A NEW STRATEGY OF FUZZY-NEURAL NETWORK FOR THAI NUMERAL SPEECH RECOGNITION

C. Wutiwiwatchai, S. Jitapunkul, V. Ahkuputra, E. Maneenoi

Digital Signal Processing Research Laboratory, Department of Electrical Engineering,
Faculty of Engineering, Chulalongkorn University, Bangkok 10330, THAILAND
e-mail : jsomchai@chula.ac.th

S. Luksaneeyanawin

Linguistic Research Unit, Department of Linguistics, Faculty of Arts,
Chulalongkorn University, Bangkok 10330, THAILAND
e-mail : sudaporn@chulakn.car.chula.ac.th

ABSTRACT

In this research, a new strategy of Fuzzy-Neural Network system was proposed for Thai numeral speech recognition. Instead of using the fuzzy membership input with class membership desired-output during training procedure as proposed by several researches, we used the fuzzy membership input with fundamental binary desired-output. This can reduce the misunderstood training, decrease the training time and also improve the recognition ability. The system was tested on the Thai ten-numeral speech (0-9) recognition. The error rate for speaker-independent test achieved 9.2% compared to 14% error rate of conventional neural network system while the error rate of the system using class membership desired-output is quite high because of misunderstood training.

1. INTRODUCTION

It has been proven by many researches that the use of fuzzy techniques can improve the accuracy of speech recognition. A main reason is that the nature of human speech is always ambiguous, and can be overcome by fuzzy system. A major problem of fuzzy system is an increment in amount of computation required. The use of neural network models, which have an ability of parallel computation, can suitably offset this problem. Hence, many techniques based on fuzzy-neural network (fuzzy-NN) system are still implemented for speech recognition.

A novel fuzzy-NN system used the fuzzy membership input with the class membership desired-output for neural network training, e.g., the works in [1], [2]. A similar fuzzy-NN system used only the class membership desired-output while the input was still normal, e.g., the works in [3], [4]. The goal of using class membership values during training is to make a soft-decision system like human decision. However, some problems occurred when using class membership values. First, with useful features extracted from any ambiguous speech, e.g., Linear Predictive Coefficients (LPC), the values of class membership computed were not corresponding to the input pattern. This problem will increase when number of training speech is increased. Second, although computed class membership values are corresponding to the input pattern, an over-soft decision still occurred in the worst case. This will spent too much training time and maybe divergent. Pal.[1] called this problem "the fuzziest case" and proposed an INT-fuzzy modification used to improve the class membership values. This technique can offset the second problem, but not for the first problem.

In this paper, a new approach to overcome both problems described above is presented. Instead of using fuzzy membership input with class membership desired-output for neural network training, the use of fuzzy membership input with general binary desired-output is proposed. The implemented system is tested on Thai numeral speech recognition and compared to the conventional neural network system and the fuzzy-NN system using class membership value.

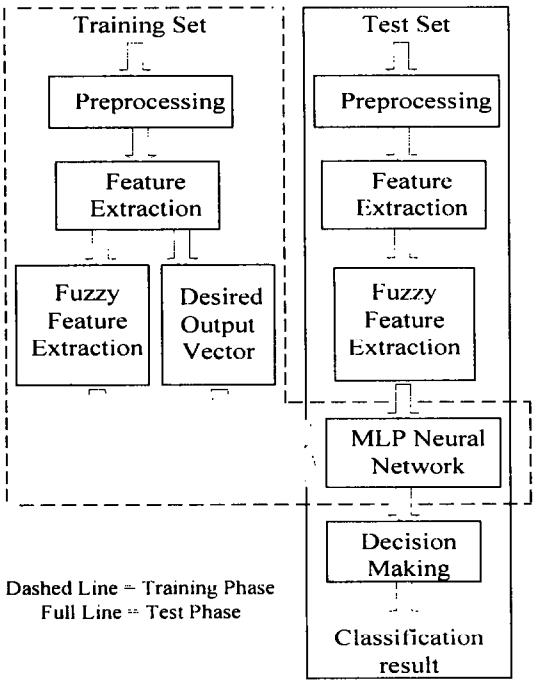

Figure 1. A diagram of fuzzy-NN model

2. A FUZZY-NN MODEL

Figure 1 shows a diagram of fuzzy-NN model, which is consisted of training and test procedure. In training procedure, speech signals are passed through the preclassification, which classify each signal to a sub-network. The raw speech signals are then detected the endpoints and normalized in the preprocessing step. Useful features, such as Linear Prediction Coefficient (LPC), are extracted after preemphsis, frame blocking, and windowing in the feature extraction step. Then, the fuzzification of exact features and desired-output computation are done to form new input and desired-output vectors for the neural network. The trained neural network is used to test the unknown speech signals later.

2.1 Multilayer Perceptron Neural Network

In this research, Multilayer Perceptron (MLP) neural network is used in order to achieve an isolated numeral speech recognition. MLP consists of an input layer with number of nodes equal to number of input features, an output layer with number of nodes normally equal to number of recognized patterns, and number of hidden layers with number of nodes depended on the complexity of patterns. Backpropagation learning algorithm is used for neural network training. Other details of training and test algorithms are presented in [1], [5].

2.2 Fuzzy Membership Representation of Input Pattern in Linguistic Form

In the conventional MLP model, some prominent features are extracted from each input pattern such as LPC from speech sample. These features arranged in vector form are directly used as the MLP's inputs. However, real processes may posses imprecise or incomplete input features, which can be represented using fuzzy membership values. An exact input feature will be converted to fuzzy membership values on N-linguistic properties, e.g., {Low, Medium, High} for N = 3. Therefore an n-dimensional input vector $\underline{i} = [i_1 \quad i_2 \ldots i_n]$ will be represented as a 3n-dimensional vector

$$\underline{\tilde{i}} = [\mu_L(i_1) \quad \mu_M(i_1) \quad \mu_H(i_1) \quad \mu_L(i_2) \quad \mu_M(i_2) \quad \mu_H(i_2) \ldots .] \tag{1}$$

where $\mu_L(i_j), \mu_M(i_j)$ and $\mu_H(i_j)$ are three membership functions converting feature i_j to be fuzzy membership values in three linguistic properties. Figure 2 shows the overlapping structure of three trapezoidal functions for particular input feature i_j where i_{max} and i_{min} denote the upper and lower bounds of the observed range of feature i_j in all l pattern points. Hence we can easily define any of three-membership function by multi-linear equations such as

$$\mu_L(i_j) = \begin{cases} 1, & i_{min} \le i_j < b \\ (c - i_j)/(c - b), & b \le i_j < c \end{cases} \tag{2}$$

the values of a - e can be simply defined to be the edge points of six-equal intervals clustered from the feature space. Instead of using exact feature vector, we use this fuzzy membership vector as the MLP's input vector.

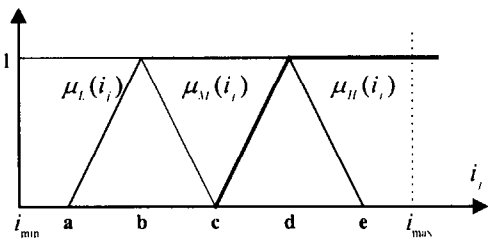

Figure 2. Overlapping Structure of Trapezoidal Function for Linguistic Properties {low, medium, high}

2.3 Class Membership Desired-Output

Another fuzzy technique often incorporated to the fuzzy-NN model is the use of class membership values instead of binary values in desired-output vector during training. The class membership value, lying in range [0, 1], indicates the degree of belongingness of each input pattern to each recognized class. Normally, the class membership value at the output node corresponding to incoming input pattern should be the highest value while the others are small and maybe nonzero values that indicate the degree of belongingness to the other classes. This enables the model to more efficiently classify fuzzy data with overlapping class boundaries.

To calculate the class membership value, we first calculate the weighted distance of training pattern $\underline{i} = [i_1 \quad i_2 \ldots i_j \ldots i_n]$ to the k^{th} class in a c-class problem

$$z_k = \sqrt{\sum_{j=1}^{n} \left(\frac{i_j - o_{kj}}{v_{kj}} \right)^2} \tag{3}$$

where i_j is the j^{th} exact feature, o_{kj} and v_{kj} denote mean and standard deviation of the j^{th} feature for the k^{th} class respectively. The class membership value of this input pattern to the k^{th} class is then computed using value of weighted distance as

$$\mu_k(\underline{i}) = \frac{1}{1 + \left(\frac{z_k}{f_d} \right)^{f_e}}, \quad \mu_k(\underline{i}) \in [0,1] \tag{4}$$

where f_d and f_e is the denominational and the exponential fuzzy generator respectively. These parameters control an amount of fuzziness in this class membership set.

3. DISADVANTAGES OF USING CLASS MEMBERSHIP VALUE

Two disadvantages are founded in the fuzzy-NN using class membership desired-output:

First, the speaking variation problem may cause an overlapping of each class. Especially in the worst case, an occurrence of over-ambiguous input speech pattern, which produces a desired-output vector, consisted of the highest value at the wrong node. This will cause a misunderstood training, which absolutely gives incorrect recognition.

Second, although the output node containing the highest score of class membership value is corresponding to the input pattern, another problem occurs with an ambiguous speech pattern. This speech pattern may correspond to two classes or more in the input feature space. Hence, the computed class membership desired-output vector may consists of two high values or more. This case may conducts an ambiguous training causing too much training time and unexpected classification results. Pal.[1] called "the fuzziest case" and suggested to modify these membership values using INT-fuzzy modification equation as

$$\mu_{INT}(i) = \begin{cases} 2[\mu(i)]^2, & 0 \le \mu(i) \le 0.5 \\ 1 - 2[1-\mu(i)]^2, & \mu(i) \ge 0.5 \\ 0, & otherwise \end{cases} \quad (5)$$

This modified strategy will produce new membership output values that provide less class overlapping. However, this strategy cannot overcome the first problem described above. In order to solve both problems, we propose to use binary desired-output values, which are always "1" at the correct node and "0" at the others, instead of class membership desired-output values. This strategy is suitable for the MLP based system, which have an ability of input-output mapping although there are many input variations. This is similar to human recognition, which can remember any over-ambiguous speech by experience. This proposed strategy can also decrease training time, an amount of computation, and allows more input variation.

4. IMPLEMENTATION AND RESULT

The system is implemented in C on Pentium-Pro personal computer for Thai numerals, zero to nine. The speech samples are recorded twice at 16-bit and 11 kHz sampling rate from 60 speakers; 50 for training and speaker-dependent test, the rest for speaker-independent test. The speech samples are passed through the signal preprocessing where speech samples are preemphasized and smoothing windowed using 20 ms Hamming window with 5 ms frame shift. The 10-order LPCs are applied for speech feature extraction from each speech frame as a feature vector. This vector is used as the input vector, and a corresponding binary vector is used as the desired-output vector of the conventional MLP. In the fuzzy-NN system, the obtained feature vector is converted by three trapezoidal functions on 3-linguistic properties as described above. A new converted vector will become a new input vector of the MLP. The classical fuzzy-NN system uses a class membership vector as a desired-output vector, while the proposed fuzzy-NN or the modified fuzzy-NN as we called uses a desired-output vector of binary values. All systems are tested with three test sets: training test set which is the training set, speaker-dependent test set, and speaker-independent test set. Various sets with different number of training speaker are used to show the decreasing trend of error rate. Several parameters used in neural network are 0.9 momentum rate, 0.1 learning, 0.01 and 0.00001 error threshold depending on fundamental LPC input or fuzzy membership input, f_d of 5 and f_e of 1. Number of hidden node is varied to obtain the best result for each system.

Table 1. Error rate results of the Modified Fuzzy-Neural Network system

No. of Training speaker	Error Rate (%)		
	Training Test Set	Dependent Test Set	Independent Test Set
10	0.0	27.3	36.3
20	0.2	23.5	24.7
30	0.3	17.5	19.3
40	0.5	12.0	13.5
50	0.5	9.0	9.2

Table 2. Error rate results of the Conventional Neural Network system

No. of Training speaker	Error Rate (%)		
	Training Test Set	Dependent Test Set	Independent Test Set
10	0.0	27.7	44.0
20	0.0	23.3	32.0
30	0.0	21.4	27.0
40	0.0	16.2	19.7
50	0.0	13.0	14.0

Table 3. Error rate results of the Classical Fuzzy-Neural Network system

No. of Training speaker	Error Rate (%)		
	Training Test Set	Dependent Test Set	Independent Test Set
10	34.0	56.0	64.0
20	41.0	59.5	66.0
30	48.5	62.0	69.5
40	52.5	64.5	71.0
50	59.0	67.5	71.5

Table 1, 2, and 3 show the error result results of the modified fuzzy-NN system, the conventional NN system, and the classical fuzzy-NN system respectively. Comparison of error rate result is shown prominently in figure 3. It can be seen that the modified fuzzy-NN system can achieve the best result of 9.2% error rate for speaker-independent test. There is 4.8% decrease from the best result of fundamental

NN system without fuzzy. The classical fuzzy-NN system using class membership conducts quite high error rate although it is tested with the training test set. Furthermore, an abnormal trend of error rate occurs when number of training speaker is increased. The reason is that there are some misunderstood desired-output vectors occurred when using class membership as described in section 3. This case will happen frequently when number of training speaker is increased. In contrast to the modified fuzzy-NN system using binary desired-output vector, the desired-output vector can be defined correctly although the input pattern is very ambiguous especially in Thai numeral speech set, which contains several ambiguous vocabularies such as the words /su:n4/, /s@:ng/, and /sa:m4/ (mean "0", "2", and "3" respectively). Moreover, this can decrease number of computation required and the training time. However, we have considered that there are two possible solutions to solve the problem of misunderstood desired-output vector. First, the feature that we used, LPC, is maybe unsuitable for fuzzy system because it only maps the curve of speech envelope, which can be ambiguous feature. The way to use a more efficient feature such as distinctive feature [7] will possibly result better. This idea is now in implementation. The second solution is which we use in this experiment, to use binary desired-output vector instead of class membership desired-output vector.

5. CONCLUSION

A strategy of using fuzzy theory can improve the result of Thai numeral speech recognition, but it needs a modification of using binary desired-output vector instead of class membership desired-output vector. A new fuzzy-NN system, which uses binary desired-output and the fuzzy membership input vector, can overcome the problem of misunderstood training vector occurred when using the class membership value and can also enhance the error rate result compared to the fundamental NN system without fuzzy. The best result of Thai numeral speech recognition for speaker-independent test is 9.2% error rate, which is 4.8% decreasing from the non-fuzzy NN system. This modified system can also decrease number of computation required and the training time compared to the classical fuzzy-NN system.

6. ACKNOWLEDGEMENT

The authors would like to acknowledge the Digital Signal Processing Research Laboratory on the support of these researches.

7. REFERENCES

[1] Pal S. K. and Mitra S., "Multilayer Perceptron, Fuzzy Sets and Classification". *IEEE Transactions on Neural Networks*, Vol 3, September 1992, pages 683-697.

[2] Carlos A. R., Carlos A. G. and Wyllis, B., "The Use of Trapezoidal Function in Linguistic Fuzzy Relational Neural Network for Speech Recognition". *IEEE International Conference on Neural Networks*, Vol 7, July 1994, pages 4487-4492.

[3] Komori Y., "A Neural Fuzzy Training Approach for Continuous Speech Recognition Improvement". *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol 1, March 1992, pages 405-408.

[4] Gurgen F. S., Aikawa, K. and Shikano, K., "Phoneme Recognition with Neural Network using a Novel Fuzzy Training Algorithm". *IEEE International Joint Conference on Neural Networks*, Vol 1, November 1991, pages 572-577.

[5] Pornsukchandra W. and Jitapunkul S. "Speaker-Independent Thai Numeral Speech Recognition using LPC and the Back Propagation Neural Network". *Electrical Engineering Conference*, Khonkaen, Thailand, November 1996, pages 977-981.

[6] Ahkuputra V., Jitapunkul S., Pornsukchandra W. and Luksaneeyanawin S. "A Speaker-Independent Thai Polysyllabic Word Recognition Using Hidden Markov Model". *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, Canada, Aug. 1997, pages 593-599.

[7] Tubtong N. "Thai Word Recognition Using Phoneme Distinctive Feature". *Master's Thesis*, Chulalongkorn University, 1996.

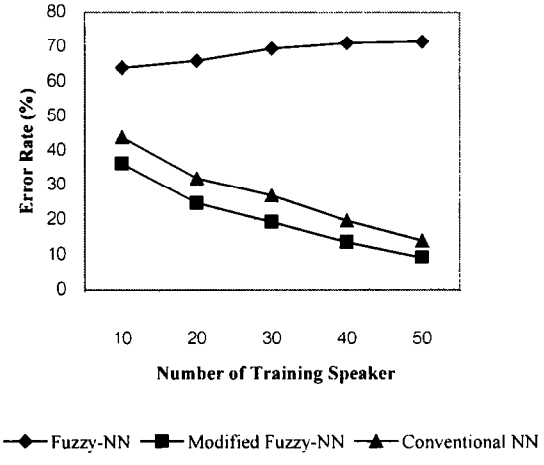

Figure 3. Error rate comparison for speaker-independent test