AN ISOLATED CHINESE WORD RECOGNITION SYSTEM USING HIERARCHICAL NEURAL NETWORK WITH APPLICATIONS TO TELEPHONE DIALING

Hing C. Ng, Shu H. Leung and Andrew Luk

Department of Electronic Engineering
City Polytechnic of Hong Kong

ABSTRACT - This paper is to present a neural network based isolated word recognition system for monosyllabic language especially for Cantonese. The features are extracted from FFT-based filter bank that is designed according to the formant characteristics of the Cantonese phonemes. A hierarchical neural network is used for recognizing feature vectors with good recognition rate and moderate computational complexity.

INTRODUCTION

The goal of speech recognition is to make machines understand normal human speech and, in turn, be able to perform some task based on the understanding. The applications have been found in various areas such as word processing and remote control.

For monosyllabic speech, isolated word recognition is a quite difficult problem because the acoustic information is much smaller than those polysyllabic languages. Due to this fact, we intensively study the acoustic characteristics of the phonemes of the words in the vocabulary and the details of the work is reported in (Leung et. al., 1992). Such information provides the basis to design filter bank for extracting the features for recognition. In order to make real time recognition using digital signal processor possible, we limit the number of frequency bands to nine that requires only moderate system complexity but provides sufficient frequency information.

The Cantonese is a tonal language that has nine tones of variation. Such feature provides further information for differentiating isolated words. For most of the common words, the acoustic pattern is almost in the form of "CVVC" where C and V stand for consonant and vowel, respectively. In order to provide the tonal variation in the feature vector, we subdivide the word linearly into ten time-segments that could be sufficient to show the formant trajectories.

For monosyllabic language, it is quite reasonable to assume that the acoustic patterns vary almost accordingly to the word duration. The even partition of the word into ten segments can thus, in certain extent, retain the acoustic pattern without aligning the time duration of the test words. For improving the detection of the end points of the test words, we allow each test word to have a number of boundary sets. The score of the test word in the recognition is the best one among all the allowed boundary sets.

For using fully connected neural network, the computational complexity and the connections for the input dimension of ninety is seemed too large for real time implementation. In the paper, we use two approaches to reduce the neural network structure for lowering the computations: (1) use hierarchical structure to cluster the feature space; and (2) reduce the input dimension by subdividing the network into subnetwork structure. Besides, the hierarchical structure provides a means to improve the

feature space for enhancing the recognition rate.

PHONEMIC CHARACTERISTICS OF CANTONESE

Cantonese is one of the dialects speaking in the southern part of China and is the local dialect of Hong Kong and Canton. Just like most of the Chinese dialects, Cantonese is a tonal language with nine tones of speaking and seven basic vowel phonemes. The formant frequencies of these phonemes are found in (Leung et. al., 1992) and tabulated in Table 1.

	F_1 (Hz)	F_2 (Hz)	F_3 (Hz)
[i]	279	2197	3108
[y]	266	2080	3267
[I]	510	1720	2310
[ɛ]	533	1767	2284
[œ]	490	1472	2242
[ɵ]	529	1248	3774
[a]	664	1174	2085
[ɐ̯]	628	1219	2273
[u]	268	695	1104
[U]	452	815	2513
[ɔ]	492	813	3551

Table 1 Formant frequencies of Cantonese vowel phonemes

The words chosen for telephone dialing with their phonetic transcriptions are summarized in Table 2:

Cantonese word	IPA symbols	Cantonese word	IPA symbols
zero	liŋ^4	six	luk^9
one	jɐ̯t^7	seven	tsʰɐ̯t^7
two	ji^6	eight	bat^8
three	sam^1	nine	gʷɐ̯u^6
four	si^3	*	siŋ^1
five	ŋ^5	#	tsɛŋ^1

Table 2 Phonetic Transcription of Cantonese Words

The number in the superscript denotes the tone of speaking of the word.

FEATURE EXTRACTION

The speech signal is passed through a low pass filter with the cutoff frequency of 3.2 kHz and sampled at 8 kHz. An automated endpoint detector is used to locate the boundaries of the test word based on the zero-crossing and the energy envelope. Every 32 ms frame (about 256 samples) is pre-emphasized and windowed by Hamming window. We use 50% overlap and thus the frame interval is about 16 ms.

The speech samples are transformed by FFT and the first 128 frequency components are used to form nine energy groups according to the frequency bands listed as follows:

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

The frequency ranges are carefully chosen so as to minimize the ambiguity for identifying the formant frequencies of the Cantonese phoneme vowels. The logarithm of the energy vector forms the basic feature vector to the neural network for recognition.

Since the duration of a test word is not always the same, certain time alignment is needed so that all words have the same feature length. Dynamic time warping (DTW) is an efficient linear time warping algorithm and is commonly used for doing the job. However, the DTW requires quite intensive computation that prohibits its use in real time applications despite its accuracy of time alignment.

In order to simplify the time alignment, we subdivide the time segments into ten time groups in which the time segments are averaged out. Therefore, we use the resulting ten time segments to represent the word irrespective of the time duration. For using ten segments, we could obtain sufficient time information to describe the tonal variation which is essential for recognizing Cantonese words.

For further improving the endpoints, we allow the boundary tolerance on both ends about one frame interval. Therefore each test word has four boundary sets. The score of the test word in the recognition is the maximum among the four.

NEURAL NETWORK STRUCTURE

Neural network has been widely applied in many pattern recognition problems (Morgan & Scofield, 1991). Sometimes it is referred to artificial neural network (ANN) that contains cells whose function is identical to real neurons: they must integrate input from other cells and communicate the integrated signal.

ANN is a very attractive structure for solving complex problem not only because of its "brain-like" appeal but also it has automatic design capability through some learning strategies. There are a number of learning algorithms to train neural network to recognize the desired features. In this paper, we adopt back propagation learning algorithm for learning the connection weights in the network. Mathematically, back propagation is a gradient descent method that uses the output error to calculate the gradients of the error function with respect to the connections and then to update the connection weights.

The back propagation learning algorithm is briefly summarized as follows:

(1) Input a training pattern ξ_k to the input layer,

$$v_k^0 = \xi_k \quad .$$

(2) Propagate the signal forward and compute the neuron outputs

$$v_i^m = g(h_i^m) = g\left(\sum_j w_{ij}^m \, v_j^{m-1} \right) .$$

(3) Compute the deltas for the output layer

$$\delta_i^M = g'(h_i^M)(d_i - v_i^M) , \qquad d_i = \text{desired signal}$$

(4) Compute the deltas for the preceding layers by propagating the errors backwards

$$\delta_i^{m-1} = g'(h_i^{m-1}) \sum_j w_{ij}^m \, \delta_j^m .$$

(5) Update the weights according to

$$w_{ij}^{new} = w_{ij}^{old} + \eta \, \delta_i^m \, v_j^{m-1} .$$

(6) Repeat (2) to (5) for next pattern.

The complexity of neural network is in proportion with the input and output dimensions. In order to reduce the computations while maintaining the recognition rate, we use the following methods to modify the neural structure:

(1) Use hierarchy structure as shown in Figure 1 to reduce the output dimension - We make use of the formant characteristics of the phoneme vowels to partition the words into six groups as follows:

group	word	group	word
1(i)	2, 4	4(ɒ)	1, 7, 9
2(I, ɛ)	0, *, #	5(u)	6
3(a)	3, 8	6(ŋ)	5

A neural network is trained to identify the phoneme vowels for which there are only six output nodes and we use those six time segments in the middle portion of feature vector as training pattern. Since the phonemes have very distinctive features, we can obtain very high classification rate.

After the first network has selected the group, a second network is used to recognize the elements or to reject the word. In this partition, the second network has at most three output nodes that lowers the network complexity quite substantially.

(2) We use subnet structure as shown in Figure 2 to reduce the connectivity - In the second network of the hierarchy, we use the full feature vector as the input. Then the connectivity will be too high for real time implementation if fully connected network is used. Thus we propose to divide the input vector into three vectors and each of them contains four time segments (there is one time segment in overlap). These vectors can be considered to consist of the acoustic features regarding the onset, phoneme vowel and ending, respectively. These vectors are inputted to three subnets that have only two layers and the subnets are fully connected to a hidden layer and then to the output layer.

The network is shown to have very close performance as the fully connected network

RESULTS

We use five different voices to train the recognition system. The system is implemented in real time on a TMS320C30 development system. The performance of the system is quite satisfactory and the recognition rate is about 95%.

CONCLUSIONS

An isolated word recognition system for Cantonese has been successfully implemented on a TMS320C30 development system using hierarchical neural network. The performance is quite satisfactory and the recognition rate is about 95%.

REFERENCES

Leung, S. H. et. al., (1992) "An ARMA Model for Extracting Cantonese Phoneme Characteristics", to be presented at SST-92.

Morgan, D. P. & Scofield, C. L. (1991) Neural Networks and Speech Processing, Kluwer Academic Publishers.

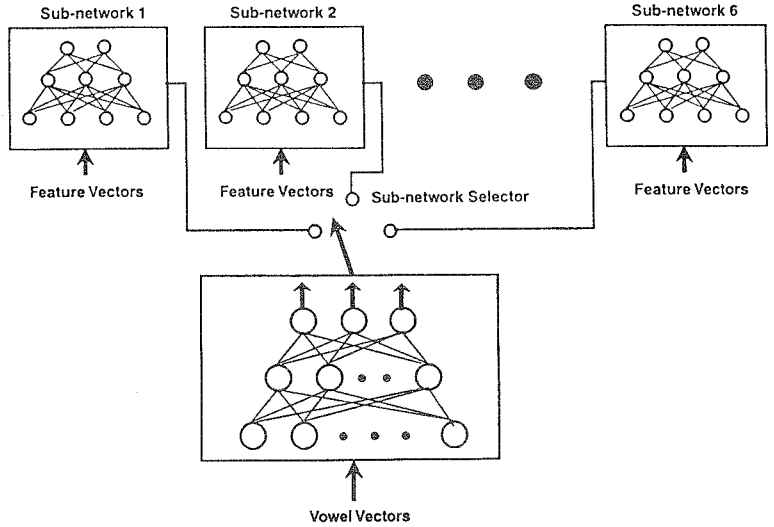

Figure 1 : Hierarchical Structure

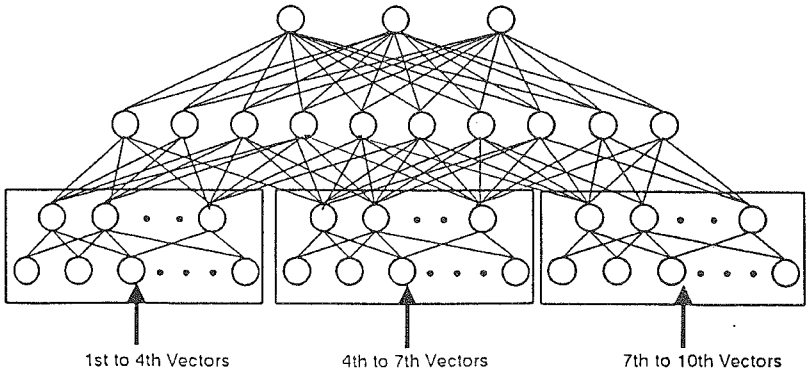

Figure 2 : Sub-network Structure