Use of a Self-Learning Neuro-Fuzzy System for Syllabic Labeling of Continuous Speech

Ching-Tang Hsieh, Mu-Chun Su and Shih-Chieh Chienn

Department of Electrical Engineering, Tamkang University, Taiwan, R.O.C.

Abstract- For reducing requirement of large memory and minimizing computation complexity in large-vocabulary continuous speech recognition system, speech segmentation plays an important role in speech recognition systems. In this paper, we formulate the speech segmentation as a two-phase problem. Phase 1 (frame labelling) involves labeling frames of speech data. Frames are classified into three types : (1) silence; (2) consonants and (3) vowels according to two segmentation features. In phase 2 (syllabic unit segmentation) we apply the concept of transition states to segment continuous speech data into syllabic units based on the labeled frames. The novel class of hyperrectangular composite neural networks (HRCNN's) is used to cluster frames. The HRCNN's integrate the rule-based approach and neural network paradigms, therefore, this special hybrid system may neutralize the disadvantages of each alternative. The parameters in the trained HRCNN's are utilized to extract both crispy and fuzzy classification rules. Four speaker's continuous reading-rate Mandarin speech are given to illustrate the proposed two-phase speech segmentation model. In our experiments, the performance of the HRCNN's is better than the "Distributed Fuzzy Rule" approach based on the comparisons of the number of rules and the correct recognition rate.

I. Introduction

In general, there are two kinds of speech segmentation. One is phonemic unit segmentation [1] and the other is syllabic unit segmentation [2]. Either phonemic or syllabic unit segmentation, most of the works are based on the thresholds of parameters to segment the speech data. The thresholds of segmentation features (e.g. the zero crossing rate and the average value of the first formant band for each frame) are usually set by personal experience. Therefore, the segmentation performance is not very reliable. This motivated us to formulate the decision of thresholds as a classification problem in order to automate the threshold setting procedure. The classification problem may be solved either by crispy rules or by fuzzy rules.

Several approaches have been proposed for classification problems. Some works focused on conventional probabilistic and deterministic classifier [3],[4],[5]. These approaches attempt to determine the exact decision region of a class from its prototypes. In addition, it is usually assumed that a pattern can belong to one and only one class. One approach uses neural networks to classify patterns. For backpropagation networks, classification knowledge is encoded in the parameters of trained networks, but it is hard to analyze the network and the learning process is relatively slow. Recently, several approaches focus on generating fuzzy if-then rules directly from numerical data. In most of fuzzy systems, construction of fuzzy rules from numerical data for classification problems consists of two phases: (1) fuzzy partition of a pattern space and (2) identification of a fuzzy rule for each fuzzy subspace. The major restriction of this approach is that the number of divisions of each input variable must be preselected. In addition, the degree of partitions will affect the classification power and the number of generated fuzzy rules. One approach to remedy the mentioned disadvantages is to use the concept of distributed representation of fuzzy rules which is implemented by supercomposing many fuzzy rules corresponding to different fuzzy partitions of a pattern space [6]. However, this approach will still result in a lot of unnecessary fuzzy rules. The genetic algorithm has been proposed for choosing an appropriate set of fuzzy rules [7]. In [8], [9] fuzzy rules with variable fuzzy regions are extracted for classification problems. These approaches do not need to define the number of divisions of each input variable in advance. In [8] each class is represented by a set of hyperboxes, in which overlaps among hyperboxes for the same class are allowed, but no overlaps are allowed between different classes. However this approach may not easily handle patterns where complicated separative boundaries exist. To overcome this problem, two types of hyperboxes (1) activation hyperboxes and (2) inhibition hyperboxes were proposed in [9]. These hyperboxes are defined recursively. It allows inhibition hyperboxes to be stored inside activation hyperboxes, and inhibition to inhibition can be nested any number of levels deep. However, during the training procedure, if the activation hyperboxes at level n is identical to the inhibition hyperbox at level (n-1), they need to define a set of activation hyperboxes which include only one datum. Therefore it will require a large memory for storing these parameters defining these hyperboxes and length computation time for classifying patterns. In [10], [11], [12] a novel class of hyperrectangular composite neural networks (HRCNN's) has been proposed to extract both crispy and fuzzy rules from numerical data. The parameters of the trained HRCNN's are easily utilized to represent a set of if-then rules. This kind of self-learning neuro-fuzzy system has advantages over backpropagation networks and conventional fuzzy systems in the following way: (1) the learning process is fast; (2) it is easy to analyze the trained HRCNN's; (3) it is easy to apply HRCNN's to problems in which the number of input variable is large; and (4) we do not need to define the number of division of each input variable in advance.

After we have finished the frame labeling phase, the next phase is to correctly segment the speech data into syllabic units. If we assume that every frame be classified correctly into one of the three types: silence (0), consonants (1), and vowels (2), then the frames of speech data can be represented by the labeling sequence as

$$[00...0][11...1]22...2000...0. \qquad (1)$$

Therefore it is easy to segment the sequence into syllabic units by detecting any transition in the labeling sequence. However, it is usually not possible to achieve 100% recognition rate for any classification algorithm even if we use the HRCNN's. In order to offset the non-perfect recognition performance of HRCNN's, the concept of transition states is applied in the second phase. From a set of labeling sequences, several transition states can be concluded. Then the segmentation errors carried in phase I can be remedy in the phase II.

Effectiveness of this approach has been substantiated by segmentation experiments for samples of continuous, radio news uttered in Mandarin by two female and two male. This paper is organized as follows. In section 2 we discussed the characteristics of Mandarin. Section 3 briefly describes the novel class of hyperrectangular composite neural networks and the distributed representation of fuzzy rules. The experimental results are given in Section 4. Finally, some concluding remarks are presented in Section 5.

II. Characteristics of Mandarin

The characteristics of mandarin speech is very different from other languages, such as English and French. Basically, a Mandarin word can be expressed as

$$[\text{Consonant,}][\text{median,}]\ \text{Vowel}[\text{,tail}] \qquad (2)$$

where [.] is optional. The Median may be pronounced as /i/, /u/, or /iu/ and the Tail may be produced as /n/ or /ng/. In most of Mandarin speech recognition systems, it is usually to regard the combination of the Median, vowel, and Tail as a Vowel. Thus, a Mandarin word is usually represented as

$$[\text{Consonant,}]\ \text{Vowel.} \qquad (3)$$

Since every Mandarin word may be expressed by Eg. (3), Mandarin speech can be partitioned into three different syllabic units: (1) silence; (2) consonant, and (3) vowel. Table I illustrates the relationship among phonemes and syllabic units.

The speech signal is a slowly time varying signal in the sense that when examined over a sufficiently short period of time, (between 5 and 100 msec) its characteristics are fairly stationary; however, over long periods of time (on the order of 1/5 seconds or more) the signal characteristics change to reflect the different speech sounds being spoken [13]. There are several different ways of characterizing the speech signal and representing the information associated with the sounds. In general, the speech signal is first sampled through an A/D converter. After sampling the signal, a spectral analysis based in fast Fourier transformation (FFT) and further calculations for extracting features are performed. The choice of features is made mostly based on ad hoc consideration. In our experiments, two features extracted from every frame, the average value of first formant band (VI) and the zero crossing rate, are used for frame classification involved in the first phase. The first parameter is based on the

Table I. Relationship among phonemes and syllabic units

Label	Class	Phoneme
0	silence	silence
1	consonants	f,d,t,g,k,h,j,b,p,y,m,N,ch,sh,tz,ts,s,r,l
2	vowels	i,u,a,o,e,io,ai,ei,ai,ei,au,ou,el,ia,ie,iau,iou,ua,uo,uai,uei,ue,iue,iua,iu,n,ng

spectral envelopes of speech obtained by the unbiased log spectral estimation. This parameter has been used in voiced/unvoiced detection, and with good performance in [1]. It represents the degree of periodic characteristic of vowels. The second parameter, usually to be used to decide the startpoints and endpoints of syllables [2], represents the degree of frictions. These two parameters are used as the input variables to the HRCNN's.

III. Rule Extraction

A. Hyperrectangular Composite Neural Networks

The construction of an rule-based expert system involves the process of acquiring production rules. Production rules are often represented as "IF condition THEN act". The backpropagation networks may not arrive at an inference structure close to this kind of high level knowledge representation. Here a novel class of hyperrectangular composite neural networks is presented. The class of hyperrectangular composite neural networks provides a new tool for machine learning. The classification knowledge is easily extracted from the weights in a hyperrectangular composite neural network. The symbolic representation of a neural node with hyperrectangular neural-type junctions is shown in Fig. 1(a) and it is described by the following equations :

$$net_j(\underline{x}) = \sum_{i=1}^{n} f((M_{ji} - x_i)(x_i - m_{ji})) - n \qquad (4)$$

and

$$out_j(\underline{x}) = f(net_j(\underline{x})) \qquad (5)$$

where

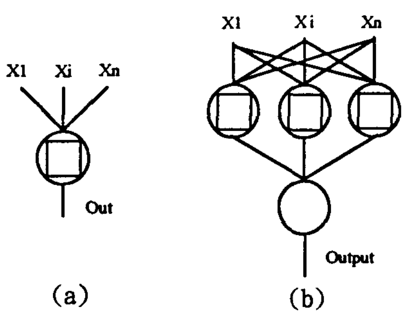


Fig. 1 (a) a neural node with "hyperrectangular neural-type" junctions and (b) a hyperrectangular composite neural network

$$f(x) = \begin{cases} 1 & if\, x \geq 0 \\ 0 & if\, x < 0 \end{cases} \qquad (6)$$

M_{ji} and m_{ji} R are adjustable weights of the jth neural node, n is the dimensionality of input variables, and is an output function of a neural node with hyperrectangular neural-type junctions.

The supervised decision-directed learning (SDDL) algorithm generates a two-layer feedforward network in a sequential manner by adding hidden nodes as needed. As long as there are no identical data over different classes, We can obtain 100% recognition rate for training data. Fig. 2 illustrates the training procedure. The flow chart of the SDDL algorithm is given in Fig. 3. A more detailed description of the training procedure is given in [10].

Extraction of Crispy Rules from HRCNN's

According to Eqs. (4) - (6), we know that a neural node with hyperrectangular neural-type junctions outputs one only under the following condition:

$$\begin{cases} m_{j1} \leq x_1 \leq M_{j1} \\ m_{j2} \leq x_1 \leq M_{j2} \\ \quad\cdots \\ m_{jn} \leq x_1 \leq M_{jn} \end{cases} \qquad (7)$$

Therefore, the classification knowledge can be described in the form of a production rule. The IF...THEN...rule has an antecedent (premise) consisting of the one condition as well as a single consequent.

$$IF((m_{j1} \leq x_1 \leq M_{j1}) \cap \ldots \cap (m_{j1} \leq x_1 \leq M_{j1})) \qquad (8)$$

$$THEN \text{ OUT} = 1.$$

The domain defined by the antecedent of Eq. (8) is a hyperrectangle in n-dimensional space. Owing to characteristics of data, such as **dispersion characteristic**, data may present an existence of many distinct clusters in input space. It is efficient to cope with, this situation by using a set of hyperrectangles with different sizes and locations to fit the data. Each hyperrectangle corresponds to an "intermediate rule". The final classification rule is

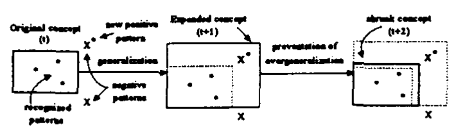


Fig. 2 The training procedure

the aggregation of the intermediate rules. This kind of expression is then mapped to a tow-layer hyperrectangular composite neural network. In the composite neural network configuration shown in Fig. 1(b), input variables are assigned to input nodes, intermediate rules correspond to hidden nodes, and the final classification rule is assigned to an output node. Each hidden node is connected, with weight valued 1.0, to an output node. The output node implements the "OR" function. Therefore, if there are K hidden nodes in the composite neural network, the classification rule is expressed in the following form:

$$
\begin{aligned}
&IF((m_{11} \le x_1 \le M_{11}) \cap \ldots \cap (m_{1n} \le x_1 \le M_{1n}))\\
&\quad THEN \text{ OUTPUT} = 1;\\
&\quad \ldots \qquad\qquad\qquad\qquad\qquad\qquad (6)\\
&ELSE\ IF((m_{k1} \le x_1 \le M_{k1}) \cap \ldots \cap (m_{kn} \le x_n \le M_{kn}))\\
&\quad THEN \text{ OUTPUT} = 1;\\
&ELSE \text{ OUTPUT} = 0;
\end{aligned}
$$

Extraction of Fuzzy rules

After having found a set of crispy if-then rules, we may fuzzify these crispy rules by using a reasonable fuzzy membership function. The membership f

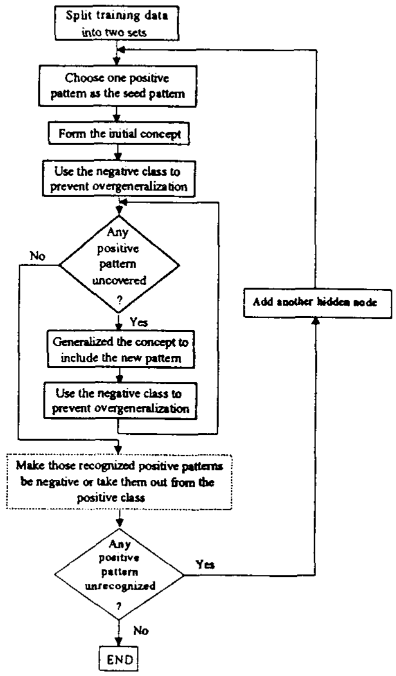


Fig. 3 The flow chart of the SDDL algorithm

unction $m_j(\underline{x})$ for the jth hyperrectangle, 0 $m_j(\underline{x})$ 1, must measure the degree to which the input pattern $\underline{x}$ falls outside of the hyperrectangle. As $m_j(\underline{x})$ approaches 1, the patterns should be more contained by the hyperrectangle, with the value 1 representing complete hyperrectangle containment. In addition, the measurement should reflect the degree of importance of the input feature on a dimension by dimension basis. The membership function that meets all these criteria is defined as

$$m_j(\underline{x}) = \frac{V_j}{V_j + s(Vol(\underline{x}) - V_j)} \tag{9}$$

where

$$V_j = \prod_{i=1}^{n}(M_{ji} - m_{ji}) \tag{10}$$

=volume of the jth hyperrectangle,

$$\text{Vol}(\underline{x}) = \prod_{i=1}^{n} \max(M_{ji}\text{-}m_{ji},\ x_{ji}\text{-}m_{ji},\ M_{ji}\text{-}x_{ji}), \tag{11}$$

and

s_j = the sensitivity parameter that regulates how fast the membership values decreases as the distance between $\underline{x}$ and the jth hyperrectangle. An example of this membership function for two dimensional case is shown in Fig. 4. The connection between the jth hidden node and the output node represents the fraction of positive examples that fall within the jth hyperrectangle The **"maximum-membership defuzzification scheme "** is utilized at the last stage of fuzzy systems [14]. This scheme was motivated by the popular probabilistic methods of maximum-likelihood and maximum-aposteriori parameter estimation. The other approach to find w_j is to use the LMS or the backpropagation algorithm.

B. Distributed representation of fuzzy rules

In [6], the concept of distributed representation of fuzzy rules was introduced. This

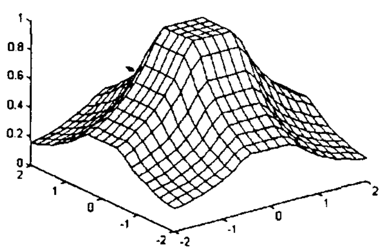


Fig. 4 An example of $m_j(\underline{x})$ when s_j is small for two-dimensional case

Fuzzy partitions	No. of rules	class 0 (%)	class 1 (%)	class 2 (%)	average (%)
2	4	71.53	48.06	99.9	73.16
3	13	87.96	58.71	99.8	82.16
4	29	88.32	60.97	99.6	82.96
5	54	90.51	61.61	99.6	83.91
6	90	91.15	64.84	99.6	84.86
7	139	90.15	66.45	99.6	85.4
8	203	88.69	68.06	99.5	85.42
9	284	89.05	69.53	99.6	86
10	384	89.42	70.79	99.5	86.63
11	505	89.05	72.26	99.5	86.94
12	649	89.05	72.9	99.5	87.15
13	818	88.69	73.55	99.4	87.21
14	1,014	88.69	73.87	99.4	87.32
15	1,239	89.05	74.52	99.4	87.66

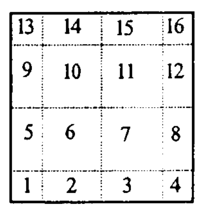
Table II Generalization performance achieved by the "Distributed Representation Fuzzy Rules" approach

approach was applied to extract fuzzy rules for classification pattern. In this approach, the fuzzy rules corresponding to various fuzzy partitions are simultaneously utilized in fuzzy inference. Assume there be L partitions on each input feature. All the fuzzy rules corresponding to the fuzzy partitions from 2 to L are simultaneously used in fuzzy inference. Therefore the total number of the distributed fuzzy rules is $2^2+3^2+...+L^2$. Fig. 5 illustrates the difference between the ordinary approach and the distributed approach.

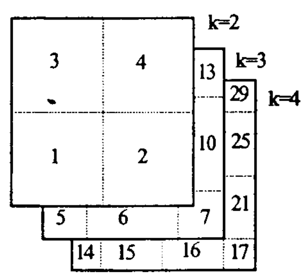
IV. Experimental Results

The overall experimental data are the news uttered in Mandarin by two females and two males. The speech signal is sampled at a rate of 10k Hz. The frame length is 256ms. There are total 22590 frames in our database. First we apply the distributed fuzzy rules method to classify patterns. Table II. tabulates the number of fuzzy rules and its corresponding performance. Then the HRCNN's were trained to extract both crispy and fuzzy classification rules. Table III shows the results. From the comparisons of Table II and Table III, we found that the HRCNN's are batter than the "distributed fuzzy rules" approach based on the comparisons of (1) the number of fuzzy rules and (2) the correct recognition ratio.

After frames have been classified, the speech signal can be represented by a labeling sequence, e.g.



(a) Ordinary Fuzzy Rules



(b)Distributed Fuzzy Rules

Fig. 5 Representations of different fuzzy rules (L=4)

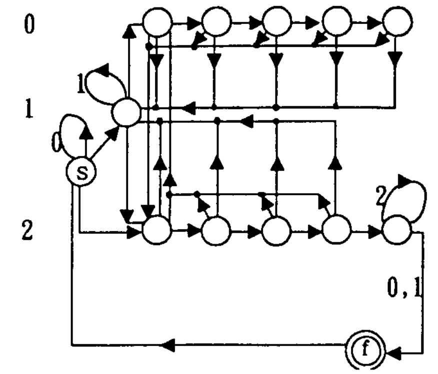


Fig. 6 Transition States

[02...10],[1012...1].2212022 (12)

As mentioned before, the achieved performance of classification is usually not perfect (i.e. 100% correct). If we segment labeling sequence into syllabic units according to the simple strategy - a transition in a labeling sequence represents an existence of a new syllabic unit, a lot of unnecessary syllabic units will be added. Therefore we need some more complicated control strategies to limit the range of segmentation points and label possibilities. Here we propose a method to offset the nonperfect performance of classification. From observations of the labeling sequence, we found the following regularities.

(1)The continuation of label 0 will not excess 5 frames for consonants. This is the difference between consonants and silence.

(2)The consonants usually start with label 1. Sometimes, the syllable will begin with label 2, such as /r/, /b/, /l/, /g/, /d/, /m/, and /N/. These properties can indicate the start-point of syllable.

(3)At least 5 frames of label 2 of sub-sequence should be existed in unit syllable. Table IV illustrates

Table III Generalization performance achieved by the trained HRCNN's : training set (8.85%), testing set (91.14%), and total number of extracted rules is 132

class / approach	0	1	2	average
Crispy (%)	93.8	79.7	93.3	88.9
Fuzzy (%)	94.2	84.7	94.7	91.2

Table IV Some forms of syllables

the forms of syllables	
consonants	vowels
110 112 11...1 11...10 11...101000 11...100000 1112001 12 122 22211...1 20210	2222...2

some possible labeling sequence. We conclude the transition states, which is based on the NFA (Nondeterministic Finite Automata), as Fig. 6.

However, the segmentation work is not finished yet even if we have finished the two phases . Owing to different speaking rate and characteristics of Mandarin words, a long sequence of 2's will exist. This results in the "connecting-syllable problem". That is, only one syllabic unit will be detected even if several vowels are pronounced sequentially. We solve this problem by detecting the peaks and concavities of the smoothing energy curve within the continuation of label 2 sequence [15]. By recursively searching the maximum peak in the label 2 sequence and locating the points in front of and behind the maximum peak, most of the boundaries between syllables can be found. Fig. 7 illustrates the concept of this solution. The maximum point (1) in the sequence will be found first. Searching back and forward to locate the boundary between r1 and r2 and the boundary between r2 and r3. This procedure will be continued until the length of all syllables are less than the length of one syllable (the length of one syllable will almost not exceed 20 frames in our experiment) Fig. 8.

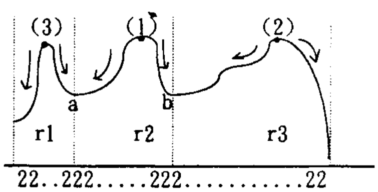


Fig. 7 Segmentation in connecting syllables with energy curve

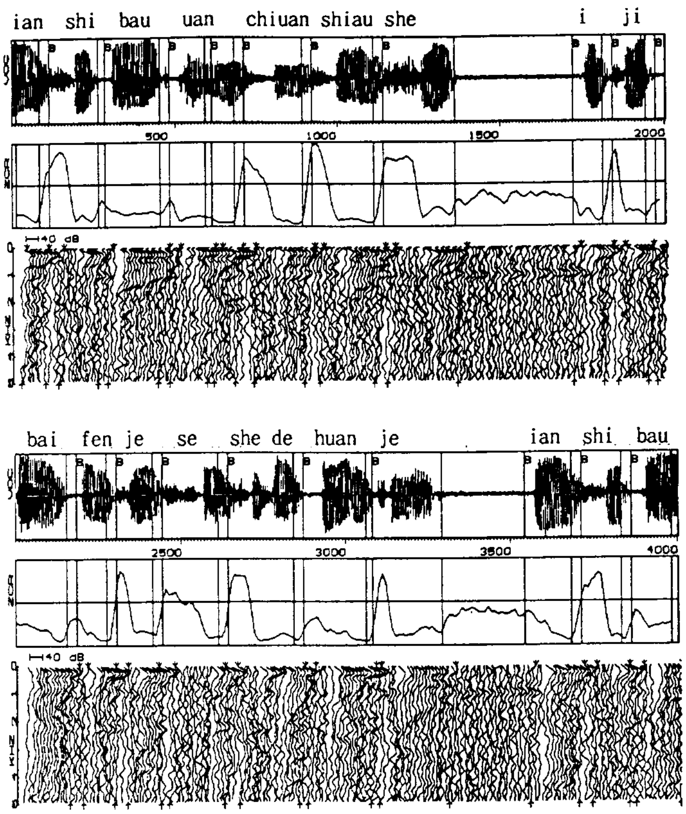


Fig 8. Female speaker "癌細胞完全消失 以及百分之四十的患者癌細胞"
(cancer cells completely disappeared and 40% cancer cells of patients)
The syllabic representation and lexical tone number are
"ian-2 shi-4 bau-1 uan-2 chiuan-2 shiau-1 she-1 i-3 ji-2 bai-3

In this case, only one undesired vowel syllablic unit is added. The whole result is very encouraging.

V. Concluding Remarks

In this paper, a model for speech segmentation is presented. The system formulations the speech recognition as a two-phase procedure. In the first phase, the novel class of hyperrectangular composite neural networks is utilized to classify frames of speech signal into three different class: (1) silence, (2) consonant, and (3) vowel. From experimental results, the HRCNN's is shown to be more efficient than the distributed fuzzy rules approach. In the following step, the concept of transition states is utilized as an appropriate control strategy to make right segmentation decisions. The experimental results seem rather encouraging. However, the transition states were found from observation of labeling sequences. In the near future, we will try to apply the HRCNN's to automate the process of finding transition states directly from given prototypes in order to make a real self-learning speech segmentation expert system.

References

[1]C. T. Hsieh and J. T. Chien, "Segmentation of Continuous Speech into Phonemic Units," IEICS pp. 420-424, 1991.

[2]L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances," THE BELL SYSTEM TECHNICAL JOURNAL. Vol. 54, No. 2, pp. 297-315 February 1975.

[3]R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1993.

[4]J. T. Tou and R. C Gonzalez, Pattern Recognition Principles, Addison-Wesley, Massachusetts, 1974.

[5]B. Everitt, Cluster Analysis, Wiley, New Your, 1974.

[6]H. Ishibuchi, K. Nozaki and H. Tanaka, "Distributed representation of fuzzy rules and its application to pattern classification," Fuzzy Sets and System 52, pp. 21-31, 1992.
[7]H. Ishibuchi, K. Nozak and N. Yamamoto, "Selecting fuzzy Rules by Genetic Algorithm for Classification Problems," 2nd IEEE International Conference on Fuzzy Systems, pp. 1119-1124, 1993.
[8]P. K. Simpson, "Fuzzy min-max neural networks-part1: Classification," IEEE Trans. on Neural Networks, Vol. 3, pp. 776-786, Sept. 1992.
[9]S. Abe and M. S. Lan, " A classifier using fuzzy rules extracted directly from numerical data," Proc. of the Second IEEE Int. Conf. on Fuzzy Systems, 1993.
[10]M. C. Su, A Neural Network Approach to Knowledge Acquisition, Ph.D. thesis, University of Maryland, August, 1993.
[11]C. T. Hsieh, M. C. Su, and C. T. Tseng, "A Mandarin recognition system based on a novel class of hyperrectangular composite neural networks, " IASTED Int. Conf. Modeling, Simulation and Identification, 1994.
[12]M. C. Su, C. T. Hsieh and M. N. Shiau, "A neural network approach to crispy and fuzzy if-then rules extraction from experimental data," 2nd National Conf. on Fuzzy Theory and Applications, 1994.
[13]L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, 1978.
[14]B. Kosko, Neural Networks and Fuzzy Systems : A Dynamical Systems Approach to Machine Intelligence, Prentice-Hall, NJ, 1992.
[15]C. T. Hsieh and S. C. Chien "Speech segmentation and clustering problem based on fuzzy rules and transition states," Twelfth IASTED Internation Conf. Applied Informatics, 1994.