# Applications of Fuzzy set theory in speech recognition

Peidong SONG Yoshinao AOKI
Faculty of Engineering, Hokkaido University, Sapporo-shi, 060 JAPAN

Abstract

In this paper , a new method is proposed based on fuzzy specch recognition for a speaker-independent . First the paper states the significance of applying the fuzzy set theory to the speech recognition , then briefly introduces the fundamental principle of fuzzy speech recognition , and states the method of setting up a reference dictionary by combining the fuzzy classification with linear matching and DP matching in speech recognition theory . Then it gives the fuzzy set expression of the speech recognized feature , and defines the resemblance degree between the input words and the standard pattern of the reference dictionary according to that expression . Finally , the paper gives the principle block and the implement block of this system . The average recognized rate 98.3% is taken by means of the statistical results from the experiments in this system .

Keywords: Fuzzy expression of the feature, A reference dictionary feature

## 1.Introduction

With the advance of computer technologies , the artificial intelligence technology is continuously permeating every field . People have fully felt the importance of artificial intelligence . The purpose of studying artificial intelligence is to have computer do some intelligent tasks which humans can do . The ultimate task of speech recognition is to transform computer into machines with ears .Therefore , speech recognition is an important component of the artificial intelligence technology .

In recent years , speech recognition technology has developed rapidly , new recognition methods have been brought forward one after another and various recognition systems have appeared[2] , yet , to apply them to practical systems without any problems , we can say we still have a long way to go . One of the main reasons is that the present speech recognition technology is based on accurate mathematical results and binary logic , and this method does not fit in with the human thought and judgement patterns which are based on fuzzy logic or the characteristics of the natural language with fuzziness .

Speech features are usually related to health , character of occupation , age , sex , physiological state , psychological state and mood . Therefore , We can say that fuzziness is one of the essential attributes of the natural language . Correctly describing the characteristics of a speech will directly affect the recognized rate , and so , it is one of the main subjects of study for which many people work . Because the fuzzy theory is an effective means of dealing with fuzziness problems , this paper applies the fuzzy theory to the field of speech recognition , and puts forward a speaker-independent fuzzy speech recognition method .

## 2.Speaker-independent fuzzy speech recognition [2] ~ [5]

### 2.1 The summaization of the basic method

The procedure for computer speech recognition is as follows : First , the computer transfirs the acoustic waves sent by a human which stand for a word meaning into numerical electronic signals , and extracts the features of the speech from these signals as input patterns . Then , according to a matching principle , it matches these input patterns with the patterns in the reference dictionary which were stored in the computer before , and selects the moot similar the pattern from them as the final recognized result . In terms of different reference dictionaries speech recognition can be classified into specified speaker speech recognition and speaker-independent speech recognition . What we call "specified speaker speech recognition " means that , before the computer recognizes a human's speech , we need to store the standard patterns of that human's speech in the computer , and form a reference dictionary . Then , the computer can only recognize correctly that human's speech . When the speaker is changed , the reference dictionary needs to be changed , too . Speaker -independent speech recognition has no limitation on the speaker which means that the reference dictionary is universal .

There are some differences in the speech generation organs of different speakers , so even though the word is the same , the speech features of different humans are different . Therefore , when we set up the reference dictionary of a speaker-independent speech recognition system , we can't analyse only one or two humans' speech , we must analyse a lot of humans' speech features . Since the internale storage space is limited , it is not possible to register many human's speech features of each word . Usually, the way of setting up the reference dictionary of the speaker-independent speech recognition system is to extract a lot of humans' speech features first , and then , to classify them , and finally , to take the average value of the different classes . This kind of classification is not precise , and it brings a great influence to the recognized rate . Therefore , this paper uses the fuzzy classification method in setting up the reference dictionary , i , e . classifying reasonably the speech features of many humans speaking the same word into different classes ,then , taking the average value within each class . Before classification , in order to make the monolization of the length of the different human's speech feature for the same word and the internal length of relativity in this word , we adopted the linear matching and the DP matching method . The purpose of this is to find out the precise and real resemblance relation which is needed when classifying features .

In this paper the speech signal feature used , is Fourier transform coefficients . As mentioned above , human speech is affected by many fuzzy factors , therefore these coefficients do not give an absolutely accurate expression in this system . We use the fuzzy set which is defined in the value field of the Fourier transform coefficients to describe the speech features . In terms of the description , we defined the resemblance degree of the recognition procedure for this system .

## 2.2 The setting up of reference dictionary

The interval used to measure and analyse the speech signals which vary with time is called a frame , the length of the interval is called the frame length , and the distance between the frames is called the frame interval . Since the speech signals which vary with time can be thought of as constants within 5∼20ms , and we measure and analyse the speech signal every 10ms or so , the speech signal's essential feature can be truly reflected . Therefore , the value taken for the frame length is 5∼20ms , and for the frame interval is 10ms or so usually .

Suppose , the H features of H humans speaking the same word is $\{x_1, x_2 ......, x_H\}$. In order to resonably classify these features into several classes by using fuzzy classification , we need first to set up the fuzzy resemblance relations among the H features , and we use the following

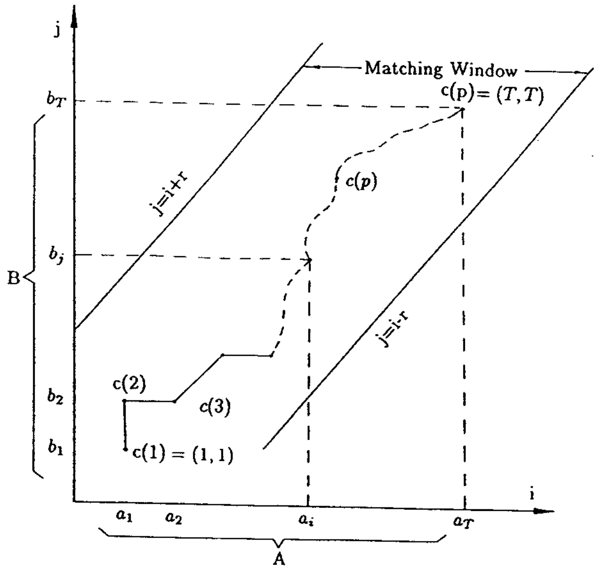

Fig. 1. DP algorithm

formula in this system[1] :

$$r_{i,j} = e^{-D(x_i,y_j)} \tag{1}$$

Where , $D(x_i, x_j)$ is the absolute value distance between $x_i$ and $x_j$ , and this paper , in terms of the speech features , adopts the DP distance by using the DP method after linear matching [3]. Linear matching makes the H different length features monolized , i, e . to make the frame number of each feature equal to the average frame number of the H feature of the word .The method is simple , so we won't mention it here . The DP distance is the distance after eliminating the influences of the variation of each corresponding length within the features . By the following algorithm , we know that the DP distance can reflect precisely and truly the differences among different speaker speech features for the same word .

Suppose A and B are arbitrarily two features in feature $\{x_1, x_2 ..., x_H\}$ after the linear matching . T is the frame number , then A and B can be expressed as :

$$A = \{a_1, a_2, ..., a_T\}$$
$$B = \{b_1, b_2, ..., b_T\}$$

where , $a_i (1 \leq i \leq T)$ is the Fourier transform coefficient of the ith frame in the feature A , and $b_j$ is the Fourier transform coefficient of the jth frame in the feature B .

The basic idea of the DP method is to find a path F = c(1) c(2)......c(p) from the matching window in Fig(1) , and to have the right side of equation (2) take the minimum value , and this minimum value is the DP distance between feature A and B , and is expressed as D(A , B) .

$$D(A,B) = Q^{-1} \bigwedge_{F} \left[ \sum_{p=1}^{P} d(c(p) \cdot \omega(p)) \right] \tag{2}$$

where ,

$$\omega(p) = (i(p) - i(p-1)) + (j(p) - j(p-1))$$

$$d(c(p)) = d(i,j) = |a_i - b_j|$$

$$Q = 2T$$

In order to find the discrete functional extrema value , we adopt dynamic programming . The method of doing this is as follows :

(1) The initial value setting :

$$g(1,1) = 2d(1,1) = 2d(a_1, b_1), i = 1$$

$$g(1,0) = g(0,1) = 0$$

(2) To take $j_1 = max(1, i-r), j_2 = min(i+r, T)$ to have j increased from $j_1$ to $j_2$ , its increment is 1 , and we use the recursion formula (3) to calculate g(i ,j).

$$g(i,j) = min \begin{bmatrix} g(i,i-1) & +d(i,j) \\ g(i-1,j-1) & +2d(i,j) \\ g(i-1,j) & +d(i,j) \end{bmatrix} \tag{3}$$

(3) if $i \leq T$ , then take i+1 as the new value of i to execute (2) , if i=T , then $D(A,B) = g(T,T)/Q$ and the calculation of (3) is finished .

From formula (2) , we find out the DP distance between $x_i$ and $x_j (1 \leq i,j \leq T)$ , where $x_i$ and $x_j$ are the different features of two human's pronounciation of the same word , then by using formula (1) , we can get the fuzzy resemblance relation between these features . After that , we do the fuzzy classification by using the maximum tree method

[7], and classify reasonably the H feature into a certain number of classes , and find the average value within each class . Finally , express these average values by using fuzzy set , and store them in the computer as a standard pattern in the reference dictionary . Considering the real-time of computer processing , the specific method of fuzzy expression is to suppose the total number of the words in the reference dictionary to be I , and by fuzzy classification , classify the reference patterns of each word into K classes , take $a_{ik}{}^{(t,l)}$ as the arithmetic mean value of the corresponding lth Fourier trans-form coefficient of the tth frame . Where , $1 \leq i \leq I)$ , $1 \leq k \leq K$ , $1 \leq t \leq T$ , $1 \leq l \leq L_a$ , $L_a$ is the total number of the Fourier transform coefficients in the tth frame . The corresponding standard pattern in the reference dictionary in the computer is a fuzzy set , and the definition of the membership function is as follows :

$$\mu_{a_{ik}{}^{(t,l)}}(m) = \begin{cases} 1 - (1/5)|m - a_{ik}{}^{(t,l)}| & , |m - a_{ik}{}^{(t,l)}| \leq 5 \\ 0 & , otherwise \end{cases} \quad (4)$$

If the feature of every word is processed by the above method the reference dictionary of the speaker-independent fuzzy speach recognition system mentioned by this article is formed . So we can see that the reference dictionary for this system is the multi-structure expressed by fuzzy set . i . e . each word corresponds to some standard fuzzy pattern . Therefore , this system has developed the structure of the speaker-independent speech recognition system of the usual multi-structure reference dictionary .

### 2.3 The definition of the resemblance degree .

In what way the inputed speech patterns match with the standard pattern in the reference dictionary , is the problem which is going to be solved in this section . The method adopted by this system is first to do the linear matching to the inputed recognized speech feature , and to have its frame number equal to the frame number of the ith word . The Fourier coefficients in each frame are expressed by using fuzzy set , and , in this way , form the fuzzy pattern X(i) of the waiting to be recognized speech which matches the standard pattern of the ith word ,and then , use formula (5) to match it to the standard pattern of the ith word in the reference dictionary .

Suppose , the lth Fourier coefficient in the tth frame of the feature of the waiting to be recognized speech is $a^{(t,l)}$ . The fuzzy expression calculated by formula (4) is $\mu_{a^{(t,l)}}(m)$ then the resemblance degree of $\mu_a{}^{(t,l)}(m)$ and the standard fuzzy pattern of the Ith word in the reference dictionary $S_{x(i),i}$ is defined as follows:

$$S_{x(i),i} = \bigvee_{k=1,2..:\mathrm{k}} \left[ 1/(L \cdot T) \sum_{t=1}^{T} \sum_{l=1}^{l} P_k(t,l) \right] \tag{5}$$

where ,

$$P_k(t,l) = \bigvee_{m} \left[ \mu_{a^{(t,l)}}(m) \bigwedge \mu_{a_{i,k}{}^{(t,l)}}(m) \right]$$

$$L = L_a \bigwedge L_{x(i)}$$

After we have calculated the corresponding resemblance degree of each word in the reference dictionary by using formula (5) , if there is a $I_0$th word in I words , and it satisfies the following formula(6):

$$S_{x(I_0),I_0} = max[S_{x(1),1}, S_{x(2),2} \cdots, S_{x(I),I}] \tag{6}$$

then we think that the $I_0$th word and the waiting to be recognized patterns are the most similar , i . e . we think that the waiting to be recognized pattern and the $I_0$th word are the same , therefore , we get the $I_0$ as the recognized output results .

## 3. System constitution

This paper puts forward the system constitution of the speaker inpdependent fuzzy speech recognition method as shown in Fig(2) , where the broken line expresses the setting up procedure of the reference dictionary . The pre-processing part is implemented by a special purpose hardware processing unit . Its function is , after smoothly filtering the electronic signals of the speech , to transfer the analogue speech signals into the numerical speech signals (12 bit ) under the 10 kHz physical sampling frequency , ready for the computer's read-save . The maximum tree principle is adopted in fuzzy classification [1] , and we use the Kurskal principle to find the maximum tree by using Fuzzy resemblance matrix . In the flow chart , the procedure from the speech feature extraction to the end is implemented on an FM-7 computer by using software . The Hamming window method is used in frame classification : the frame length is 12.8ms , and the period of frame is 10ms .

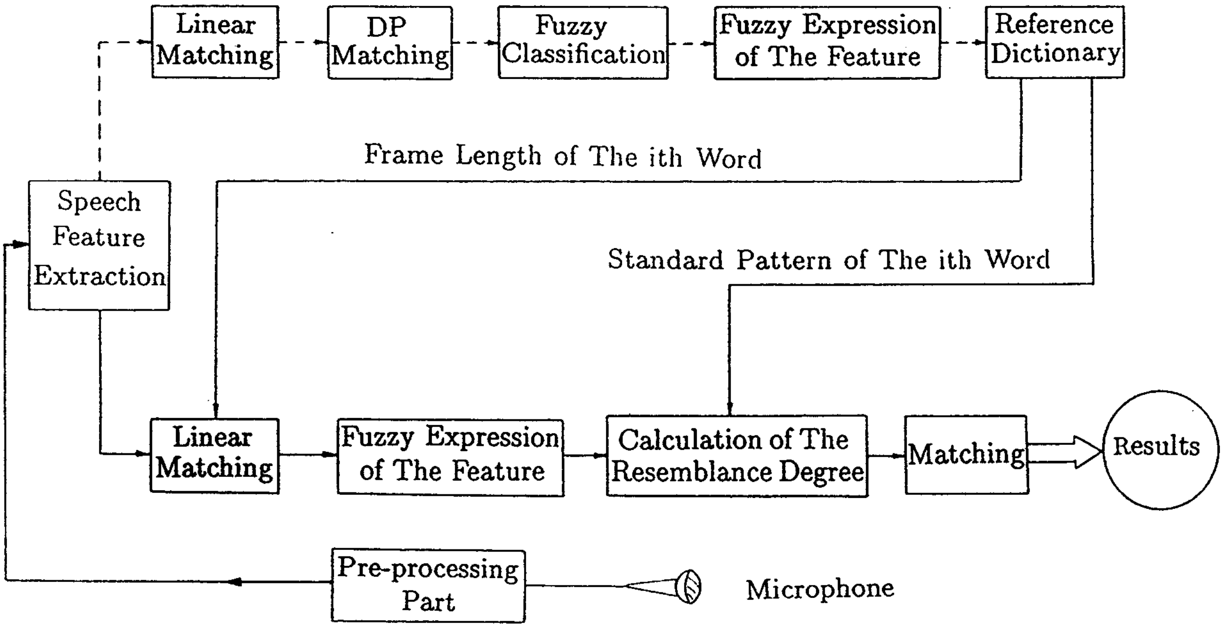
Linear Matching
DP Matching
Fuzzy Classification
Fuzzy Expression of The Feature
Reference Dictionary
Frame Length of The ith Word
Speech Feature Extraction
Standard Pattern of The ith Word
Linear Matching
Fuzzy Expression of The Feature
Calculation of The Resemblance Degree
Matching
Results
Pre-processing Part
Microphone

Fig. 2. System constitution of the speaker-inpdependent fuzzy speech recognition

## 4. The result of recognition

While setting up the reference dictionary , each word altogether 12 humans speech features including males and females , and registers 50 geographical terms . The speech environment is a quiet room without noise . The specific procedure is that the sound is transfered to the hardware interface board through a microphone , then , the computer stores the numerical speech signals , as the source voice data , on disk , ready for forming the reference dictionary . Fuzzy classification classifies the speech features of 12 humans speaking each word into three classes , i . e . the standard fuzzy pattern of each word in the reference dictionary is three . The fuzzy patterns expressed by fuzzy set in the reference dictionary are also stored on the disk ready for real-time invocation when needed for recognizing .

The recognition experiment is done by inviting 3 men and 3 women individually to do the practical analogue experiment . From the experiments , we find that the recognition rate is different for different people .The average recognition rate for females is a bit higher than that for males .The average recognition rate for this experiment is 98.3% . One other point which should be emphasized is that the six speakers all speak standard chinese .

## 5. Conclusions

This paper applies fuzzy theory into speech recognition technology , and puts forward a new method of speaker-independent fuzzy speech recognition .According to the experiments , we know that , because the fuzzy classification is used in setting up the reference dictionary , and the fuzzy set is adopted in speech features' description , the universality of the reference dictionary is improved . The recognition procedure is done by using the close degree between the fuzzy sets to recognize the results , therefore , this system gets a higher recognized rate . This paper is a very significant effort in the field of speech recognition .

The problems which still need to be solved , in order to make the whole system mone practical , are to realize the hardware description of the part algorithms and the modularization of the processing unit .

REFERENCES

1. WANG PEI-ZHUANG, Fuzzy set theory and applications , *The publisher of Science and technologie , shang-hai CHINA*
2. NAKAGAWA and NAKAMOTO Speaker-independent large vocabulary word recognition based on syllable by syllable input, *The transactions of the institute of electronics, information and communication engineers of JAPAN J65-D 1982*
3. NITTA and MURATA, Speaker-independent word recognition using multiple similarty method , *The transactions of the institute of electronics, information and communication engineers of JAPAN J67-A 1984*
4. MUROI and NAKAGAWA, Speaker-independent word recognition using partial linear expansion and weighted average templates *The transactions of the* it institute of electronics, information and communication engineers of JAPAN *J69-A 1986*
5. NAKAGAWA and ENOMOTO, Speaker-independent phoneme and word recognition by statistical classification method for time-sequential patterns *The transactions of the institute of electronics, information and communication engineers of JAPAN J71-A 1988*