Cantonese Phonemes Recognition via the Gated Neural Network

Andrew Luk, C.P. Cheung, S.H. Leung, and W.H. Lau

Department of Electronic Engineering
City University of Hong Kong (designate)
Email: ee000alu@CityU.edu.hk

ABSTRACT - This paper examines the idea and structure of a gated neural network that can be used to recognize Cantonese phonemes. The idea is to partition a fully connected multi-layered feedforward neural network (MLFNN) into a number of functional areas, each of which is a fully connected MLFNN and responsible for a subset of the original problem. These functional areas are activated via gated signals generated from either some other functional areas or external sources. Preliminary results have indicated that such network structure can achieve better performance than the MLFNN.

INTRODUCTION

In a previous correspondence (Leung, Luk, Liu and Lun, 1992), the formant characteristics of the eleven vowel phonemes were examined using a new autoregressive moving average (ARMA) algorithm. The analysis then was not automatic, phonemes had to be labelled manually. An automated system that could identify and label Cantonese phonemes was envisioned. One possible approach would be to use a single fully connected multi-layered feedforward neural network (MLFNN) (Domany, van Hemmen and Schulten, 1990). Unfortunately, automatic Cantonese phoneme recognition via such network is a very difficult task because of the heavy computational complexity both in the learning and recall stages. Previous work (Zahorian and Jagharghi, 1992) has shown that for English (American) vowel phonemes classification, only about 75% of accuracy can be achieved via back-propagated learning on a fully connected MLFNN.

This paper proposes an alternative structure, referred to as the gated neural network, which divides the original network into smaller subunits, each of which is activated via neuronal output(s) from the previous subunit structures or external sources. Thus the network behaves as a neuronal circuit where outputs from the previous subunits act as neuronal gates to switch on (or off) subsequent subunits. This biologically motivated modification greatly simplify the learning of each subunit and make it easier to partition the original network.

The next section of this paper will introduce the idea and structure of the gated neural network. The section that followed will concentrate on the application of such structure in Cantonese phoneme classification; comments and discussions on the experimental results will also be included. Finally, a brief conclusion will be provided at the end of this paper.

GATED NEURAL NETWORKS

Neural network architectures are motivated by models of human brains and the nerve cells. In many cases, we always simplify the brain as a single densely interconnected neural network. But, biological research has verified that the brain, instead of being a huge single piece of computational machine, is in fact highly localized and can be partitioned into a number of functional areas/parts (Noback and Demarest, 1981). These functional areas are responsible for functions such as hearing, speech, vision, and other locomotion activities. For instance, using the Brodmann's classification scheme (Brodmann, 1910), the brain can be parcelled into 47 areas. The Broca's speech areas are located at areas 44 and 45. This idea can clearly be applied to the design and construction of (artificial) neural networks, and many different interpretations of these localization effects have been given in the literature. For instance, a simple binary tree structure (neural tree network) can be constructed (Rahim, 1992) for English (American) phonemes classification, where each leaf or

terminal node at the bottom of the tree represents a phoneme. All other nodes, including the root of the binary tree, serve as bifurcation (decision) indicators only. In such a structure, input only occurs at the root of the network. Alternatively, a hierarchical neural network can be constructed, as proposed by the authors earlier (Ng, Leung and Luk, 1992), for the classification of words. In that network, the inputs are channelled to different subnetworks for classification operation.

Another useful idea that can be explored in the design of a neural network is stemmed from the basic electronics. In simple sequential logic, a sequence of gates are switched on and off via some control signals. A portion of the circuitry will be activated by one or more control signals. These control signals can be the outputs from another portion of the circuitry or from an independent signal.

These two ideas can be combined together in the design and construction of a neural network. We term this type of networks *gated neural networks* (GNN). The basic concept of a gated neural network is to allocate different parts of a job to different functional areas (or units or subnetworks) of a hierarchical neural network. Each of this functional area will be activated by one or more control gate signal. On the other hand, the outputs of each functional area can be control and/or decision signals. This is analogous to the switching circuit, where each functional area can serve as both output and control signals. A typical example is the output of a 3-to-8 decoder, where the output can be used as chip-select signal for some other circuitry (like a RAM module) or as an output signal to drive, say, an LCD indicator. Figure 1 illustrates some of the basic idea of a gated neural network.

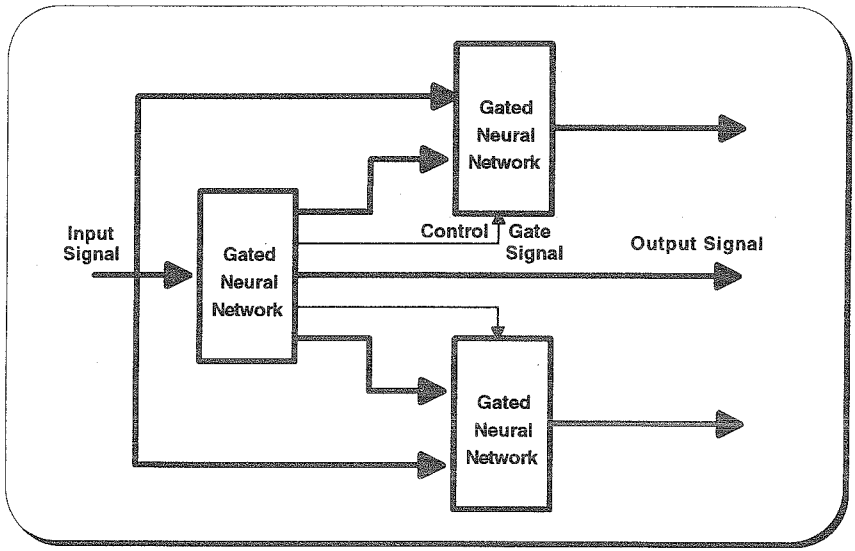

Figure 1: Basic Idea of a Gated Neural Network (GNN).

One of the main advantages of the proposed neural network is that not every functional area of the neural network needs to be activated at the same instant; it is possible to activate only part of the network depending on the type of input signals. Another major advantage is that each functional area can accept different input signals.

In the next section, preliminary results on the application of this type of gated neural network on Cantonese phonemes recognition will be presented.

In our system, the raw speech data from a single subject is sampled at 10 kHz and high frequency emphasized. For each 256-point high frequency emphasized data, it is windowed by a 3-term Blackman-Harris window of size 512 points with zero padding. The windowed data is then Fourier transformed into frequency domain, its logarithmic magnitude is then computed. The 256-point logarithmic magnitude coefficients are then grouped into a number of perceptual frequency bands via the analysis of Cantonese phonemes (Leung, Luk, Liu and Lun, 1992). A direct application of the perceptual frequency bands proposed earlier (Ng, Leung and Luk, 1992) will not partition the frequency space sufficiently for good phoneme classification. It follows that additional partitioning is needed for region where two or more loci of phonemes are closely clustered in the F1-F2 or F2-F3 diagrams (or planes), where Fi denotes the i-th formants frequency (in Hz). Such partitioning can be used directly as input to different functional areas of the proposed gated neural network model. Figure 2 illustrates the overall architecture of the gated neural network for Cantonese phoneme classification.

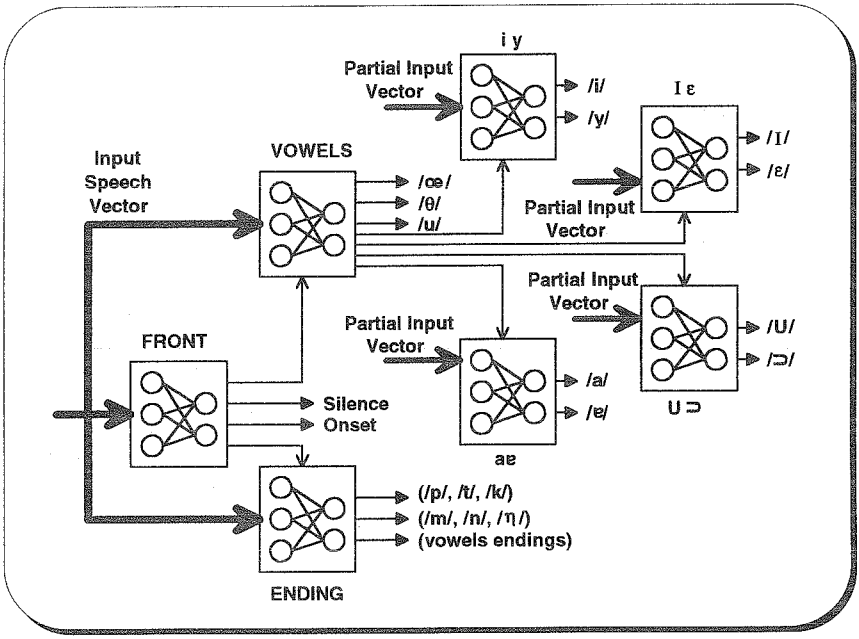

Figure 2: Gated Neural Network for Cantonese Phonemes Recognition

Note that the functional area referred to as FRONT is used to classify the incoming frame of speech data as either silence, onset, vowel or ending. In this preliminary work, the onset consonants are not further subdivided into their corresponding phonemes. Since the Cantonese word usually ends with a stop (/p/, /t/, and /k/), a vowel or a nasal (/n/, /m/ and /ŋ/), the functional area labelled as ENDING is used to classify the incoming frame as member(s) of these three categories of endings. On the other hand, the vowels are recognized using five functional areas. In Figure 2, the functional area VOWELS is activated by the *vowel* gate signal which is output from the FRONT functional area (thin line with an open triangular head). This VOWELS functional area is then used to classify three vowel phonemes (namely, /oe/, /θ/ and /u/) and generate four gate signals. These four gate signals are, in turn, used to activate the four functional areas labelled as 'i y', 'I ɛ', 'a ɐ' and 'U ɔ'.

As mentioned earlier, the input to each functional area may be different. In Figure 2, this is denoted by the two lines of different thickness. The inputs to functional areas FRONT, VOWELS and ENDINGS are the same average logarithmic spectral values from the nine frequency bands as shown in Table 1.

Band	Frequency Range (Hz)
1	200 - 400
2	400 - 600
3	600 - 800
4	800 - 1000
5	1000 - 1300
6	1300 - 1600
7	1600 - 2000
8	2000 - 2400
9	2400 - 3200

Table 1: Frequency band partitioning scheme used in FRONT, VOWELS and ENDINGS functional areas.

However, the frequency band partitioning scheme for the other functional areas are all different. For example, only four frequency bands are used as input in the 'i y' and "ɪ ɛ' functional areas. They are tabulated as shown below in Tables 2 and 3 respectively.

Band	Frequency Range (Hz)
1	1950 - 2150
2	2150 - 2300
3	3000 - 3200
4	3200 - 3400

Table 2: Frequency band partitioning scheme used in 'i y' functional area.

Band	Frequency Range (Hz)
1	1550 - 1750
2	1750 - 1950
3	2150 - 2300
4	2300 - 2450

Table 3: Frequency band partitioning scheme used in 'ɪ ɛ' functional area.

For the other two functional areas, another two frequency band partitioning schemes are used. The rationale for these partitioning schemes stems from the fact that these phonemes are closely spaced in the F1-F2 plane but not necessarily in the F2-F3 plane (Leung, Luk, Liu and Lun, 1992). Consequently, only the fine details of these phoneme pairs in the F2-F3 plane need to be examined by this set of functional areas.

Preliminary result of the FRONT functional area on Cantonese phonemes recognition is given in the following modified confusion matrix. This matrix is slightly different from the standard one because the rows and columns do not necessary add up to a hundred percent. This is due to the fact that if any of the output of FRONT is not greater than say 0.9, then that frame of speech data will be considered as belonging to multiple number of categories if the values of these outputs are greater than a predefined threshold. For instance, if no single output of the FRONT functional area is greater

than 0.9 and there are two outputs, say onset and vowel outputs, with a value of 0.45 (which is greater than a pre-set threshold, say, 0.25), the program will consider that particular speech frame as both an onset and a vowel frame. On the other hand, if none of the output values in the FRONT functional area are greater than the pre-set threshold, that frame is unclassified and will be skipped. Consequently, the rows and columns of the confusion matrix will not necessarily add up to a hundred percent.

	Silence	Onset	Vowel	Ending
Silence	98.59%	1.67%	0.00%	0.29%
Onset	0.00%	43.00%	53.11%	11.21%
Vowel	0.00%	19.61%	75.65%	13.34%
Ending	0.02%	0.37%	2.87%	92.63%

Modified confusion matrix for the FRONT functional area

Note that the FRONT functional area has a very good discriminating power (over 90%) with the silence and ending speech frame. On the other hand, the onset is frequently confused with the vowels and a number of vowel frames are also considered as onset or ending. This is quite natural because the speech frame are non-overlapping and it is possible that the speech frame of 256 sampled points may include more than one phoneme. For instance, the stop onset in the Cantonese word 炒 (/t/ /a/ /p/) is very short. Therefore, when sampled at 10 kHz, it may include the phoneme /a/ in its speech frame.

Once the input speech vector has passed through the first functional area, a number of gate signals will be activated. If the input is a vowel frame, a vowel gate signal will be output. This signal will then activate the VOWELS functional area. The modified confusion matrix for the VOWEL functional area is shown below:

	i y	I ɛ	oe	θ	a ɐ	u	U ɔ
i y	99.77%	0.12%	0.16%	0.00%	0.00%	0.14%	0.00%
I ɛ	0.10%	99.66%	0.15%	0.06%	0.00%	0.00%	0.00%
oe	0.19%	0.34%	99.28%	0.28%	0.21%	0.00%	0.00%
θ	0.00%	0.00%	0.31%	99.33%	0.39%	0.00%	0.12%
a ɐ	0.00%	0.00%	0.11%	0.25%	99.59%	0.00%	0.24%
u	0.33%	0.01%	0.00%	0.00%	0.00%	99.48%	0.30%
U ɔ	0.00%	0.12%	0.00%	0.17%	0.16%	0.17%	99.72%

Modified confusion matrix for the VOWELS functional area

Similarly, in the preliminary results, high recognition rate (over 99% correct classification) are obtained for the rest of the functional areas. This may not seem surprising because the FRONT functional area has already identified the incoming speech frame as belonging to one of the four categories of output. Consequently. the potential error frames (i.e. those that belong to multiple categories or none of the categories) have already been screened out by the first FRONT functional area. Subsequent recognition by the VOWELS and other functional areas will be more accurate since these speech frames can be easily classified.

CONCLUSIONS

This paper introduces the idea and structure of the gated neural network. Localization in brain functions and electronic switching circuitry are combined to form a flexible network. Its application in Cantonese phonemes recognition is outlined. Preliminary results have indicated that such network can perform better than the standard MLFNN. Furthermore, the decomposition of a single huge

network into smaller functional areas (or subnetworks) enables a simple training strategy to be used in each of the functional area. Each functional area only requires to train on its specific outputs, which is a subset of the original problem. Different levels of generalization can be introduced at various stages of the gated neural network. Indeed, the gated neural network can be viewed as a generalized tree structure with independent activations at each node.

The preliminary results reported in this paper is based on a single male human subject and a corpus of about a hundred specially selected Cantonese words, which include all the possible know combination of the Cantonese phonemes. Consequently, good recognition rate can be achieved for this small database. Future work will including expanding the current database to include other commonly used Cantonese words and multiple speakers. Additional functional areas are also added to the current network for consonant recognition. The performance of these additional units will be reported in future correspondences.

ACKNOWLEDGEMENT

The authors would like to express their sincere thanks to Dr. C.F. Chan and Dr. A. MacDonald for their invaluable and helpful advices. We would also like to thank Mr G.K.F. Liu for his help in preparing the Cantonese word database and Ms S.W.Y. Lien for proof reading and preparing this manuscript. This work is supported by the CPHK strategic research grant 700-194.

REFERENCES

Brodmann, K. (1910) *Feinere anatomie des Grosshirns*, Lewandowsky's Handbuch der Neurologie, Vol. 5, 206-307.

Domany, E., van Hemmen, J.L. & Schulten, K. (1990) *Models of Neural Networks*, Springer-Verlag, Berlin.

Leung, S.H., Luk, A., Liu, G.K.F. & Lun, C.S. (1992) *An ARMA model for extracting Cantonese phoneme characteristics*, Proceedings of the Fourth Australian International Conference on Speech Science and Technology, St. John' College, The University of Queensland, Brisbane, Australia, 692-697.

Ng, H.C., Leung, S.H. & Luk, A. (1992) *An isolated Chinese word recognition system using hierarchical neural network with applications to telephone dialling*, Proceedings of the Fourth Australian International Conference on Speech Science and Technology, St. John' College, The University of Queensland, Brisbane, Australia, 816-821.

Noback, C.R. & Demarest, R.J. (1981) *The Human Nervous System: Basic Principles of Neurobiology*, Third Edition, McGraw-Hill Book Company, New York.

Rahim, M. (1992) *A neural tree architecture for phoneme classification with experiments on the TIMIT database*, Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing, San Francisco, California, USA, Vol. 2, 345-348.

Zahorian, S.A. & Jagharghi, A.J. (1992) *Minimum mean-square error transformations of categorical data to target positions*, IEEE Trans. on Signal Processing, Vol. 2, 1992.