THAI POLYSYLLABIC WORD RECOGNITION USING FUZZY-NEURAL NETWORK

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, the Fuzzy-Neural Network (fuzzy-NN) model was proposed for Speaker-Independent Thai polysyllabic word recognition. Various fuzzy membership functions on linguistic properties were used to convert exact features extracted from input speech to the fuzzy membership values. The fuzzy membership values were arranged to be new input vector of Multilayer Perceptron (MLP) neural network. The binary desired outputs were used during training. 70 Thai words consist of ten numerals, the others were single-syllable, double-syllable and triple-syllable, 20 words in each group, were used for system evaluation. In order to improve recognition accuracy, number of syllable and tonal level detected were conducted for speech preclassification. The Pi fuzzy membership function provided the best recognition accuracy among other functions; Trapezoidal, and Triangular function. Under an optimal condition, the achieved recognition error rates were 5.6% on dependent test and 6.7% on independent test, which were respectively 3.3% and 3.4% decreasing from the conventional Neural Network system.

1. INTRODUCTION

Unlike English language, Thai language structure consists of 44 consonants, 26 vowels, and 5 tonal levels. These can be combined to a lot of words and also provide many ambiguous words, e.g., /s@:ng4/ and /sa:m4/ (mean "two" and "three" respectively) or /cet1/ and /pet1/ (mean "seven" and "duck" respectively). Recognition of these words is always incorrect although human does it.

Fuzzy logic techniques have been developed to overcome this problem. Hence, it is reasonable to use fuzzy techniques in Thai speech recognition. However, a prominent weak point of fuzzy system is an increase of number of computation required. This can be offset by the parallel computational ability of Neural Network (NN). Therefore, several speech recognition researches were based on fuzzy-NN. A novel fuzzy-NN system, presented in [1] and [2], uses the fuzzy membership functions on linguistic properties and the class membership function to improve the neural network's input and desired-output data respectively. Another system proposed in [3] and [4] uses only the fuzzification on desired-output data. In this research, the neural network's input data are converted to the fuzzy membership values while the desired-output data are still normal binary values. This can overcome the problem of misunderstood training occurring with the system using class membership desired-output.

The Multilayer Perceptron (MLP) neural network is used to achieve isolated-word recognition. Generally, a MLP can recognize better, if number of recognized vocabularies is decreased. Hence, several preclassification approaches are used. In this research, number of syllables per word is used to first classify. With a prominent feature of tones in Thai syllable, the tone detection algorithm based on fundamental frequency approach is used to classify later.

Two procedures of experiment are set up. First, an experiment of numeral speech recognition is conducted and compared among using various types of fuzzy membership functions. The last is an experiment of polysyllabic word recognition with two preclassification methods described above.

2. A FUZZY-NN MODEL

A Proposed fuzzy-NN model 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 $\underline{i} = \left[i_1 \quad i_2 \, ... \, i_n\right]$ will be represented as a 3n-dimensional vector

$$\begin{aligned} \underline{\tilde{i}} = [\mu_L(i_1) \quad \mu_M(i_1) \quad \mu_H(i_1) \\ \mu_L(i_2) \quad \mu_M(i_2) \quad \mu_H(i_2) ....] \end{aligned} \qquad (1)$$

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.

2.3 Types of Fuzzy Membership Function

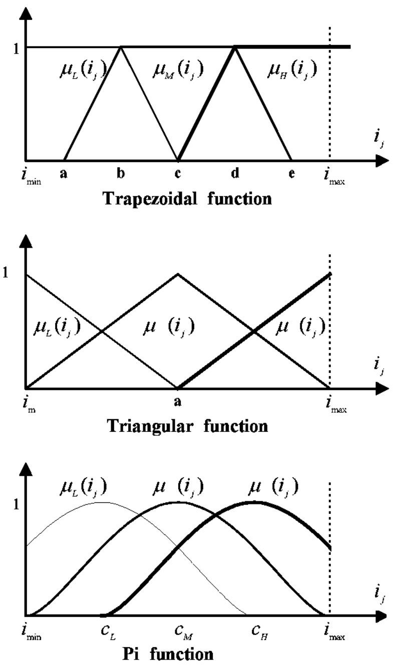

Figure 1. Three types of fuzzy membership function on 3-linguistic properties

There are several kinds of fuzzy membership function that have ever proposed such as the Pi function in [1], the Trapezoidal function in [2]. Pal. [1] suggested to use the Pi function instead of Trapezoidal function by a reason that it is a more cost-effective representation with fewer inputs dedicated to a particular input feature. It can be noted that there is no convenient suggestion in using what type of membership function. Hence, three membership functions, which are Pi, Trapezoidal, and Triangular function, are used and compared in this research. Figure 1 shows overlapping structures of three membership functions on 3-linguistic properties {Low, Medium, High}.where $i_{\min}$ and $i_{\max}$ denote the minimum and maximum value of exact feature or LPC space. The values of a – e in figure 2 are simply defined to cluster the feature range into equal sections. Then each Trapezoidal or Triangular function can be defined easily in multi-linear equation.such as

$$\mu_L(i_j) = \begin{cases} 1 \quad , i_{\min} \le i_j \le a \\ (b - i_j)/(b - a), \, a \le i_j \le b \\ 0 \quad , \mathit{otherwise} \end{cases} \qquad (2)$$

which is the trapezoidal function of Low-linguistic property. Pi function can be defined as

$$\mu(i_j) = \begin{cases} 0 \quad , \mathit{otherwise} \\ 2(1 - \frac{\left|i_j - c\right|}{\lambda})^2, \frac{\lambda}{2} \le \left|i_j - c\right| \le \lambda \\ 1 - 2(\frac{\left|i_j - c\right|}{\lambda})^2, 0 \le \left|i_j - c\right| \le \lambda \end{cases} \qquad (3)$$

where c is the center of function shape and λ is a parameter controlling the width of shape. C and λ of the function on each linguistic property can be calculated as following

$$\begin{aligned} \lambda_M &= \frac{1}{2}(i_{\max} - i_{\min}), \, c_M = i_{\min} + \lambda_M \\ \lambda_L &= \frac{1}{fdenom}(c_M - i_{\min}), \, c_L = c_M - \frac{1}{2}\lambda_L \\ \lambda_H &= \frac{1}{fdenom}(i_{\max} - c_M), \, c_H = c_M + \frac{1}{2}\lambda_H \end{aligned} \qquad (4)$$

where $fdenom$ denotes the denominator adjusting the degree of overlapping of each function. This parameter directly effects to the fuzziness of membership values.

3. PRECLASSIFICATION

Normally, a neural network can recognize better if number of recognized vocabularies is reduced. Therefore preclassification techniques are often conducted. Two preclassification techniques are used in this research.

3.1 Preclassification by number of syllable

It is reasonable to preclassify the incoming words by number of syllable of each word, because of two reasons. First, vocabularies used in this research are polysyllabic words, which consist of single, double, and triple-syllabic words. Time duration of each word should be increased when number of syllable per word is increased. Therefore, if the words are classified into three groups being different in number of syllables, we can apply a time-normalization process to each vocabulary group with number of normalized samples suitably. Second, counting of number of syllable per word is obtained by product of the syllable detection, which is a useful technique and still developed continuously as seen in many works. This technique is also implemented for Thai language using several approaches such as energy and frequency-based [5], [8].

3.2 Preclassification by tone

After a word is passed through the syllable detection, all syllables can be separated. In order to classify finely, the tone detection is conducted to extract tone of the first syllable. Thai language is a tonal language like Mandarin. Two groups of Thai tone are static tone, which is consisted of high tone /3/, middle tone /0/, and low tone /1/, and dynamic tone, which is consisted of falling tone /2/ and rising tone /4/. All five tones can be distinguished easily by using the direction and the level of fundamental frequency (F0). Several algorithms of Thai tone detection have been implemented [7].

4. IMPLEMENTATION AND RESULT

The system is implemented in C on Pentium-Pro personal computer. 70-Thai words consisted of 10 numerals zero to nine, 20 single-syllabic words, 20 double-syllabic words, and 20 triple-syllabic words are used in experiment. 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 emphasized and smoothing windowed using 20 ms Hamming window with 5 ms frame shift. The 10-order LPC is applied for speech feature extraction for each speech frame and used to form input vectors for fundamental NN system. For fuzzy-NN system, new vectors are generated which are consisted of fuzzy membership values of LPCs on 3-level linguistic properties {Low, Medium, High}. Both systems use the same binary desired-output vector during training. There are two experimental procedures in this research as following.

4.1 Comparative experiment of three fuzzy membership functions

Three useful fuzzy membership functions; pi, trapezoidal, and triangular function are used for fuzzification. This experiment is implemented for ten Thai numeral speech recognition and also compared to the conventional NN using LPC input. Several parameters used in this experiment are momentum rate of 0.9, learning rate of 0.1, error threshold of 0.01 and 0.00001 for LPC input and fuzzy membership input respectively, and $fdenom$ of 0.8. Number of hidden nodes is selected optimally to obtain the best result as possible. The error rate results with respect to number of training speakers for speaker-independent test are shown in figure 2.

The results show that the use of pi membership function is the best and the use of LPC is the worst. This can be analyzed that the objective of using fuzzy membership input instead of exact feature input is to cover any imprecise or incomplete input feature, which always occur in speech processing. Hence, a variation of feature value should be indicated by a change of its fuzzy membership value. This objective can be achieved when using pi and triangular function, except trapezoidal function. However, a large change of triangular membership value is not appropriate for speech feature compared to the other functions. Although the pi membership input can achieve the best solution, the training time required when using the trapezoidal function is the less. The reason is that the use of pi or triangular function will generate an input vector, which is consisted of every non-zero element while there are many trapezoidal membership values that equal to zero. Therefore, the computation required for pi or triangular membership values are more complicated than trapezoidal membership values.

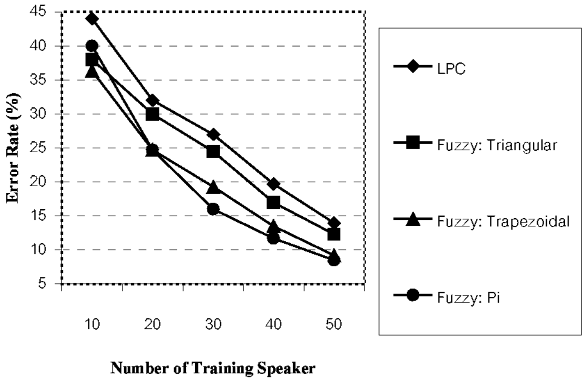

Figure 2. Comparison graph of error rate results using various types of input

4.2 Thai polysyllabic word recognition

The fuzzy-NN system using pi membership input with binary desired-output is selected for polysyllabic word recognition. The system is trained and tested with speaker-dependent and speaker-independent test set (Test1 and Test2 respectively). With preclassification by number of syllables, vocabularies can be separated into three groups; single, double, and triple syllabic word. It can be notes here that Thai numeral words, zero to nine, are all single-syllabic words, then they are included in the single-syllable group. The other parameters used are the same as section 4.2. The error rate results compared to the conventional NN system using LPC input are shown in table 1. The experiment on single-syllable group using fuzzy-NN system is not available in this research, because a large number of vocabularies contained in this group will conduct too much training time. Average values of error rate shown in the table are computed regarding number of vocabularies in each group as

$$Average\ error\ rate = \frac{\sum_{all\ group}(error\ rate \times No.\ of\ vocab)}{Total\ No.\ of\ vocab} \qquad (5)$$

According to results, the error rates are generally higher when number of vocabularies is increased. Average error rates of speaker-dependent and speaker-independent test using pi membership input are 6.2% and 9.7% decreasing from the test using LPC input. The reason of improvement when using fuzzy system is that the use of fuzzy technique can allow the recognition system to overcome many incomplete feature values. Hence, some ambiguous words such as /s@:ng4/ and /sa:m4/ (mean “2” and “3” respectively), or /cet1/ and /pet1/ (mean “7” and “duck” respectively), which usually provide many incomplete feature values and be always recognized incorrectly, can be fixed better with this fuzzy system.

Table 1. Error rate result of vocabulary group preclassification by number of syllable

Error Rate (%)					
Vocab-Group		**1-syll**	**2-syll**	**3-syll**	**Average**
No. of Vocab		**30**	**20**	**20**	**70**
Fuzzy:	**Test1**	N.A	9.9	10.3	10.1
	Test2		12.0	12.5	12.2
LPC	**Test1**	21.2	12.8	12.4	16.3
	Test2	27.5	15.7	19.8	21.9

In order to decrease error rate results, tone of the first syllable in each word is extracted for more preclassification. With this procedure, vocabularies can be classified into fifteen groups. Again, we have no experiment on some groups that contain only one vocabulary or none. Table 2 shows the results.

Table 2. Error rate of vocabulary group preclassification by number of syllable and tonal level

Error Rate (%)					
Vocab Group	**No. of Vocab**	**Fuzzy: Pi**		**LPC**	
		Test 1	**Test 2**	**Test 1**	**Test 2**
1-syll tone0	**7**	7.8	7.5	12.9	11.9
1-syll tone1	**8**	7.0	8.3	12.5	15.7
1-syll tone2	**7**	7.6	10.2	11.2	13.7
1-syll tone3	**3**	3.9	5.8	4.7	7.8
1-syll tone4	**5**	6.9	4.5	12.9	6.0
2-syll tone0	**10**	5.7	7.0	8.5	10.3
2-syll tone1	**3**	3.5	2.8	4.4	3.3
2-syll tone2	**2**	0.0	0.0	2.0	0.0
2-syll tone4	**5**	5.0	3.7	7.1	5.3
3-syll tone0	**8**	6.6	11.9	9.5	8.7
3-syll tone1	**5**	1.6	0.0	4.4	0.0
3-syll tone3	**5**	5.4	10.0	7.2	2.7
Average	**70**	5.6	6.7	8.9	10.1

For fuzzy-NN system, the average error rate results can achieve the best 5.6% and 6.7% for speaker-dependent and speaker-independent test. There are 3.3% and 3.4% decreasing from the fundamental LPC system. As seen that the percentages of decrement of error rate are less than the results from table 1. This can be explained that with small size of vocabularies, the results of LPC system are quite good themselves, the use of fuzzy technique can certainly improve not much.

5. CONCLUSION

A fuzzy-NN system, which uses the fuzzification of exact feature to form new input vector for neural network, can improve the error rate of Thai speech recognition compared to the fundamental NN system using exact feature input such as LPC. In this research, pi membership function can achieve the best function above other functions; trapezoidal and triangular function. However, several preclassification techniques are required in order to reduce the size of vocabularies to be recognized by each sub-network. Number of syllables for peach polysyllabic words and especially tonal level of Thai syllable are two outstanding features used for preclassification techniques.

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.

[8] Pratumtan T. “Thai Speech Recognition Using Syllabic Based”. *Master’s Thesis*, Chulalongkorn University, 1987.