정보저장시스템학회논문집, 제 6 권 제 1 호, pp. 18~24

소프트컴퓨팅 기법을 이용한 다음절 단어의 음성인식

Speech Recognition of Multi-Syllable Words Using Soft Computing Techniques

이종수†, 윤지원*

Jongsoo Lee and Ji Won Yoon

(2010 년 3 월 12 일 접수; 2010 년 3 월 22 일 심사완료; 2010 년 3 월 23 일 게재확정)

Abstract

The performance of the speech recognition mainly depends on uncertain factors such as speaker's conditions and environmental effects. The present study deals with the speech recognition of a number of multi-syllable isolated Korean words using soft computing techniques such as back-propagation neural network, fuzzy inference system, and fuzzy neural network. Feature patterns for the speech recognition are analyzed with 12th order thirty frames that are normalized by the linear predictive coding and Cepstrums. Using four models of speech recognizer, actual experiments for both single-speakers and multiple-speakers are conducted. Through this study, the recognizers of combined fuzzy logic and back-propagation neural network and fuzzy neural network show the better performance in identifying the speech recognition.

Key Words: Speech Recognition Algorithm(음성인식알고리즘), Neural Network(신경회로망), Fuzzy Logic(퍼지논리), Fuzzy Neural Network(퍼지신경망), Multi-Syllable Word(다음절단어)

1. Introduction

Speech is one of the most practical and easiest ways for the communication of information. The communication through speech or voice would be sometimes the only way to hands-disabled persons who have hard time in manipulating automated devices such as wheelchair, automobile door and window, etc. There have been a number of efforts in the development of speech recognition systems and algorithms [1,2]. According to this trend, speech recognition algorithms using hidden Markov model and dynamic time warping have been studied. The performance and accuracy of speech signals depend on speaker's physical/psychological conditions, time, noise and other environmental effects. Such uncertain factors would be handled with intelligent soft computing techniques. Artificial neural network and fuzzy logic theory have received considerable attention in the speech recognizer since the latter can deal with uncertain linguistic expressions and the former can accommodate stationary patterns along with parallel processing. The objective of the present study is to realize the speech feature pattern recognition of a number of multi-syllable isolated Korean words using intelligent soft computing methods such as back-propagation neural network (BPN), fuzzy inference system (FIS), and fuzzy neural network (FNN). There are four different architectures suggested for the speech recognition of multi-syllable Korean words; the first is a back-propagation neural network (BPN), the second is the if-then rule based fuzzy inference system (FIS), the third is a combined method of fuzzy inference system and back-propagation neural network (FIS+BPN) by which the pattern recognition can be trained faster and easier. The construction of a conventional fuzzy inference system requires much of efforts due to the manual rule matching processes so that a fuzzy neural network (FNN) is employed as a fourth approach, wherein self-membership function parameters are adjusted by using delta rule based error back-propagation process to avoid the conventional manual rule matching efforts and gain logically adjusted membership function parameters.

In the present study, a total of 8 multi-syllable Korean words for use in the wheelchair application are employed as a test-bed in speech recognition. Feature patterns for the speech recognition are analyzed with 12th order thirty frames that are normalized by the linear predictive

† 연세대학교 기계공학부
E-mail: jleej@yonsei.ac.kr
TEL: (02) 2123-4474
* 연세대학교 대학원 기계공학과

coding (LPC) and Cepstrums. Using actual voice signals obtained from both single-speakers and multiple-speakers, proposed soft computing based algorithms are examined to predict the speech recognition capabilities.

2. Features for Speech Recognition

The speech recognition system is a kind of pattern recognition, and generally consists of following three levels: the first is the speech signal pre-processing level, the second is the speech feature extraction level, and the last is the pattern classification and recognition level [2]. The speech feature extraction is a process extracting usable speech recognition components from speech signals. In the present study, 12th order thirty frames normalized by LPC & Cepstrum are considered [3]. A wheelchair application for the hand-disabled person is considered as a test-bed in the present study. As an example of Cepstrum pattern extracted by the above procedure, a three-syllable Korean word, “A-pu-ro” that means “move forward” is shown in Fig. 1. The speech feature patterns have a certain level of variations, even though they are extracted from the same word due to voice variation and environmental effects. Such diversities and uncertainties make a speech recognizer difficult to generate accurate results. Intelligent soft computing methods such as artificial neural networks and fuzzy logic facilitates are considered to effectively handle diversities and uncertainties in speech recognition. The target words to be recognized in the present study are summarized in Table 1. Cepstrum results of target words (obtained from multiple-speakers) are demonstrated in Figs. 1 to 8, wherein horizontal and vertical axes denote the number of frames and the strength of voice signals, respectively.

Table 1 Target words for wheelchair application

Function	Target word	Meaning
Movement	A-pu-ro	Move forward
	Jeong-ji	Stop
Direction	Dol-a-ra	Turn back
	Jowa-ro	Turn left
	U-ro	Turn right
Speed	Bbal-li	Fast
	Cheon-cheon-hi	Slow
	Bo-tong	Medium

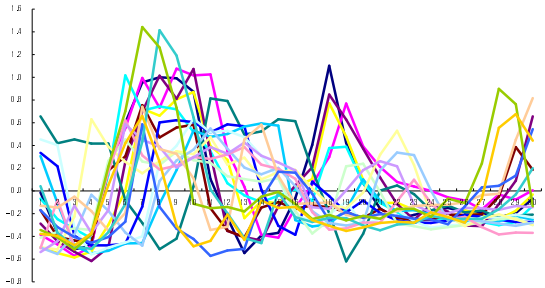

Fig.1 Cepstrums for A-pu-ro

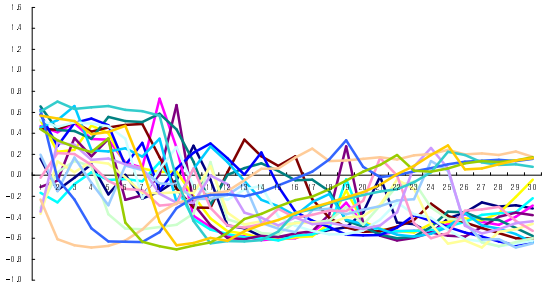

Fig. 2 Cepstrums for Dol-a-ra

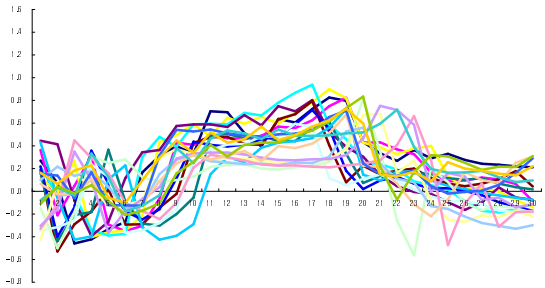

Fig. 3 Cepstrums for Jeong-ji

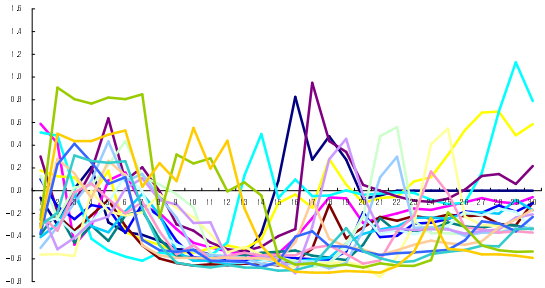

Fig. 4 Cepstrums for Jowa-ro

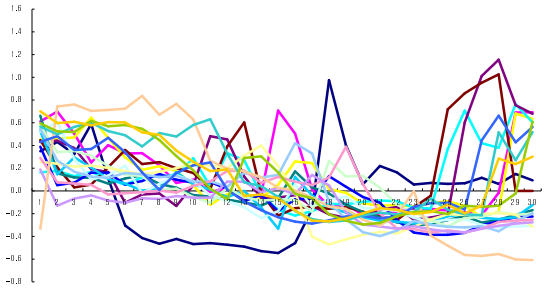

Fig. 5 Cepstrums for U-ro

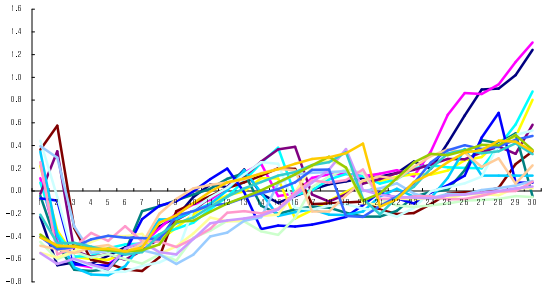

Fig. 6 Cepstrums for Bbal-li

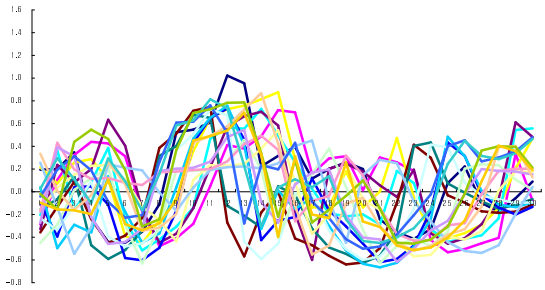

Fig. 7 Cepstrums for Cheon-choen-hi

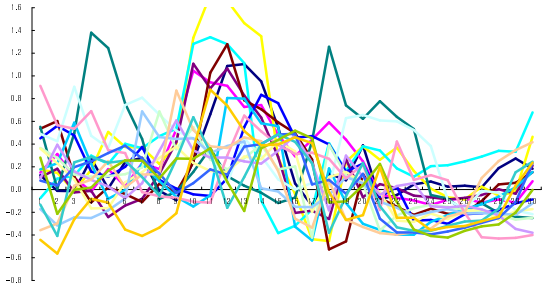

Fig.8 Cepstrums for Bo-tong

3. Soft Computing Based Algorithms

3.1 Back-Propagation Neural Network

Three-layered BPN architecture is employed to generate approximate response surfaces for the speech recognition. A number of input data used to train and test the BPN are 12th order thirty frames normalized by Cepstrums of target words as shown in Table 1. Numerical values of each frame are assigned to input nodes in the network. It is noted that thirty frames correspond to 30 neurons in the input layer. Output target values are zero and one as shown in Table 2. In a case where Cepstrums are a feature of "A-pu-ro", the output target of Y_1 to Y_8 is a vector of [1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]. Therefore, there are a total of 8 neurons (i.e., 8 target words in Table 1) in the output layer.

Generally, it is very difficult or even impossible to exactly predict an output target as 1.0. So, the present study assumes that the approximated target value that is larger than 0.8 can be considered as the exact value of 1.0. The training efficiency of BPN is determined by a number of control parameters. The first one is the number of neurons in the hidden layer(s), and the second one is the learning rate. The present study uses a three-layered network, which means there is a single hidden layer. For two control parameters, the best number of neurons in a hidden layer is selected from 10, 20, 30 or 40, and the learning rate for the best BPN training is selected between 0.1 and 0.9 with an interval of 0.1. That is, there are a total of 36 combinations of BPN training to determine the best values of the number of hidden neurons and learning rate. The best result was selected by considering both average recognition rates and individual recognition rates. First, choose the highest average recognition rate among the average recognition rates. If the average recognition rates were similar then the deviations of the individual recognition rates were taken into consideration. If the deviations were smaller then it was considered as the better result.

Table 2 Target values for use in BPN

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8
A-pu-ro	1	0	0	0	0	0	0	0
Dol-a-ra	0	1	0	0	0	0	0	0
Jeong-ji	0	0	1	0	0	0	0	0
Jowa-ro	0	0	0	1	0	0	0	0
U-ro	0	0	0	0	1	0	0	0
Bbal-i	0	0	0	0	0	1	0	0
Cheon-cheon-hi	0	0	0	0	0	0	1	0
Bo-tong	0	0	0	0	0	0	0	1

3.2 Fuzzy Inference System

Fuzzy logic enables to make it possible to handle not only numerical expressions but also uncertain and linguistic expressions [4]. Fuzzy theory can be used to

take care of vagueness and complexness in the speech features. Among the various types of fuzzy logic and inference systems, Mamdani's min-max operation based inference model is used to establish the speech recognizer of FIS. The rules used to organize the systems are given critically, just as the concept of a look-up table, by analyzing every training data used. So, the number of rules used varied proportionally by the number of training data. The training data are gained by calculating the average values of the raw thirty frame normalized Cepstrums by three frames. That means X_1 subjected to the premise of the fuzzy inference system is the average value of the Cepstrums from frame one to three. In this study, the number of the rules is varied between 40,960 and 163,840 by the conditions of speech recognition experiments. All fuzzy membership functions used in this study are Gaussian functions. A number of premise membership functions are given between nineteen and twenty four for every input variable. The number of conclusion membership functions is fixed at eight, same as the number of targeting words to be recognized. Centroid of area (COA) is used to defuzzify the fuzzy output. The output COA has to match the value which represents the targeting word to be considered as the word that the output value represents. Table 3 shows the values which represent the targeting words.

Table 3 Numerical values for representing target words

word	value	word	value
A-pu-ro	10	U-ro	50
Dol-a-ra	20	Bbal-i	60
Jeong-ji	30	Cheon-cheon-hi	70
Jowa-ro	40	Bo-tong	80

Table 4 Typical example of COA from BPN data

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	35.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	39.9	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	49.5	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	59.6	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	69.9	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	79.8

3.3 Combination of FIS and BPN

The objective of this recognizer is to enhance the speech recognition performance by utilizing advantages of both FIS and BPN. In the recognizer of FIS+BPN, the FIS uses more simplified data than raw values of Ceptrums in order to learn (i.e., train and test) the BPN. Each frame is individually corresponding to a premise membership function of the Mamdani's min-max fuzzy inferences system, which is identically used in the recognizer of FIS. The values of centroid of area obtained from the defuzzification process are presented in Table 4.

Such gained COA is subsequently used to learn the three-layered BPN architecture. That is, the COA is then broadcasted into input layers that correspond to 8 neurons in the BPN. For instance, when the COA value is the result of the word "A-pu-ro", the input data for training the BPN is [10, 0, 0, 0, 0, 0, 0, 0]. Accordingly, a word of "U-ro" is translated into [0, 0, 0, 0, 50, 0, 0, 0] as well. Such vectors of simplified COA values are easier and faster to train the neural network rather than raw data of Cepstrums. The target values used in FIS+BPN are the same as those in the BPN as shown in Table 2. The approximated target value that is larger than 0.8 is considered as the exact value of 1.0 as in the case of BPN.

3.4 Fuzzy Neural Network

In general, FNN has the structure of multi-layered neural network combined with fuzzy membership functions. In this structure, input layer refers the input values, output layer refers the output value and hidden layer refers the fuzzy membership functions and if-then rules. The back-propagation learning process adjusts the parameters of fuzzy membership functions, center and width of the Gaussian functions, to make fuzzy inference system suitable to fuzzy rules. These fuzzy rules are given by specialists or experts of the field and initial parameters of the membership functions are given randomly before learning process begins [5,6]. In this study, the rules are given by analyzing the training patterns critically, just as the concept of a look-up table. The number of rules varied between, 40,960 and 163,840 depending on the conditions of speech recognition experiments. The number of premise membership functions is eleven for every input variable. The number of conclusion membership functions is eight, same as the number of targeting words to be recognized. Because the rule matching process is held automatically by the back-propagation learning process, it is reasonable to use fewer premise membership functions compared with the recognizer of FIS. The initial parameters of premise membership functions are as follows. The center values are 0 to 100 with an increment of 10, and their width is chosen as 2.5. Initial parameters of conclusion membership functions are as follows. The center values are 10 to 80, increasing by 10 and their width are 4.2. And the targeting values of centroid of areas are the same as the center values of the conclusion membership functions. That is, if input-Cepstrums are extracted from a word of "A-pu-ro", then the fuzzy neural network had to be trained to generate the value 10 as the centroid of area.

Sorting the training parameters are as follows. In case of single speaker recognitions, nine different learning rates from 0.1 to 0.9, increased by 0.1, are used to train the neural networks. But in case of multi speaker recognition, training is very sensitive to learning rates, so the efficient learning rate is selected by several times of trial and error. Unlike the back-propagation neural networks, the

number of hidden layer nodes of the fuzzy neural network is fixed as the number of rules. The difference between the output COA and the target value has to satisfy a constant level, in this study. smaller than 3, to be considered as right recognition.

4. Speech Experiments

Actual speech recognition experiments are conducted to verify the performance of soft computing based speech recognizers. A total of six male persons of 25 to 28 years old have participated in recording their voices. Participants are asked to speak as naturally as possible in order to obtain more realistic, meaningful data. Every spoken word has been saved under the 8 kHz sample rate and 16 bit resolution [7]. The target words for recognition are shown in Table 1. A total of 160 training data are used to construct fuzzy inference systems and to learn back-propagation neural networks and fuzzy neural networks. The numbers of testing data for the single speaker recognition and the multiple-speaker recognition are 160 and 120, respectively.

5. Results & Discussion

5.1 Single-Speaker Results

Two cases of single-speaker experiment are conducted by changing the number of training data for each of speech recognizers. First experiment used 160 training data, which means 20 training data for each targeting words to train neural networks and construct the FIS. But recording 20 same words continually is unrealistic. So, in second experiment, considering the reality of recording the speeches, we used 40 training data. That means 5 training data for each targeting words. Both experiments used same 160 testing data, 20 data per a word, which are recorded separately with training data as mentioned before. The best recognition rates of each experiment are shown in Tables 5 and 6.

Analysis of experiment results using BPN is as follows. In case of experiment using 160 training data, training neural network with learning rate of 0.8 and twenty hidden layer nodes showed the best results. In case of experiment using 40 training data, training neural network with learning rate of 0.9 and twenty hidden layer nodes showed the best results. We could confirm that BPNs show better performance, both in average recognition rates and individual recognition rates, if the number of training data is sufficient. But even if the training data are sufficient, recognition rates of words showing similar patterns relatively such as “Jowa-ro”, “U-ro”, “Cheon-cheon-hi” and “Bo-tong”, are harder and tougher for the neural networks to recognize accurately. In case of word “Jeong-ji”, the recognition rate dramatically reduced proportionally to the number of training data compared to the other recognizers. We consider that it is because the five patterns of the word “Jeong-ji”, used to train the neural network showed less diversity compared to the other words' patterns and the basic character of BPN which can only deal with crisp numerical values and cannot deal with linguistic terms that other recognizers can.

Analysis of experiment results using FIS and FIS+BPN is as follows. Just like BPN, FIS performances are proportional to the number of the training data. We considered that it is because the number of rules to construct the fuzzy inference system is proportional to the number of the training data. Since FIS has the handling ability of uncertainties, the speech recognition performance shows less reduction proportional to the reduction in the number of training data, compared with BPN. As shown in Tables 5 and 6, the recognizer of FIS+BPN shows improved performance compared with individual architectures of BPN and FIS. However, due to the manually constructed rule matching in FIS, the recognizer of FIS+BPN is lack of logicality and time efficiency. The sorted parameters of BPN are as follows. In case of the first experiment, the learning rate is 0.75 and the hidden layer nodes are twelve, and in case of the second experiment, the learning rate is 0.75 and the hidden layer nodes are eight.

Analysis of experiment results using fuzzy neural network is as follows. In case of experiment using 160 training data, training FNN with the learning rate of 0.2 showed the best results. In case of experiment using 40 training data, training FNN with the learning rate of 0.1 showed the best results. Like other three recognizers, the recognizers using FNN are also influenced by the number of training data and pattern similarities. But, they showed better performance in recognizing similar patterns relatively. We considered that it is because of the logical computational rule matching by the computational learning process. In this study, the recognizer of FNN shows relatively more efficient single-speaker speech recognition performance than other recognizers.

5.2 Multiple-Speaker Results

Multiple-speaker speech recognition experiment is conducted once. The total number of training data is 160, that is, 20 data are used for each targeting word. The number of testing data is 120, 15 of each targeting words. To realize the reality of the experiment, just like the second single speaker speech recognition experiment, six speakers shared to spoke targeting words 2 to 4 times to gain the training and testing data. Training data and targeting data are recorded separately. The experiment results are shown in Table 7, wherein multiple-speaker speech recognition rates are far lower than the single speaker cases.

Analysis of experiment result using BPN is as follows. Training neural network with the learning rate

of 0.9 and twenty hidden layer nodes shows the best results. Like mentioned above, the recognition rates of multi speaker speech are much lower than single speakers'. Just like the case of single-speakers, recognition rates of words showing similar patterns relatively such as "Jowa-ro", "U-ro", "A-pu-ro", "Cheon-cheon-hi" and "Bo-tong", are harder and tougher for the neural networks to recognize accurately. In case of "Jowa-ro" and "U-ro", it shows opposing results compared to other three recognizers. We consider that it is because of the basic character of BPN which can only deal with crisp numerical values and cannot deal with linguistic terms that other recognizers can. Speeches with relatively distinguishable patterns such as "Bbal-li" also shows lower recognition rate than that of the single-speaker results due to their complexness and diversities.

Analysis of experiment results using FIS and FIS and BPN together is as follows. Due to the complexness and diversities of the speech features, recognizer using FIS also showed relatively lower speech recognition rates than those of the single speakers' even with the same number of rules as single speaker speech recognition. As shown in Table 7, the recognizer of FIS+BPN shows an improved performance, also in the multiple-speaker speech recognition experiments, compared to the ones using BPN and FIS separately. The learning rate of 0.75 and twelve hidden layer nodes are used to construct the BPN.

Analysis of experiment results using FNN is as follows. The sorted learning rate to train the FNN is found by trial and error and it is 0.01179. Like other three recognizers, the recognizers using FNN also shows low speech recognition rates because of the complexness and the diversities of the speech feature patterns. As a whole, all the recognizers show large deviations of recognition rate by the targeting words and low average speech recognition rates due to the similarities and complexness of the speech features, compared to the single-speaker speech recognition experiments. In the multiple-speaker experiment, the recognizers of FIS+BPN and FNN show slightly better performance in terms of average recognition rate, compared with the recognizers of both BPN and FIS.

6. Concluding Remarks

The present study explores the speech recognition of multi-syllable isolated Korean words using a total of 4 soft computing based algorithms. First two are conventional methods of back-propagation neural network and Mamdani's min-max based fuzzy inference system. The third is the consecutive combination of fuzzy inference system and back-propagation neural network. Lastly, the fourth is architecture of fuzzy neural network. Each soft computing based algorithm is validated through single-speaker and multiple-speaker experiments. A back-propagation neural network shows a dramatic decrement in speech recognition rate due to the number of training data. Fuzzy inference system requires lots of efforts and time during the manual rule matching. The combined method of fuzzy inference system and back-propagation neural network show an increment in speech recognition rate, but it still needs much of efforts and time during the manual rule matching. Unlike above three algorithms, the architecture of fuzzy neural network is able to save the conventional rule matching time and efforts along with gaining reasonable and logical self-adjusted fuzzy membership functions parameters. Through single- and multiple-speaker experiments, the recognizers of combined fuzzy logic and back-propagation neural network and fuzzy neural network show the better performance in identifying the speech recognition. In the present study, four important factors influencing the ability of a speech recognition system are also identified; they are the speech recognition algorithm, the number of training data, similarities and diversities between Cepstrum patterns.

Table 5 Single-speaker results with 40 training and 160 testing data (unit: %)

	BPN	FIS	FIS+BPN	FNN
A-pu-ro	70%	65	70	70
Jeong-ji	20	80	80	80
Dol-a-ra	75	85	85	90
Jowa-ro	90	75	80	90
U-ro	65	80	80	85
Bbal-i	90	100	100	100
Cheon-cheon-hi	85	80	85	70
Bo-tong	85	70	80	85
Average	72.5	79.4	82.5	83.8

Table 6 Single-speaker results with 160 training and 160 testing data (unit: %)

	BPN	FIS	FIS+BPN	FNN
A-pu-ro	100	90	95	100
Jeong-ji	100	100	100	100
Dol-a-ra	95	100	100	95
Jowa-ro	80	80	80	90
U-ro	70	75	80	95
Bbal-i	100	100	100	100
Cheon-cheon-hi	85	70	90	85
Bo-tong	85	80	90	80
Average	89.4	86.9	91.9	93.1

Table 7 Multiple-speaker results with 160 training and 120 testing data (unit: %)

	BPN	FIS	FIS+BPN	FNN
A-pu-ro	60	67	67	93
Jeong-ji	80	87	87	87
Dol-a-ra	80	60	60	67
Jowa-ro	73	40	40	40

U-ro	53	80	87	67
Bbal-i	73	80	87	87
Cheon-cheon-hi	47	73	80	67
Bo-tong	87	67	67	67
Average	69.1	69.3	71.9	71.9

References

[1] Oh, T. H., 1998, Speech Language Information Processing, Hong-Neung Science Publishing.

[2] Lee, H. S., 1999, Speech Recognition Technique, Chong-Moon-Gak.

[3] Rabiner, L., and Juang, B. H., 1993, Fundamentals of Speech Recognition, Prentice-Hall.

[4] Jang, J.-S. R., Sun, C.-T., and Mizutani, E., 1997, Neuro-Fuzzy and Soft Computing, Prentice Hall, Upper Saddle River, NJ.

[5] Choi, M. G., and Lee, S. B., 1996, "The Study on the Algorithm for Design of Fuzzy Logic Controller Using Neural Network," Proceedings of Fall Conference on Korean Society of Fuzzy Logic and Intelligent System, pp. 243-248.

[6] Wang, L-X., 1997, A Course in Fuzzy Systems and Control, Prentice-Hall, Upper Saddle River, NJ.

[7] Kim, J. H., Ryu, H. S., Kang, J. M., Kang S. I., Kim, K. H., and Lee, S. B., 2002, "Wheelchair System Design on Speech Recognition Function," Proceedings of Fall Conference on Korean Society of Fuzzy Logic and Intelligent System, pp. 1-5.

[8] Kasabov, N. K., Kozma, R., and Watts, M. J., 1998, "Phoneme-Based Speech Recognition via Fuzzy Neural Networks Modeling and Learning," International Journal of Informatics and Computer Science, Vol. 110, Issue 1-2, pp. 61-79.

[9] Halavati, R., Shouraki, S. B., and Zadeh, S. H., 2007, "Recognition of Human Speech Phonemes Using a Novel Fuzzy Approach," Applied Soft Computing, Vol. 7, No. 3, pp. 828-839.

[10] Helmi, N., and Helmi, B. H., 2008, "Speech Recognition with Fuzzy Neural Network for Discrete Words," Proceedings of 4th International Conference on Natural Computation, Vol. 7, pp. 265-269, Jinan, China.

[11] Jang, C.-F., Chiou, C.-T., and Lai, C.-L., 2007, "Hierarchical Singelton-Type Recurrent Neural Fuzzy Networks for Noisy Speech Recognition," IEEE Transactions on Neural Networks, Vol. 18, No. 3, pp. 833-848.

[12] Othman, A. M., and Riadh, M. H., 2008, "Speech Recognition Using Scaly Neural Networks," International Journal of Intelligent Systems and Technologies, Vol. 3, No. 2, pp. 71-76.

[13] Kasabov, N., and Iliev, G., 2001, "Hybrid System for Robust Recognition of Noisy Speech Based on Evolving Fuzzy Neural Networks and Adaptive Filtering," IEEE/ENNS/ENNS International Joint Conference on Neural Networks.

[14] Lippmann, R. P., 1989, "Review of Neural Networks for Speech Recognition," Neural Computation, Vol. 1, No. 1, pp. 1-38.