6

Using Genetic Algorithm to Improve the Performance of Speech Recognition Based on Artificial Neural Network

Shing-Tai Pan[1], Chih-Chin Lai[2]
National University of Kaohsiung[1], National University of Tainan[2]
Taiwan

1. Introduction

The development for speech recognition system has been for a while. The recognition platform can be divided into three types. Dynamic Time Warping (DTW) (Sakoe, 1978), the earliest platform, uses the variation in frame's time for adjustment and further recognition. Later, Artificial Neural Network (ANN) replaced DTW. Finally, Hidden Markov Model was developed to adopt statistics for improved recognition performance.

Besides the recognition platform, the process of speech recognition also includes: recording of voice signal, point detect, pre-emphasis, speech feature capture, etc. The final step is to transfer the input sampling feature to recognition platform for matching.

In recent years, study on Genetic Algorithm can be found in many research papers (Chu, 2003a; Chen, 2003; Chu, 2003b). They demonstrated different characteristics in Genetic Algorithm than others. For example, parallel search based on random multi-points, instead of a single point, was adopted to avoid being limited to local optimum. In the operation of Genetic Algorithm, it only needs to establish the objective function without auxiliary operations, such as differential operation. Therefore, it can be used for the objective functions for all types of problems.

Because artificial neural network has better speech recognition speed and less calculation load than others, it is suitable for chips with lower computing capability. Therefore, artificial neural network was adopted in this study as speech recognition platform. Most artificial neural networks for speech recognition are back-propagation neural networks. The local optimum problem (Yeh, 1993) with Steepest Descent Method makes it fail to reach the highest recognition rate. In this study, Genetic Algorithm was used to improve the drawback.

Consequently, the mission of this chapter is the experiment of speech recognition under the recognition structure of Artificial Neural Network (ANN) which is trained by the Genetic Algorithm (GA). This chapter adopted Artificial Neural Network (ANN) to recognize Mandarin digit speech. Genetic algorithm (GA) was used to complement Steepest Descent Method (SDM) and make a global search of optimal weight in neural network. Thus, the performance of speech recognition was improved. The nonspecific speaker speech recognition was the target of this chapter. The experiment in this chapter would show that the GA can achieve near the global optimum search and a higher recognition rate would be

Source: Robust Speech Recognition and Understanding, Book edited by: Michael Grimm and Kristian Kroschel,
ISBN 987-3-90213-08-0, pp.460, I-Tech, Vienna, Austria, June 2007

obtained. Moreover, two method of the computation of the characteristic value were compared for the speech recognition.
However, the drawback of GA used to train the ANN is that it will waste many training time. This is becasue that the numbers of input layer and output layer is very large when the ANN is used in recognizing speech. Hence, the parameters in the ANN is emormously increasing. Consequently, the training rate of the ANN becomes very slow. It is then necessary that other improved methods must be investigated in the future research.
The rest of this chapter is organized as follows. In section 2, the speech pre-processing is introduced. Then, in section 3 we investigate the speech recognition by ANN which is trained by genetic algorithm to attain global optimal weights. Section 4 present the experiment result of the speech recognition. Finally, in section 5, we make some conclusions about this chapter.

2. Speech pre-processing

The speech signal needs be pre-processed prior to entering the recognition platform. The speech pre-processing includes point detection, hamming windows, speech feature, etc. Each process is illustrated as follows.

2.1 Fixed-size frame and Dynamic-size frame

With fixed-size frame, the number of frame varies with speech speed due to different length of voice signal. This problem does not exist in DTW recognition system. However, this study with artificial neural network ANN has to use dynamic-size frame to obtain a fixed number of frames. There are two methods to get a fixed number of frames: (1) dynamic numbers of sample points (2) dynamic overlap rates (Chen, 2002). Either of the two methods can lead to a fixed numbers of frames, meeting the requirement by ANN recognition system.

2.2 Point Detection

A voice signal can be divided into three parts: speech segment, silence segment and background noise. How to differentiate between speech segment and silence segment is called point detection. After removal of unnecessary segments, the time frame in comparison is narrowed and the response time is shortened.
There are a number of algorithms for speech end point detection. In general, there are three types based on parameters: (1) time domain point detection method, (2) frequency domain point detection method, (3) hybrid parameters point detection method. Among the three, the time domain point detection is the simplest and the most frequently used, but has the shortcoming of lower noise resistance. On the other hand, frequency domain point detection and hybrid parameters point detection have higher noise resistance and better accuracy, but is more complicated in calculation. In this chapter, we adopted the time domain point detection method for shortening the computation time.

2.3 Hamming Window

The purpose to fetch hamming window is to prevent discontinuity in every frame and both ends of every frame. When the voice signal is multiplied by hamming window, we can reduce the effect of discontinuity (Wang, 2004).

2.4 Feature capture

In general, there are two types of features capturing of voice signal for speech recognition, time domain analysis and frequency domain analysis. Of the two, time domain analysis is more straightforward and involves less calculation, so it saves time. Frequency domain analysis needs to go through Fourier transform first, so it involves more calculation and complexity and takes more time. Common speech features capturing include Linear Predict Coding (LPC) (Chen, 1994), linear predict cepstrum coefficient (LPCC), Mel-frequency Cepstrum coefficient (MFCC) (Chu, 2003b) etc. With consideration of recognition accuracy, this study selected MFCC as the speech feature capturing method.

Using MFCC to obtain solutions involves three steps: 1. Using Fast Fourier Transform (FTT) to obtain power spectrum of the speech signal; 2. Applying a Mel-space filter-bank to the power spectrum to get logarithmic energy value; 3. We conduct the discrete cosine transform (DCT) of log filter-bank energies to obtain MFCC.

3. Speech recognition platform

BPNN is the most commonly used structure in ANN. Although ANN has fast recognition rate and fault tolerance, it is not perfect because its SDM has a problem with Local Optimum. To prevent this from happening, GA is adopted to assist in SDM for obtaining optimal weight and improved recognition performance.

3.1 Back-propagation neural network

In principle, back-propagation neural network uses Multiple-Layer Perception as system framework and SDM as training rule. Such a system is called back-propagation neural network. Multiple-layer in Multiple-Layer Perception model indicates it is composed of many layers of neurons. Besides, the signal transmittance mode between neurons in two layers is the same as that for a single layer. The study adopted three-layer structure (input layer, hidden layer, output layer) as the basic framework for speech recognition, which is depicted in Fig. 1.

3.2 Genetic algorithm

Genetic algorithm (Goldberg, 1989; Michalewicz, 1999) involves Selection, Reproduction and Mutation.

The purpose of selection is to determine the genes to retain or delete for each generation based on their degree of adaptation. There are two types of determination: (1) Roulette-Wheel Selection (2) Tournament Selection. The study adopted tournament selection. It is to follow their fitness function sequence for each gene set to determine whether they are retained. The fittest survives. Reproduction is a process to exchange chromosomes to create the next generation according to distribution rule. In general, there are one-point crossover, two-point crossover and uniform crossover, etc. The evolutionary process of GA is depicted in Fig. 2.

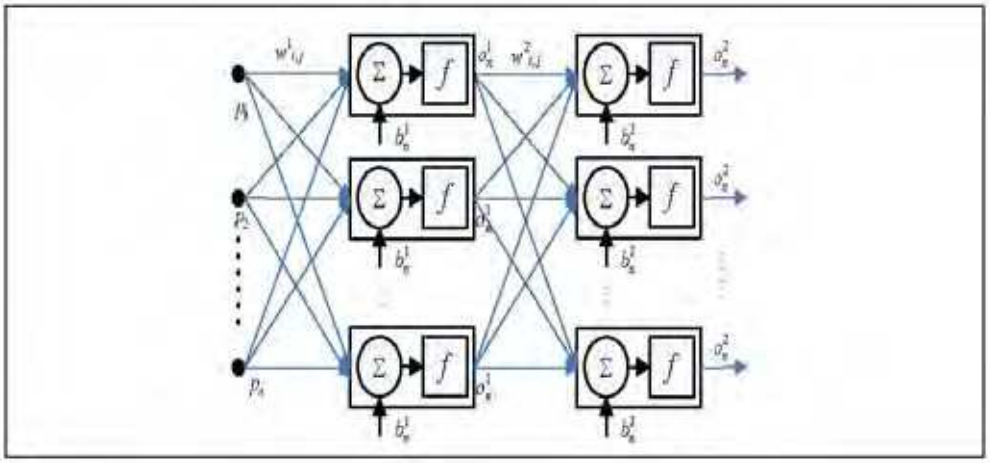


Figure 1. The three-layer structured ANN

The way of mutation is not very different from that for crossover. There are one-point mutation and two-point mutation. Mutation process depends on conditions; for example, mutation can start as adaptation function and stop changing after several generations. Mutation rate cannot be too high. Otherwise, convergence will not occur.

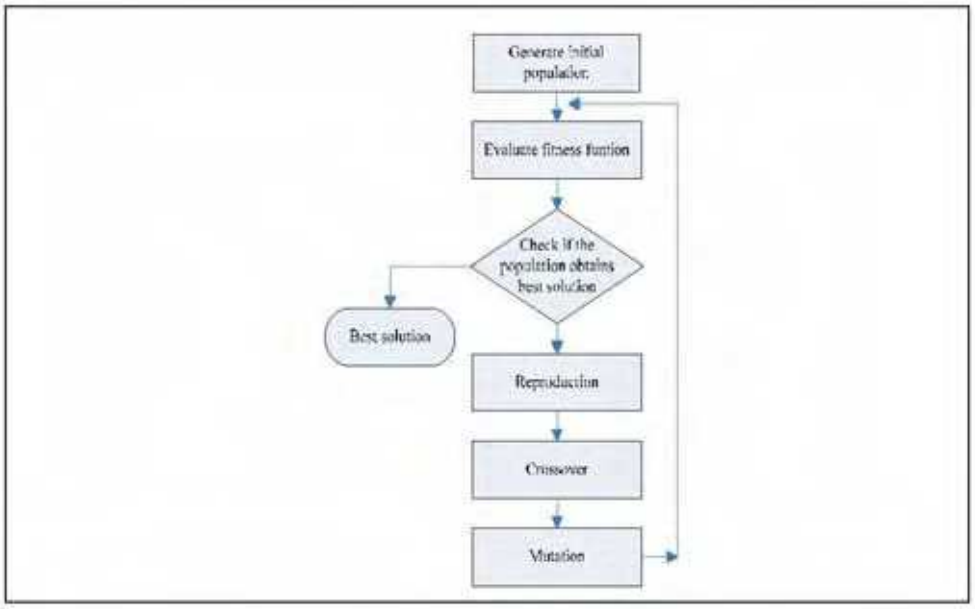


Figure 2. The evolutionary process of GA

4. Experimental results

ANN with SDM training (i.e., BPNN) was first adopted to be speech recognition platform. The result was compared to GA assisted ANN platform. Although genetic algorithm has advantages that SDM cannot provide. It has one drawback, i.e. the crossover, reproduction and mutation process needs more time to seek optimal solutions.

In the initial speech, voice files for ten people were collected. This file contains the voice of numeral 1 ~ 9. Each people recorded four sets, three of which were used for training and one of which was for test. Recording format had sampling frequency 8 kHz, mono channel 16 bit sampling point. After recording, point detection was to remove silence segment, followed by pre-emphasis. Then, speech segment was divided to 20 fixed frames. Feature parameter was extracted from each frame. One frame had 10 features. Thus, each number would have 200 features.

In the aspect of speech frame, because ANN recognition platform was adopted, speech segment needs the same number of frames regardless of its length of time. The adopted dynamic-size frame was different from the DTW fixed frame. The study also adopted dynamic sampling point (fixed overlap rate). The frame sampling point can be expressed by the following equation (1) (Chen, 2002):

$$L = \mathrm{Fix}\{\, l_s / [(N\text{-}1)(1\text{-}R)+1] \,\} \tag{1}$$

L represents the number of frame sampling points, l_s is the total signal length, N is the number of frames, R is overlap rate (%), while Fix(x) the maximum integer smaller than x. Through such function, the same number of frames can be obtained for different length of speech. Besides, the point detection in the study still adopted time-domain point detection due to the reason of less calculation load. It took the average of energies for the first few silence segments (background noise) and added it to 5%~10% of the maximum frame energy to set the threshold value for point detection, as shown in the following equation (2) (Chen, 2002):

$$Threshold = 7.5\% \times \max[E(n)] + \frac{1}{K}\sum_{i=1}^{K} E(i), \quad 1 \le n \le \tilde{N} \tag{2}$$

Threshold is the point detection threshold value. N is the number of frames prior to point detection. E(n) is the sum of energies for sampling points in the nth frame prior to point detection. K represents captured number of silence segment frames. E(i) is the sum of energies for sampling points in ith silence segment. 7.5% of the maximum frame energy was added with the average energy for five silence segment frames (K = 5) to establish point detection threshold calculation.

After completion of point detection, pre-emphasis and Hamming window were carried out to capture features. With consideration of recognition accuracy, MFCC parameter was adopted for recognition. MFCC level in this experiment was 10. After obtaining features, they were input to recognition platform to start speech recognition.

4.1 BPNN Experiment Results

In the design structure of artificial neural network, there are two output modes. One used binary coding to express output, for example, the system has 16 corresponding outputs to four output neurons. Thus, the number of output neurons was reduced. The other was one-

to-one output. For example, 9 frames needed 9 outputs neurons. Although binary coding can minimize the number of neurons, it not only had low recognition rate, but difficulty in convergence after experiment comparison with one-to-one mode. Therefore, one-to-one output was adopted here. The entire ANN structure had 200 inputs, 30 neurons in hidden layer and 9 neurons in output layer.

In the experiment, each speech had 20 frames. Each frame had ten levels of features, indicating 200 input parameters as input layer, while 30 neurons were in the hidden layer. The number of neurons cannot be too many; otherwise, it cannot obtain convergence. If the number is too small, recognition error will be large. The number of neurons in the hidden layer (N_no) is expressed by the following equation (3) (Yeh, 1993) :

$$N\_no = (\text{In\_number} \times \text{Out\_number})^{1/2} \tag{3}$$

In_number represents the number of input layer units, while Out_number represents the number of output layer units.

Because the number of ANN inputs is as high as 200, there will be a problem, i.e. when the input approaches a relatively large number, the output will be at the extreme condition. To solve the problem, the speech features were all downsized prior to input. It was to multiply the speech feature by 0.01 to prevent transfer function from over-saturation. From Fig. 3, it is known that once the number of training generation is over 1,500, it then fails to breakthrough and the recognition rate is 91% for the existing database. Even further training is continued, root mean square error does not progress, and recognition rate does not improve either. Under this situation, genetic algorithm is used to assist in seeking the weight.

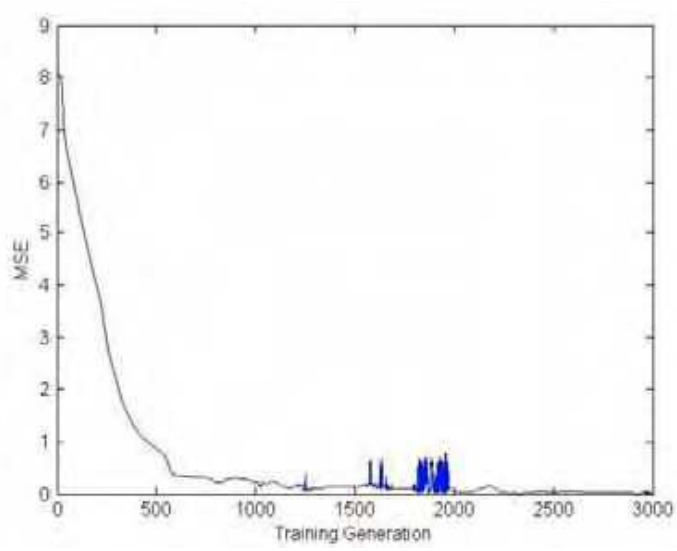


Figure 3. MSE for Output in ANN with SDM Training Process

4.2 Experiment Results with Genetic Algorithm to assist in Training

SDM in the above-mentioned artificial neural network was used to seek the weight. Back-propagation neural network tended to have problems with local optimum, so genetic algorithm was used to seek the weight for the entire domain (Chu, 2003a).

At first, the study used SDM to train ANN to obtain the weight. Then converged weight and bias were entered into genetic algorithm. It improved the initial speed and also helped SDM out of the local optimum. Fig. 4 shows the entire training structure. The experiment has shown that SDM training followed by GA training would help error break the SDM limit and greatly improve recognition rate. Fig. 5 shows the error range for SDM training followed by GA.

Through 3000 generations of GA, MSE value drops to 0.001 with recognition rate up to 95%, as shown in Table 1.

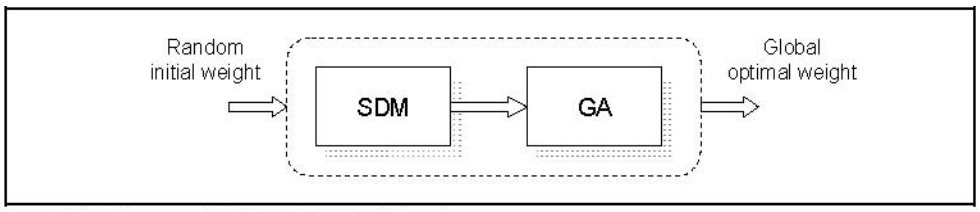


Figure 4. The Proposed ANN Training Structure

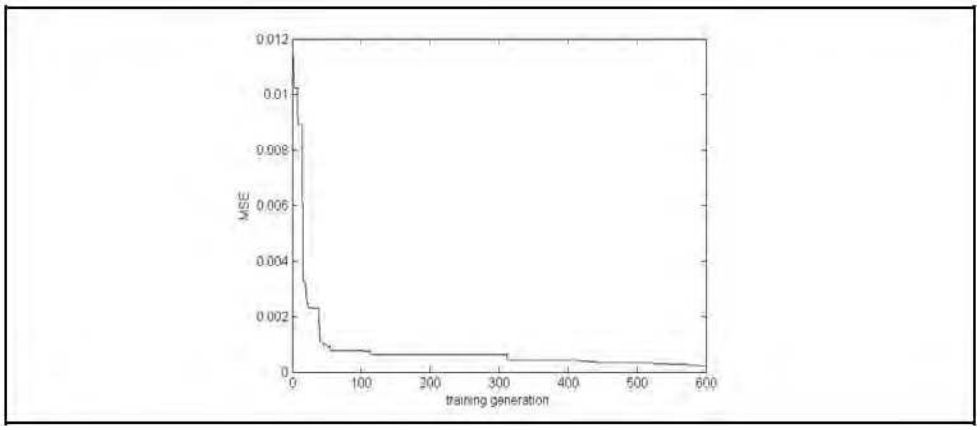


Figure 5. MSE for Output in Genetic Algorithm Training after SDM Training

5. Conclusion

In this chapter, we have seen that the recognition rate through the SDM in BPNN is up to 91% under the MFCC feature. This recognition rate is not the optimum because that the SDM can always get local optimum. To solve this problem, GA was adopted and following SDM to improve MSE. By this two stage (SDM then GA) training scheme, the recognition rate can be increasing up to 95%. However, under the condition of adopting only MFCC parameters, speech recognition rate still has room for improvement. For the future, other

modes of features capturing method can be adopted, such as Cepstrum Coefficient or LPC parameter together with pitch parameter, to improve recognition rate.

Speech	Successful Recognition	Failed Recognition	Recognition Rate(%)	Total Recognition Rate(%)
1	10	0	100	
2	10	0	100	
3	7	3	70	
4	10	0	100	
5	10	0	100	95
6	9	1	90	
7	9	1	90	
8	10	0	100	
9	10	0	100	

Table 1. Speech Recognition Results after 10 sets of testing

6. References

Chen, S. C. (2003). Use of GA in CSD Coded Finite Impulse Digital Filter (FIR), Shu-Te University, MS Thesis, Taiwan.

Chen, M. Y. (1994). PC Computer Voice Operation, Chi Biao Publication, 957-717-071-4, Taiwan.

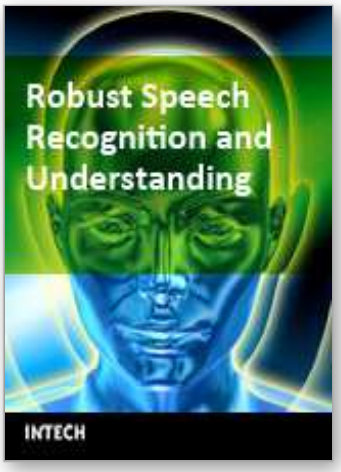
Chen, S. L. (2002). Speech Recognition based on Artificial Neural Network, National Sun Yat-Sen University, Master Thesis, Taiwan. Chu, S. H. (2003a). Combination of GA and SDM to Improve ANN Training Efficiency, Shu-Te University, MS Thesis, Taiwan. Chu, W. C. (2003b). Speech Coding Algorithms, John Wiley & Sons, 978-0-471-37312-4, USA.

Demuth, H. B., Beale, M. H., Hagan, M. T. (1996). *Neural Network Design*, Thomson Learning, 0534943322, USA. Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley Professional, 0201157675, Reading, Massachusetts. Michalewicz, Z. (1999). Genetic Algorithms + Data Structures = Evolution Programs, Springer-Verlag, 3540606769, Berlin.

Sakoe, H. and Chiba, S. (1978). Dynamic Programming Optimization for Spoken Word Recongition, *IEEE Transactions on Signal Processing*, Vol. 26, pp 43- 49.

Wang, S. C. (2004). Voice Signal Processing, Chun Hwa Publication, 9572143840, Taiwan.

Yeh, Y. C. (1993). Implementation and Application of Artificial Neural Network, Ru Lin Publication, 957499628X, Taiwan.



Robust Speech Recognition and Understanding

Edited by Michael Grimm and Kristian Kroschel

ISBN 978-3-902613-08-0
Hard cover, 460 pages
Publisher I-Tech Education and Publishing
Published online 01, June, 2007
Published in print edition June, 2007

This book on Robust Speech Recognition and Understanding brings together many different aspects of the current research on automatic speech recognition and language understanding. The first four chapters address the task of voice activity detection which is considered an important issue for all speech recognition systems. The next chapters give several extensions to state-of-the-art HMM methods. Furthermore, a number of chapters particularly address the task of robust ASR under noisy conditions. Two chapters on the automatic recognition of a speaker's emotional state highlight the importance of natural speech understanding and interpretation in voice-driven systems. The last chapters of the book address the application of conversational systems on robots, as well as the autonomous acquisition of vocalization skills.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821