SAPERE • AUDE

University of Otago
Te Whare Wananga O Otago
Dunedin, New Zealand

Phoneme Recognition with Hierarchical Self Organised Neural Networks and Fuzzy Systems

Nikola K. Kasabov

E. Peev

The Information Science Discussion Paper Series

Number 94 / 4
February 1994
ISSN 1172-6024

Phoneme Recognition with Hierarchical Self Organised Neural Networks and Fuzzy Systems - A Case Study

Nikola K. Kasabov[1]
Department of Information Science
University of Otago

E. Peev
KZIIT Sofia
Bulgaria

February 1994

[1] Address correspondence to: Dr N.K. Kasabov, Senior Lecturer, Department of Information Science, University of Otago, P.O. Box 56, Dunedin, New Zealand. Fax: +64 3 479 8311 Email: nkasabov@commerce.otago.ac.nz

1 Introduction

Neural networks (NN) have been intensively used for speech processing (Morgan and Scofield, 1991). This paper describes a series of experiments on using a single Kohonen Self Organizing Map (KSOM), hierarchically organised KSOM, a backpropagation-type neural network with fuzzy inputs and outputs, and a fuzzy system, for continuous speech recognition. Experiments with different non-linear transformations on the signal before using a KSOM have been done. The results obtained by using different techniques on the case study of phonemes in Bulgarian language are compared.

The data base used consists of a small sample of 30 seconds of continuous speech articulated by a male speaker. The speech includes 4 sentences, the ten digits, 10 short words, 20 syllables. The pronounced words contain all 25 phonemes, of them - 6 vowels and 19 consonants. The speech has been digitized with 20 kHz of frequency.

2 Using KSOM for Phoneme Recognition and the Importance of Non-Linear Transformations on Speech Signals

Non-linear transformations on the raw speech signal proved to be advantageous to the final recognition accuracy. Some are described here. The first one is log10 transformation of all the 64 Fourier coefficients. Another one is mel-scale filtering. The frequency band is divided into twenty specific bands filtered by corresponding triangular filters, where the first 10 filters are on a linear frequency scale and the other 10 are on a logarithmic frequency scale. The filter outputs are logarithmicised.

Calculating the so called mel-frequency cepstrum coefficients (MFCC) is another non-linear pre-processing transformation. A cosine transformation is calculated on the logirithmicised outputs from the 20 filters. The output vector's dimension could be different, e.g. N=5,10,20 thus having MFCC(5), MFCC(10), MFCC(20). Another non-linear transformation is the calculation of the so called linear frequency cepstrum coefficient (LFCC) which is a cosine transformation over the Fourier spectrum coefficients. A "window" is moving along the time scale and a segment of the signal in the window is taken and transformed by the FFT. The segments overlap, for example on 50% of the window. If the window is 12.8 msec wide, and the discretisation frequency is 20 kHz, then 256 points are taken and weighted through a Hamming window. As the spectrum is taken up to 5kHz, 64 FFT points are used. From the continuous speech sample in the case study, 2050 feature vectors have been extracted and processed.

After having done the pre-processing phase, a 15x15 KSOM has been trained with all the feature vectors; 5000 iterations with a learning coefficient a(0)=0.9 and neighbourhood radius Nr(0)=7 have been done. Different pre-processing methods lead to different phoneme recognition accuracy as illustrated in Table 1. The numbers in the brackets show the dimension of the feature vectors after the non-linear transformation. Obviously the results are far from the best cases reported in the literature on the phoneme recognition task. The reason is that the sample for the study case is a small one. We show here that a non-linear transformation after the FFT increases the accuracy, the MFCC being the best among those tested. The accuracy also depends on the dimension of the feature vectors. Ten-dimensional and 20-dimensional vectors lead to similar results, which are much better than 5-dimensional input vectors. The phonemic KSOM for the case of MFCC(10) is shown in Figure 1.

Table 1.

Method	Accuracy (%)	
	apparent	test
FFT(64)	61	57
FFT +log10(64)	78	74
MEL filters(20)	78	76
MFCC(20)	78	76
MFCC(10)	78	75
MFCC(5)	74	69
LFCC(20)	79	75

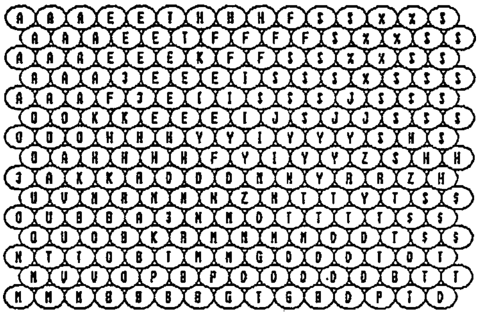
A A A E E I H H H F S S X X S
A A A E E I F F F F S X X S S
A A A E E E K F F S S X X S S
A A A J E E E I S S S X S S S
A A A F J E I I S S S J S S S
O O K K E E E I J S J J S S S
O O O H H H Y Y I Y Y Y S H S
O A H H H H F Y I Y Y Z S H H
J A K K A D D D N N Y R R Z H
U V N R N N N Z N T T Y T S $
O U B B A J N M O T T T T $ $
O U O B K R M M M M D D T $ $
N T T O B T M M G O O D T O T
M V V O P B P O O D D D B T T
M M K B B B B G T G B D P T D

Figure 1

3 Hierarchical KSOM

Instead of having a big, and therefore slow to process single KSOM, hierarchical models of KSOM can be used. The first model tested here uses one 4X4 KSOM at the first level and 16 4x4 KSOM at the second level. Every KSOM at the second level is activated when a corresponding neuron from the first level becomes active. The asymptotic computational complexity of the recognition of the two-level hierarchical model is O(2nm) where n is the number of inputs, m is the size of a single KSOM. This is much less than the computational complexity O(nmm) of a single KSOM with a size of m^2 (m=16 for the experiments). For a general r-level hierarchical model the complexity is O(rnm). The same accuracy as using MFCC and a single KSOM was achieved for the tested sample data set, but a speed-up of 8 times was achieved here.

Another hierarchical model was developed which uses both time-, and frequency-space representation of the input signals. The first-level KSOM is trained to recognise four classes of phonemes, i.e. pause($), a vocalised phoneme(@), a non-vocalised phoneme(!), a fricative segment(#). The network is trained with three time-features of the speech signal which are: MEAN (the mean value of the energy of the time-scale signal within the segment); ZERO (the number of the crossing of zero for the time-scale signal); NOISE (the mean value of the local extremes of the amplitude of the signal on the time scale). After training a small 5X5 KSOM with instances of the four classes taken from the sample speech data set, the network can recognise the four classes with accuracy of 100%, 99%, 94% and 97% respectively. After the phoneme class is successfully recognised, the feature frequency-scale vector (MFCC1,MFCC2,..., MFCC10) corresponding to the same time segment, is passed to a KSOM which corresponds to the winning class. The phonemic maps after training are shown in Figure 2.

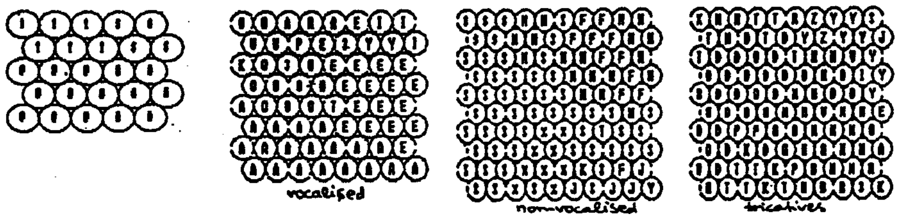

Figure 2

After experimenting with the same training and test sets, the achieved apparent accuracy was 80% and the test accuracy 78 %. The results show that combining input vectors taken from both time- and frequency-space may give better results.

4 Fuzzy Neuro Systems for Phoneme Recognition

In the previous sections KSOM was used for phoneme recognition. But is the "winner take all" paradigm appropriate here? Would fuzzy inference which produces a fuzzy decision vector instead of a 'winning neuron' be more appropriate?

The frequency input features chosen here are the 20 mel-scale filter coefficients obtained after applying triangular mel-scale filters. Every coefficient is fuzzified into three fuzzy values. Their membership functions are defined after calculating the mean of all the mel-filter values for a particular filter band over the training speech segments. A feedforward neural network trained by using the backpropagation algorithm has been used. The network is shown in Figure 3c. Figure 3a gives the membership functions of the fuzzy terms "low", "medium" and "high" for the fuzzy variable "energy of the speech segment on the mel-scale frequency band 1". Figure 3b gives the membership functions of the output fuzzy terms "phoneme /e/- beginning", "phoneme /e/- middle", " phoneme /e/- end". The outputs in this case provide not only information about the ultimate phoneme the currently processed segment belongs to, but about its fuzzy timing among the whole input phoneme signal.

An output decision block analyses the outputs from the network and 'decides' which phoneme the current speech segment belongs to. The current phoneme is not recognised until the end segment(s) of the phoneme are recognised. This is psychologically plausible as we do not decide upon the heard phoneme or word until we hear the end of it or the beginning of the next one. So, the decision is a 'delayed' one. Having three fuzzy concepts, i.e. "beginning", "middle", "end", which accompany every phoneme, helps to significantly overcome the ambiguity of phoneme recognition. For example, using the KSOM for recognising the word 'sedem' in Bulgarian language we achieve the following sequence of recognised phoneme segments: SSSSSSSSSSS %%%SN3 EEEEEEEEEEEEEE TEE3EH DDDD NI3E33 EEE AEDDG NMMMH MMMMM. For the sake of clarity we have separated the clear phoneme sequences from the begin/end ambiguous ones. When the neural network from Figure 3c is used, the following sequence is recognised: SSSSSSSSSSSSSSS N3 EEEEEEEEEEEEEEEEEE H DDDD NI EEEEEEEEE DDG MMMMMMMMMM. In this sequence the correctly recognised segments are greater than in the previous experiment. The recognition accuracy on the training set is 90% and on the test set 86%. This is significantly better than the recognition done by using a KSOM or the hierarchical KSOM model, and slightly better than the one achieved in a hierarchical KSOM - DTW system (Kasabov et al, 1993).

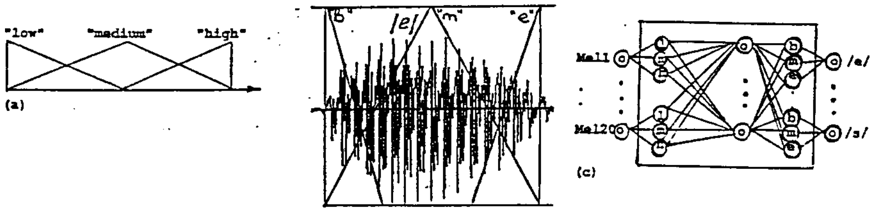

Figure 3

5 Extracting and Using Fuzzy Rules for Phoneme Recognition

Obtaining (learning) a set of fuzzy rules from a trained neural network of the type shown in Figure 3c can be done with the use of the method presented in Kasabov (1993). For the case study, 68 fuzzy rules were extracted. Instead of using the neural network, the set of fuzzy rules can be used for the classification phase. When a MAX/MIN composition inference with centroid defuzzification method was used with the same output decision block for solving the ambiguity of the final classification, an accuracy of 88% and 86% was achieved respectively for the same training and test sets of phoneme segments after initial experiments. The fuzzy system provides better generalisation for the case study. The reason may be that the big diversity in the speech signals is better approximated by 'patches' of fuzzy rules rather than by single points in the output space.

6 Conclusions

The experiments on the phoneme recognition task done here with the use of hierarchical KSOM, a backpropagation network with fuzzified data and fuzzy inference techniques, suggests that those methods are less computationally heavy and provide better generalisation for continuous speech recognition.

References

Morgan, D. and Scofield, C. (1991) Neural networks and Speech processing. Kluwer Academic Publishers

Kasabov, N. (1993) Learning fuzzy production rules for approximate reasoning in connectionist production systems, in: S.Gielen and B.Kappen (Eds) Proceedings of ICANN'93, Springer Verlag, 337-342

Kasabov, N., Nikovski, D. and E.Peev (1993) Speech recognition based on Kohonen Self Organizing Feature Maps and Hybrid Connectionist Systems, in: N.Kasabov (Ed) Artificial Neural Networks and Expert Systems, IEEE Computer Society Press, Los Alamitos, 113-117

University of Otago

Department of Information Science

The Department of Information Science is one of six departments that make up the Division of Commerce at the University of Otago. The department offers courses of study leading to a major in Information Science within the BCom, BA and BSc degrees. In addition to undergraduate teaching, the department is also strongly involved in postgraduate programmes leading to the MBA, MCom and PhD degrees. Research projects in software engineering and software development, information engineering and database, artificial intelligence/expert systems, geographic information systems, advanced information systems management and data communications are particularly well supported at present.

Discussion Paper Series Editors

The views expressed in this paper are not necessarily the same as those held by members of the editorial board. The accuracy of the information presented in this paper is the sole responsibility of the authors.

Copyright

Correspondence

This paper represents work to date and may not necessarily form the basis for the authors' final conclusions relating to this topic. It is likely, however, that the paper will appear in some form in a journal or in conference proceedings in the near future. The authors would be pleased to receive correspondence in connection with any of the issues raised in this paper. Please write to the authors at the address provided at the foot of the first page.

Any other correspondence concerning the Series should be sent to:

DPS Co-ordinator
Department of Information Science
University of Otago
P O Box 56
Dunedin
NEW ZEALAND
Fax: +64 3 479 8311
email: workpapers@commerce.otago.ac.nz