AN OFF-LINE WORKING SPEECH RECOGNITION SYSTEM EMPLOYING A COMPOUND NEURAL NETWORK AND FUZZY LOGIC

Liqing Zhou
Department of Radio Engineering, Box 96
Beijing University of Posts and Telecommunications
100088 Beijing, China

ABSTRACT

This paper introduces an off-line working speech recognition hardware system. A new compound structure of neural networks is proposed and fuzzy logic is adopted to implement the system. So the system is able to perform speaker-independent real time speech recognition in actual environments where there are heavier noises.

1. INTRODUCTION

So for as the research on neural networks is concerned, there have been many achievements in theory [1], and actual applications to various fields, including pattern recognition and speech processing, are being rapidly developed [2 ~ 6]. This paper is about the application of neural networks and fuzzy logic to speech recognition. It introduces a speaker-independent speech recognition system for a small vocabulary. A new compound structure of hierarchical neural networks is proposed and fuzzy logic is employed in order to implement the system. A high-speed digital signal processor, TMS320C30, is used as the nuclear device to make the hardware. The hardware system is computer-independent and able to perform real time speech recognition in the case of heavier ambient noises. The recognition time to an utterance is less than one second. A field test for 10 Chinese digits (0-9) to 70 persons from 18 provinces shows that a recognition rate of 91% is reached.

The system adopts LP-derived cepstrum [7][8] as the speech feature parameters. This paper mainly introduces how the neural networks and the fuzzy logic are applied to a actual task of speech recognition and how the hardware system is made.

2. APPLICATION OF NEURAL NETWORKS

In order to get a high recognition rate on the condition that involves a wide range of ambient noises and speaking styles, a great deal of work about speech parameter choosing and data processing is carried out, these jobs are not narrated in this paper, besides, a new compound structure of hierarchical neural networks is proposed based on many experiments.

Paper [9] presents a hierarchical structure of neural networks and uses it for the speech recognition of a small vocabulary of English words. The first level of the hierarchical structure distinguishes pairwise the words of a vocabulary. The level of neural networks is called basic network . The second level of the hierarchical structure performs a final judgment. The level of neural networks is called selection network.

It is known that speech features are easy to confuse. However, this means of pairwise classification is comparatively fit for distinguishing easy-to-confuse classes owing to its some fuzzy properties inherently. The basic network adopts the ordinary perceptron structure in [9].

So far as the speech recognition of single-syllable words in heavier ambient noises is concerned, because some consonant messages at the beginning of every utterance are lost, the discrimination of speech features becomes more difficult. For example, in the case of heavier ambient noises and a wide range of speaking styles, on the spectrum features of 10 Chinese digits, the clustering property for the words of the same class is poor and the confused extent for the words of different classes is large. Hence, even if such a hierarchical structure of neural networks is adopted, the results are still not very good, the recognition rate is 86.8% when 1299 templates which do not join in training are tested.

In order to raise the recognition rate, many experiments on the structure of the basic network are carried out. As a result, a new compound structure of neural networks is proposed. The basic network for the structure is composed of numbers of subnetworks. The number of the subnetworks is equal to the number of the output nodes of the basic network. These subnetworks have common input layer. Their middle layers are independent of one another. The number of the hidden nodes of these middle layers are one and the same. Every subnetwork has a output node. All outputs of these subnetworks form the outputs of the basic network. For the speech recognition of 10 words, the structure of the basic network is shown in Fig. 1. In the figure, the input layer has 132 nodes, the middle layer of every subnetwork has 6 hidden nodes, the structure has 45 subnetworks in all. In fact, in order to reduce connecting weights, the middle layer of every subnetwork has been reduced to 3 hidden nodes in the experiment. As a result of this, the overall performance is not affected.

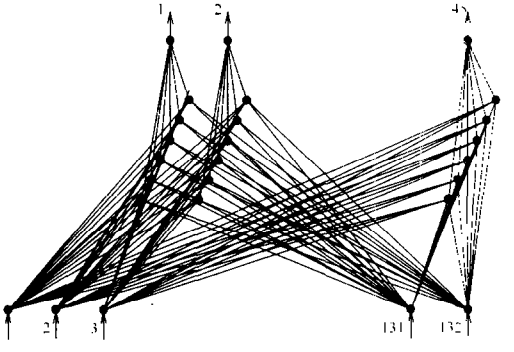

Fig. 1. The compound structure of the basic network of the hierarchical neural networks

The structure of the selection network is the same as the original one . After adopting such compound structure of hierarchical neural networks, the recognition rate to the test templates rises evidently. The recognition rate to the same 1299 templates which do not join in training is 91.5%.

3. APPLICATION OF FUZZY LOGIC

1299 templates, which do not join in training and are collected from 45 persons, are put to tests for above mentioned hierarchical neural networks. It is founded that quite a number of the mistakes occur among the easy-to-confuse words divided into several groups. In order to solve this difficulty, a tone test to some utterances is carried out. Furthermore, a set of measures on fuzzy processing and synthetical judgment to the output results of the neural networks and the tone test results is taken to reach final decisions.

First, the correctness of the recognition results of the hierarchical neural networks is processed as a fuzzy concept. Suppose maximum output value of the neural networks is q1, the corresponding word is b1; the next maximum output value is q2, the corresponding word is b2. Given x1= q1-q2, so x1∈[0,9]. For every group of the confusion words, the membership function of each fuzzy set in domain [0,9] is derived by statistical method. For example, for the perceptron structure of hierarchical neural networks, the membership function $\mu_{b1}(x1)$ of fuzzy set "b1 is correct" on confusion pair 2 and 8 is just shown in Fig. 2, and the membership function of fuzzy set "b2 is correct" on this confusion pair is

$$\mu_{b2}(x1) = 1 - \mu_{b1}(x1)$$

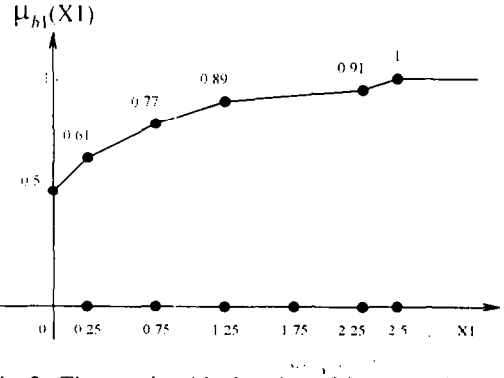

Fig. 2. The membership function of fuzzy set "b1 is correct" on confusion pair 2 and 8 for the perceptron structure of hierarchical neural networks

Second, the membership functions of every sort of tones are derived based on experiments and statistical method. For example, the membership function of the first sort of tones is shown in Fig. 3.

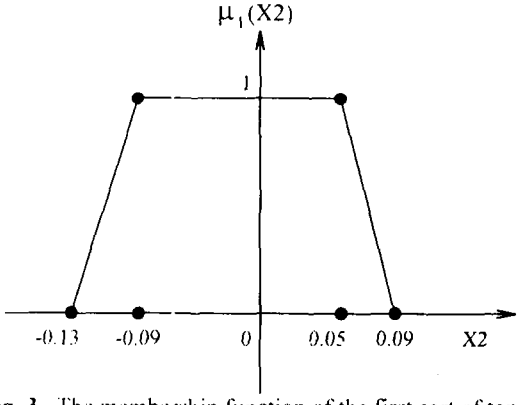

Fig. 3. The membership function of the first sort of tones

When words b1 and b2 which respectively correspond to maximum output and the next maximum output of the neural networks are the confusion words in same group, the pitch detection should be carried out and the obtained result together with the result obtained from the neural networks is used to perform a fuzzy synthetical judgment. Each membership function yet must be differently weighted according to the reliability of tone test of the corresponding word when the fuzzy synthetical judgment is performed. The weight values are obtained from great numbers of experiments.

After the fuzzy processing is adopted, although a tiny minority of templates that has originally been recognized correctly by the neural networks is judged wrong owing to the mistake of tone sorting, a large majority of templates that are judged wrong by the neural networks owing to the confusion of spectrum characteristics are finally recognized correctly owing to the right tone detection, so the total effect is that the recognition rate of the whole system is raised.

4. COMPUTER SIMULATION RESULTS OF THE RECOGNITION SYSTEM

4.1 Template Set

1859 speech templates of digits 0-9 have been collected from 45 speakers (23 males and 22 females). These persons are from 18 provinces. Their age is from 18 to 50. These templates were collected at different times, the longest interval was two years. There often were heavier noises and interference voices in the places where the templates were collected.

4.2 Training Cases of the Neural Networks

The training of the neural networks employs the error backpropagation algorithm [10] [11]. There are 560 training templates from 14 males and 14 females. These speakers are from 14 provinces. There are two templates for every digit of every speaker. When the basic network adopts the compound structure, the 45 subnetworks are trained independently of one another. The perceptron structure and every subnetwork of the compound structure converge for all 560 training templates, which means that all the 560 templates can be recognized correctly.

4.3 Test Results of the Neural Networks

There are 1299 test templates which do not join in training. Among them 710 templates belong to the speakers who take part in training and 589 templates belong to 17 speakers who do not take part in training. These templates are put to tests for the perceptron structure and the compound structure. The results for the perceptron structure are shown in Table 1. For the compound structure, in the cases of the test templates of the persons taking part in training, the templates of the persons who do not take part in training and the meanness, the recognition rates are 92.8%, 89.8% and 91.5%, respectively.

In Table 1, row C indicates the numbers of the templates that are wrong recognized as the easy-to-confusion words and the corresponding next maximum output words are the ones which are correct or in the same group, namely the numbers of the templates that are wrong recognized by the neural networks but can be further distinguished by the tone detection. Row D indicates the rate of these confused templates among the wrong

recognized templates. Row E indicates the recognition rate in each condition. for each column it is E = (A - B) / A.

4.4 Results of the Fuzzy Synthetical Judgment

For the templates whose outputs b1 and b2 from the neural networks are the confusion words in same group, the pitch detection should follow up, and the fuzzy processing and synthetical judgment are carried out finally. Under these conditions the summary test results are shown in Table 2 and Table 3 for the perceptron structure and the compound structure, respectively. For each column in the two tables, the calculating formula of the recognition rate is C = (A - B) / A.

5. DEVELOPMENT OF A COMPUTER-INDEPENDENT HARDWARE SYSTEM

A speech recognition system that is able to have actual applications should be capable of off-line working and of performing real time speech recognition in heavier ambient noises. Such a hardware equipment which adopts above-mentioned compound structure of hierarchical neural networks and fuzzy logic has been developed. Its block diagram is shown in Fig. 4.

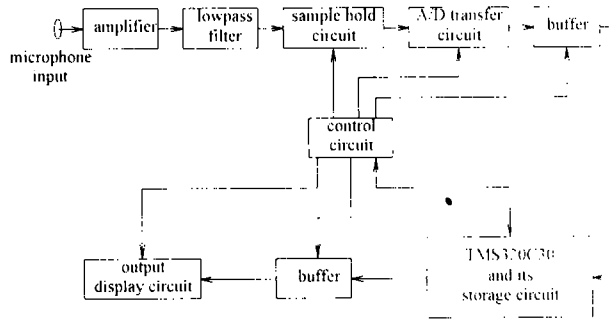

Fig. 4. The block diagram of the computer-independent hardware system

Table 1. The test results for the perceptron structure of hierarchical neural networks

items	contents	the test templates of the persons taking part in training	the templates of the persons who do not take part in training	mean cases
A	the number of templates	710	589	1299
B	the number of mistakes	79	92	171
C	the number of confusions	46	47	93
D	confusion rate	58.2%	51.1%	54.4%
E	recognition rate	88.9%	84.4%	86.8%

Table 2. The test results after adopting fuzzy logic for the perceptron structure of hierarchical neural networks

items	contents	the test templates of the persons taking part in training	the templates of the persons who do not take part in training	mean cases
A	the number of templates	710	589	1299
B	the number of mistakes	42	59	101
C	recognition rate	94.1%	90.0%	92.2%

Table 3. The test results after adopting fuzzy logic for the compound structure of hierarchical neural networks

items	contents	the test templates of the persons taking part in training	the templates of the persons who do not take part in training	mean cases
A	the number of templates	710	589	1299
B	the number of mistakes	33	47	80
C	recognition rate	95.4%	92.0%	93.8%

The hardware system is composed of three sections, that is input circuits, output circuits, TMS320C30 and its storage circuits. TMS320C30 is a high-speed digital signal processor. It can perform floating point operations. Above-mentioned speech recognition system first runs under the developing system of TMS320C30 with its assembly language program. The recognition results to above-mentioned 1299 test templates are the very same as that by high-level language program. Then, the TMS320C30 system is separated from the computer and runs independently. The requisite storage capacity is about 23k words. 3k words of which are program words and 20k or so are various data, including the weight values of the neural networks. On the input section, the speech signal is input from a microphone, then amplified, filtered by a lowpass filter, sampled and held, and A/D transformed. After passing a buffer, the obtained binary numbers go into the storage area of TMS320C30. Then, the TMS320C30 runs according to the computing program. The recognition result passes a output buffer and is shown by a Nixie light circuit finally. The recognition time to an utterance is less than one second.

The computer-independent speech recognition system is put to a noisy field test. The test result for Chinese digits 0-9 to 70 persons from 18 provinces shows that a recognition rate of 91% is reached.

6. DISCUSSION AND CONCLUSION

The compound structure of hierarchical neural networks proposed in this paper is superior to the original perceptron structure in performance. For one thing, it is about the training of neural networks. On the condition of pairwise classification, the similar or different degree of spectrum characteristics of two words in every pair is very distinct for a pair from another. Trained as a whole, a multi-layer perceptron converges only if all output errors are lower than a determined threshold. Because of serious crossed effects between the outputs and the weights, the multi-layer perceptron is very difficult to converge. However, so far as the compound structure is concerned, because its each subnetwork is trained independently of one another, and every subnetwork has only one output node, the structure is comparatively easy to converge. In fact, in the case of the same training template set and the same error threshold, the total training time for 45 subnetworks of the compound structure is still much less than the training time for the perceptron structure. For another thing, it is known by comparing Table 1 with Table 2 that, as the compound structure is used, the mean recognition rate is 4.7% higher than the perceptron structure; and for the templates of the persons who do not take part in training , the increment of the recognition rate is 5.4%, it is more than the one (3.9%) for the test templates of the persons taking part in training. These results show that both the discrimination ability and the generalized ability of the compound structure are prior to the perceptron structure.

The number of the connecting weights of the compound structure, of course, is greater, and the quantity of calculation is also larger. These mean that memory capacity needs increasing and calculative time is probably longer. But at present under the condition that high-speed computers and microprocessors are popular, it is no matter with the two effects. In fact, on the calculative time, when the speech recognition system is implemented with a high-speed digital signal processor, the difference of recognition time between the two structures is not detected at all. Actually, the compound structure of hierarchical neural networks embodies better some advantages of neural networks, such as larger-scale parallel processing ability and fault freedom.

Moreover, the speech recognition system adopts fuzzy logic, so the recognition rate is further raised. It is known by comparing Table 1 with Table 4 that, if both the compound structure of hierarchical neural networks and fuzzy logic are adopted, the recognition rate will increase by 7% compared with the original perceptron structure of hierarchic neural networks.

The hardware system implemented by the compound structure of hierarchical neural networks and the fuzzy logic algorithm is able to separate itself from the computer and to perform speaker-independent real time speech recognition in heavier ambient noises.

7. REFERENCES

[1] Hertz J. et al. Introduction to the theory of Neural Computation. Addison-Wesley Publishing Company, 1991.

[2] Waibel A., Hanazawa T., Hintom G., Shikano K., and Lang K. J. "Phoneme recognition using time - delay neural networks", IEEE Trans. Acoust., Speech, Signal Processing, Vol. 37, No. 3, pp. 328-339, Mar. 1989.

[3] Meng H. M. and Zue V. W. "A comparative study of acoustic representations of speech for vowel classification using multi-layer perceptrons", in Proc. Int. Conf. Spoken Language Processing, Kobe, Japan, Nov. 1990, pp. 1053-1056.

[4] Simpson P. K. "Fuzzy min-max neural networks--Part 1: Classification", IEEE Trans. on Neural Networks, Vol. 3, No. 5, pp. 776-786, Sep. 1992.

[5] Watrous R. L. "Speaker normalization and adaptation using second-order connectionist networks", IEEE Trans. on Neural Networks, Vol. 4, No. 1, Jan. 1993, pp. 21-30.

[6] Kwan H. K. and Cai Y. "A fuzzy neural network and its application to pattern recognition", IEEE Trans. on Fuzzy Systems, Vol. 2, No. 3, Aug. 1994, pp. 185-193.

[7] Picone J. W. " Signal modeling techniques in speech recognition", Proc. IEEE, Vol. 81, No. 9, Sep. 1993, pp. 1215-1247.

[8] Rabiner L. R. and Schafer R. W. Digital Processing of Speech Signals, Prentice-Hall, Inc., 1978.

[9] *Kämmerer* B. R. and *Küpper* W. A."Design of hierarchical perceptron structures and their application to the task of isolated-word recognition", Proc. IJCNN'89, 1989, I-243.

[10] Rumelhart D. E. et al. "Learning internal representations by error propagation ", Parallel Distributed Processing: Explorations in the Microstructures of Cognition, Vol. 1: Foundations, Combridge, MA: MIT Press, 1986, pp. 318-362.

[11] Simpson P. Artificial Neural Systems: Foundations, Paradigms, Applications and Implementation, Elmsford, NY: Pergamon, 1990.