LOW COST SPEECH RECOGNITION FOR SIMPLE DIALOGUE UNDERSTANDING

A. Kowalczyk, M. Dale & C. Rowles.
Telecom Research Laboratories

ABSTRACT - This paper describes the results of some experiments in the development and application of low cost neural networks for isolated speech recognition. Emphasis is placed on low precision weights and low memory requirements, which facilitate, in particular, the use of simple microprocessors for implementation.

INTRODUCTION

The development of practical, low cost devices for spoken language understanding has recently become one of the main aims of researchers associated with speech related industries. A wide range of applications are at stake, such as telephones with voice activated dialling, aids for handicapped people and a variety of cost efficient, user-friendly, telephone-based information services. In some applications the cost of such devices have to be minimal (e.g. < $5 for digit recognition in a basic telephone set).

This paper outlines a number of recent experiments conducted at Telecom Research Laboratories in the development of low-cost, spoken language understanding devices. For cost efficiency the speech recognition is based on a bank of frequency filters followed by a set of quantisers (thresholds) and simple artificial neural networks (ANN), mask perceptrons, for signal classification. Such ANNs are very simple to implement: the only operations required are logical "AND", comparison and addition of small integers. Experiments have shown that such networks require a small number of synaptic weights of very low precision (4-6 bits), hence have low memory requirements and could be implemented on a low-power, low-cost microprocessor or a dedicated logic-based system with a small amount of memory. Note that the memory is the main cost of such a system.

NEURAL NETWORK DESCRIPTION

Two basic architecture of neural networks used for the experiments are shown in Figure 1. The main part, the mask perceptron (Kowalczyk, Ferra & Jenkins, 1990, Kowalczyk & Ferra, to appear), consists of three layers:

- a layer of "input quantisers" (IQ's) converting the continuous outputs of frequency filters to a string of bits, $x_1, x_2, \ldots, x_n$,
- a hidden layer of logical conjunction units (higher order monomials, $x_{j1}x_{j2}\cdots x_{jk}$, and
- the output layer connected via links with real weights, $w_{i(j1,\ldots,jk)}$, to some units of the previous two layers for the purpose of "normal summation" of the weighted activation of these linked units.

A final processing stage is added on top of the mask perceptron in one of two ways. In the first case shown in Figure 1a, known as distributed encoding, each input pattern is assigned 15 non-trivial entries of a row from a 16×16 Hadamard matrix (digit recognition experiments only). These patterns of ±1's have maximal Hamming distance between each other, so that for two instances to be confused at least one of them must lead a number of errors. The distributed encoding of the output units, if implemented, requires a final stage appearing as a "Hamming net" as well as a prior "squashing" level for which we used a ramp non-linearity $f(t) = \text{sgn}(t)$ for $|t| > 1$ and $f(t) = t$ for $|t| \leq 1$. The "Hamming net" is used to calculate the cross-correlation of the threshold-modified perceptron results with the ideal code patterns, designated by the Hadamard matrix, for each word. The weights between "squashing" and "correlation" units in Figure 1 are equal to ±1 corresponding to the designated distributed patterns.

In the second case, shown in Figure 1b, we have a network with so called centralized encoding. In this case the number of mask perceptron output units is equal to the number of words, with each output unit ideally assigned 1 for the word to which it corresponds and -1 for all others, and the Hamming

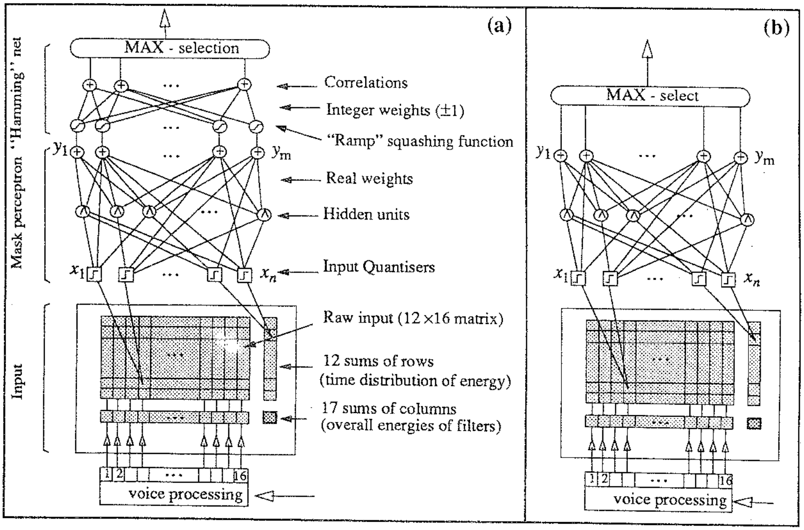

Figure 1. (a) Mask perceptron structure with distributed encoding and (b) mask perceptron structure with centralized encoding.

network layer is omitted. These networks are much smaller than the previously discussed, but in general less accurate.

Regardless of whether centralized or distributed encoding is used we employed a "MAX-selection" in the final stage in order to determine the final output of the network.

DIGIT RECOGNITION EXPERIMENTS.

We conducted a series of experiments in the recognition of 10 isolated spoken digits, which is a standard benchmark task in the speech recognition area. We used mask perceptrons with distributed encoding (c.f. Fig. 1. a) in these experiments. The generated mask perceptrons were subsequently pruned and had simplified synaptic weights in order to reduce the demand on storage memory (c.f. Kowalczyk *at al.* (1991) and Kowalczyk & Ferra (to appear) for details of the training routine). The results are as follows.

(i) Multi-lingual experiment

A neural network was trained for the recognition of 10 isolated digits spoken by 4 speakers in 4 different languages. Each of the training and independent testing data sets consisted of 800 instances (200 for each speaker). All instances consisted of 221 positive numbers representing the energies of 16 filters over 12 time windows, 16 sums for each filter, 12 sums for each time window and the total energy (c.f. Fig. 1a.).

The accuracies of obtained perceptrons with different weight resolution are shown in Figure 2.b (marked by (i)). In particular, a mask perceptron, using 6 bit precision of weights achieved the accuracy of 96% in test and required only 4k of memory to store (c.f. "A" in Figs. 2.a and 2.b).

(ii) A multiple speaker experiment

A neural network was trained for the recognition of 10 isolated digits spoken by 6 speakers in English. Each of the training and independent testing data sets consisted of 1,200 instances (200 for each speaker). All instances consisted of 231 positive numbers representing the energies of 20 filters over

10 time windows and 20 sums of energies for each filter, 10 sums for each time window and the total energy.

The results, marked by (ii), are shown in Fig. 2.b. In particular, the simplified mask perceptron using 5 bit precision weights and requiring 1.7k of memory to store, achieved an accuracy of 98% in the test (c.f. "B" in Figs. 2.a and 2.b); we observed here no significant decrease in recognition accuracy in comparison to 24 bit precision. With better signal processing, we would expect an improvement in recognition accuracy. This network was subsequently simulated on a PC and used in microphone-based tests for five new speakers. These speakers were able to easily adapt their pronunciation to the network requirements and hence improve the recognition accuracy from an initial 75% to 95%.

(iii) A single speaker experiment.

A neural network was trained for the recognition of 10 isolated digits spoken in English. Each of the training and testing data sets consisted of 200 instances and are composed of the constituents of part (ii).

The results, marked by (iii), are shown in Fig. 2. In this case, the simplified mask perceptron using 4 bit precision weights and requiring only 288 bytes of memory to store, was achieving an accuracy of 99.5% in the test (c.f. "C" in Figs. 2.a and 2.b). Analogously, for 5 bit precision we needed 330 bytes of memory and the network achieved 100% accuracy.

(a)

Network	Exper.	Fan-in	No. Bits	Memory	Accuracy	No.Bits	Memory	Accuracy
A	*i*	298	24	14k	95.7%	6	4 k	96.0%
B	*ii*	261	24	12k	98.2%	5	1.7 k	98.0%
C	*iii*	32	24	1.5k	100.0%	4	0.28 k	99.5%

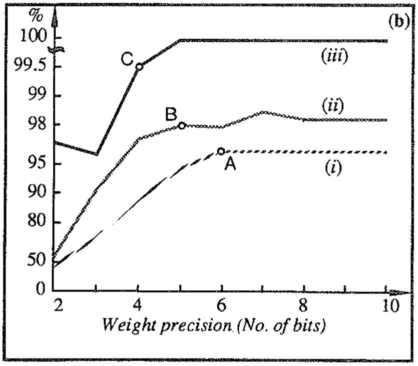

Figure 2. (a) Comparison between selected networks with the full and restricted precision of synaptic weights in the digit recognition experiments: (i) multi-lingual, (ii) multiple speaker and (iii) single speaker. "Fan-in" refers to the number of connections to output units of the mask perceptron (marked $y_1,\dots,y_m$ in Fig. 1). "Memory" means the memory required to store the network in simulation on a micro-processor.

(b) Accuracy of perceptrons versus precision of synaptic weights in digit recognition experiments.

GROUCHO - A SPOKEN DIALOGUE INTERACTION SYSTEM

This ANN approach was used to construct an information retrieval system, nicknamed Groucho, that understands simple man-machine spoken dialogues over the public switched telephone network (c.f. Fig 3). The dialogue was based on a system trained for multiple speakers, with the aim of recognizing 17 isolated words, spoken by a user in response to questions asked by the machine. Questions are asked one at a time and a list of words representing the choices are "spoken" to the caller. The output of the speech recognising ANN is passed to the dialogue manager, which in turn initiates the appropriate audio response to the user.

An important aspect of how the neural networks are used in Groucho is the use of a simple dialogue grammar. Certain types of dialogues can be readily modelled as a sequence of interactional stages. An information-seeking dialogue between a caller and the computer can be modelled as an

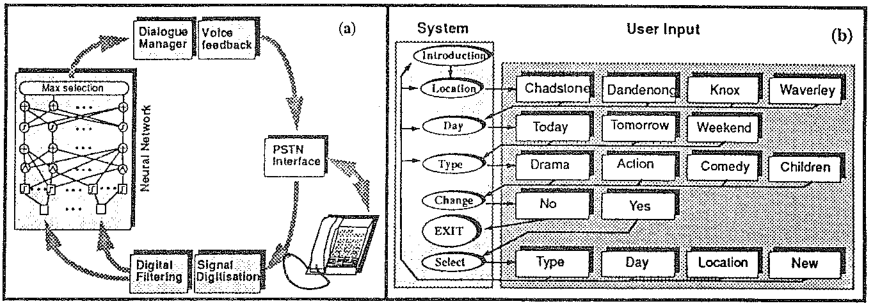

Figure 3.Groucho system: implementation schematic (a) and dialogue grammar (b).

introduction to the service, statement of an information-seeking goal, negotiation of specific query details, presentation of information retrieved, check on query satisfaction, and dialogue closure. On top of these stages can be various sub-dialogues to clarify service possibilities, query details and query satisfaction (Rowles et al. 92 this volume).

The dialogue grammar used in Groucho is shown in Figure 3b. This grammar is modelled on that of (Gagnoulet, C., Jouvet, D. & Damay, J., 1991) The grammar shows the aspects of the query domain that are relevant at each stage of the dialogue. In a small-vocabulary speech response system, an important issue is how we restrict the caller's vocabulary to that which can be recognised by the system. To ensure this, we allow Groucho to seize the conversational initiative by stating from the beginning, and at each stage, what the domain context is and what input options are expected at that stage, so that the caller knows what information to provide and how to express it.

The key advantage this structure provides is a reduced number of words expected to be uttered at each stage of the dialogue. Thus, rather than having to decide which of 17 words was uttered, we need only to decide which of a maximum of 4 words was uttered. This improves recognition robustness and simplifies the networks as a simple network is used for recognition at each stage. If none of the expected words is a strong contender, the system can try networks for other dialogue stages as a back-up.

Each of the training and testing data sets consisted of 20 instances of the 17 words from 30 speakers. This gave a total of 5100 instances for both training and testing. Each instance consisted of 187 positive real numbers representing the energies of 16 filters over 10 time windows, including the sum of each filter and time window (17x11).

Groucho was implemented on an IBM-PC using the Sound Blaster Pro card for speech digitizing and spoken feedback generation. In a simplified version, with synaptic weights of 5-6 bits precision (the full version used 24 bit floating point weights), the ANNs used for the Groucho system require only 3 k of memory to store. The accuracy of such a system is around 94-98% on tests with speech data not used in training. The reduction in precision of the synaptic weights resulted in a decrease in accuracy of less than 1% (c.f. Fig 4.).

DISCUSSION

1. Simplicity of the classifier. The ANNs considered in this paper (mask perceptrons) are very simple to implement in hardware. Practically, the only operations required are checking of logical conditions ("AND") and summations of small numbers (there are no costly multiplications). Due to this simplicity, for the tasks considered, the networks simulated even on a very simple micro-processor (6800 etc.) provides real-time word classification, and the speed of preprocessing (filters) is the most time consuming component of the recogniser.

Level	Fan-in	No. Bits	Memory	Accuracy	No.Bits	Memory	Accuracy
1	177	24	2.5 k	94.7%	6	0.8 k	94.3%
2	156	24	2.2 k	96.8%	5	0.6 k	96.7%
3	124	24	1.7 k	97.2%	5	0.5 k	97.2%
4	239	24	3.4 k	96.7%	6	1.0 k	96.0%
5	81	24	1.1 k	98.7%	5	0.1 k	98.7%

Figure 4. Comparison between implementations of Groucho with the full and restricted precision (number of bits) of synaptic weights. "Level" refers to the level of the dialogue grammar in Fig. 3. "Fan-in" and "Memory" have the meaning as in Fig. 2.

2. *Relative memory efficiency.* For comparison, one of the popular commercial systems for isolated speech recognition requires 2.4k of memory per single template (word). Hence, for a multilingual application requiring 40 templates, it will use 96 k of memory, i.e. 24 times more than the ANN in (i), not to mention the much greater computational power required by their classifier. The corresponding numbers for multiple speaker and single speaker recognition experiments requiring at least 10 templates are 14 and 83, respectively. This memory efficiency demonstrated by mask perceptrons allows a number of simple practical approaches to improve accuracy, e.g. by combining multiple independent ANNs and using a voting scheme to determine the most likely correct word.

3. *Some implementation issues.* In some potential applications the cost of speech recognition devices must be minimal. For example, an estimated cost of a practical digit recognition system for the basic telephone would need to be less than $5, while for a small business telephone system the cost would need to be less than $20. The experiments reported in this paper show that mask perceptrons can be used to build such practical devices. Two options can be considered here.

(i) A dedicated digital VLSI chip using a combination of AND gates and adders (Kowalczyk *at al.*, 1991). The processing of the speech can be implemented in analog VLSI using a combination of bandpass filters, integrators and comparators. Practically, in this case the network would have to be trained off-chip.

(ii) The use of a low cost processor for the simulation of the network (e.g. Motorolla 6811, Intel 8051, Analog Devices 2105, etc.). The front end needed for the processing of the speech may still be implemented as before. In this system, it will be relatively easy to update and adapt automatically the structure and the weights of the network to the users as the system is used and the new training data becomes available.

4. *Audio feedback generation.* In a practical implementation of a spoken language understanding system the generation of audio feedback becomes an issue. In general, some form of response is required so that the user has confirmation that the recognition system has understood them correctly. Audio feedback seems to be most natural. Currently there are VLSI devices available to convert text to speech. However, devices of acceptable quality are too expensive to be used in a low cost system. An alternative solution is to store the digital representation of the speech in a ROM (~40k for ten digits) and then to use a DAC to play back the required response.

5. *Relevant research.* The usage of low precision weights is one of the issues in practical development of ANNs (e.g Xie & Jabri, 1992)). Some authors reported that 6-8 bit resolution of synaptic weights was sufficient to obtain practical ANNs with reasonable performance (c.f. IEEE Trans. Neural Network, 1992). Fig. 2.a shows that 4-6 bits are sufficient in our case.

The idea of using a simple signal processing based on a bank of frequency filters followed by a simple feed-forward neural network for classification was also pursued by others (e.g. Unikrishnan, Hopfield & Tank, 1992).

6. Usage of real life data. It is worthwhile to mention that in our experiments apart from the simple signal processing we used relatively noisy, 'real life' data (e.g. spoken in natural office environment, over a public telephone line with a signal to noise ratio of 15 dB). This should be contrasted with "clinically pure" database recorded in sound-proof environment used often by others in speech recognition experiments.

7. Future research. More research is needed to determine how many speakers are required to produce a reasonable system with an ANN small enough to meet the accuracy and cost specifications. Also the optimal selection of filters, time windows and inclusion of other speech attributes should improve the accuracy of the system and will be a subject of future investigations.

CONCLUSION

These experiments show that a practical spoken language communication system can be implemented using an ANN approach, a simple dialogue grammar and simple low cost technology. With further effort we may expect improvement, especially in terms of larger vocabulary and accuracy in speaker independent recognition tasks.

ACKNOWLEGEMENT.

We acknowledge the contribution of Mr. R. W. Newton, of The University of Queensland, in the development of Groucho system. The permission of Director of Research, Telecom Australia, to publish this material is also acknowledged.

REFERENCES

IEEE Trans. Neural Networks (1992), 3(3), Special issue on Neural Network Hardware.

A.Kowalczyk &H.L. Ferra, and G.Jenkins (1990). *Experiments with mask-perceptrons for speech recognition.* In R.Seidl, ed., Speech Science and Technology - Proceedings 1990, pp. 16--21, Canberra. ANU Printing Service.

A.Kowalczyk, G.Aumann &H.L. Ferra, and J.Cybulski (1991). *Associative mappings with positive bounded coefficients.* In T. Kohonen at.al, eds, Artificial Neural Networks, Proc. Inter. Conf. on Art. Neural Net. (ICANN-91) Espoo (Finland), 1991. North-Holland, Amsterdam.

A. Kowalczyk & H. Ferra (to appear). *Developing Higher Order Networks with Empirically Selected Units*, IEEE Trans. on Neural Networks

Rowles C., Huang X., de Beler M., Voinwiller J., King R., Matthiesson C., Sefton P. and O'Donnell M. (1992), *Using prosody to assist in the understanding of spoken English*, this volume.

K. P. Unikrishnan, J.J. Hopfield and D.W. Tank (1992), *Speaker-Independent Digit Recognition Using a Neural Network with Time-Delayed Connections*, Neural Computation, 4 (1), 108-119.

Y. Xie and M.A. Jabri (1992), *Analysis of the Effect of Quantisation in Multilayer Neural Networks Using a Statistical Model*, IEEE Trans. Neural Net., 3 (2), 334-38.

Gagnoulet, C., Jouvet, D. & Damay, J., (1991), *Mairievox: A Voice Activated Information System.* Speech Communication, Vol. 10, No. 1.