 1991.

[6] Shoup, J.E. "Phonological Aspects of Speech Recognition," 125-138, Ch6 in Trends in Speech Recognition, W. A. Lea, Ed., Prentice-Hall, Englewood Cliffs, NJ, 1980.

[7] Rabiner. L., and Juang, B. "Fundamentals of Speech Recognition," 26-27, Ch2, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[8] Speech Tools User Manual, Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, August 1993.

[9] Stuttgart Neural Network Simulator, User Manual, Version 4.1, Institute for Parallel and Distributed High Performance Systems, University of Stuttgart, 1995.

[10] Haskey, S.J., and Datta, S. "Dynamic Speaker Adaptation for Acoustically Similar Vowel Sounds using Sub-Cluster Neural Network," *Conference and Workshop on New Ideas in Computing*, Part 2, Coventry University, May 1997, pp41-44.

[11] Haskey, S.J., and Datta, S. "Using Tongue-Hump-Position Information for Vowel Adaptation within a Subcluster Neural Network," *IEE Colloquium on Pattern Recognition*, London, 26 Feb 1997, pp9/1 – 9/6.