Automatic Log-in System by the Speaker Certification

Young-Sun Sohn

Dept. of information & communications engineering, Tongmyong University of information technology, Korea

Abstract

This paper introduces a Web site login system that uses user's native voice to improve the bother of remembering the ID and password in order to login the Web site. The DTW method that applies fuzzy inference is used as the speaker recognition algorithm. We get the ACC(Average Cepstrum Coefficient) membership function by each degree, by using the LPC that models the vocal chords, to block the recorded voice that is problem for the speaker recognition. We infer the existence of the recorded voice by setting on the basis of the number of zeros that is the value of the ACC membership function, and on the basis of the average value of the ACC membership function. We experiment the six Web sites for the six subjects and get the result that protects the recorded voice about 98% that is recorded by the digital recorder.

Key words : fuzzy inference, speaker recognition, voice, automatic login, intelligent

1. Introduction

The concern about the customer certification becomes larger complying with the abrupt increment of the service connected with the internet.[1]

Among the methods of the customer certification, the ID and password is the best useful method because of the easy realization and the special equipment is not necessary. But because of the defect that is weak to the illegal irruption by the leakage of the password, recently biometrics, which uses human's somatological characteristic that is a fingerprint, a face and a voice, etc. has been studied.[2,3] The recognition using the voice can certify the customer easily in comparing with the other somatological recognition, because it is easy to use and is not expansive for only using the microphone.[4,5] But if one records the user's voices and plays it, then the system recognizes it as the true user[6].

In this paper, we realize the speaker certification system, which protects the recorded voice by using the LPC cepstrum[7] that models the human's vocal chords.

2. An outline of the system

If the voice is inputted, the system recognizes the voice to recognize the speaker's each character by applying the voice recognition algorithm. After that, the system determines the acceptance and rejection of the recognized voice through the fuzzy inference algorithm for the similarity of the DTW. If the inputted voice is rejected as a user's voice, the system announces the reentrance of the voice to the user. In case of an ambiguous voice, the system asks the user to decide. If the recognized voice is decided as the recorded voice by the recorded voice interception algorithm, the system ended. If it is decided as the ambiguous voice, the system announces the reentrance of the voice to the user. If it is decided as the actual voice, the system logs-in the Web-site and the user can use the Web site.

3. speaker recognition algorithm[8]

We use the pattern matching method to process the voice recognition. Extracting the character of the voice pattern from an inputted voice makes a standard pattern. If a voice is inputted, then it searches and recognizes the most similar pattern compared with the preserved standard patterns. Fig. 3-1 shows the concept of the voice recognition system.

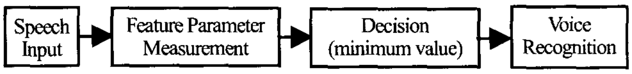


Fig. 3-1. Conception chart of speech recognizing

The whole algorithm for the recognition of the inputted voice is constructed as shown in Fig. 3-2.

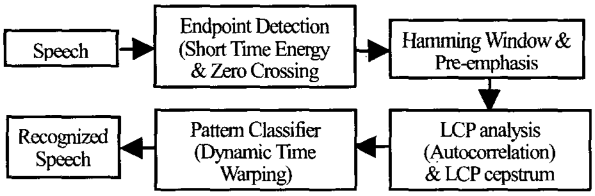

Speech
Endpoint Detection (Short Time Energy & Zero Crossing
Hamming Window & Pre-emphasis
LCP analysis (Autocorrelation) & LCP cepstrum
Pattern Classifier (Dynamic Time Warping)
Recognized Speech

Fig. 3-2. Flow chart of speech recognizing

3.1 Extraction of a block of voice in real time

We use the short time energy method and the zero-crossing method, which has smaller calculation time than other methods, to find out whether the voice is present or not under

Manuscript received Jun. 3, 2004; revised Sep. 3, 2004

the state of receiving the voice with on-line.

3.1.1 The short time energy method

This method judges whether the voice is present or not on the basis of absolute energy value within the block of the voice analysis. Because the waveform of a voiced sound has large absolute energy value and the waveform of the voiceless sound has low absolute energy value, when the absolute energy value of an analysis block with equation (1) exceeds the upper limit value of energy, we consider that there is a voice.

$$E = \sum_{i=0}^{N} | \chi(i) | \tag{1}$$

3.1.2. The zero-crossing method

The phase of the voice signal waveform passes through an axis, so this method uses the number of passing through time axis. Since a voiceless sound has not the large amplitude and is the irregular oscillation, it has large zero-crossing rate than the voiced sound. Zero-crossing rate is defined by the equation (2).

$$ZCR = \sum_{i=0}^{N-1} | \mathrm{sgn}(\chi(i)) - \mathrm{sgn}(\chi(i+1)) | *(1/2)$$

$$\mathrm{sgn}(\chi) = 1(\chi \geq 0)$$
$$\mathrm{sgn}(\chi) = -1(\chi < 0) \tag{2}$$

3.2 LPC(Line Prediction Coding) transaction process

We use the LPC coefficient to get the characteristic information of the user's voice. LPC assumes the human vocal organ as a filter, and uses the coefficient of the filter as the characteristic vector of the voice. The mathematical model of a vocal organ is shown in Fig. 3-3.

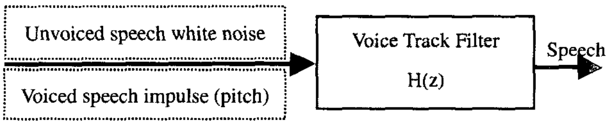


Fig. 3-3. Mathematic modeling of voice organ

The transaction process of the windowing and the pre-emphasis is necessary for the detailed quantization before the coding of LPC within the block of the voice.

3.2.1 Windowing

Windowing is an equation to analyze one part of the continuous voice signal, and is a process that separates the data with the length that the user wants uniformly. In this paper, we use Hamming windowing that presents the occurrence of the noise component by decreasing the signal distortion in the neighborhood of the boundary of the analysis block. Hamming windowing reduces the magnitude of the signal data increasingly by approaching to the edge of the analysis block. Hamming windowing is defined by equation (3).

$$W(n) = 0.54 - 0.46 \cos(\frac{2\pi n}{N-1}),\ 0 \leq n \leq N-1 \tag{3}$$

3.2.2 Pre-emphasis

The detailed high-frequency waveform, which is near the central axis within the analysis block, is very small to the low frequency, so it is hard to analyze the precise waveform. Therefore, pre-emphasis process extracts the information of the waveform by strengthening the component except the existing frequency. That is, the energy of the voice signal is reduced in the low frequency band and it is increased in the high frequency band. It is defined by equation (4).

$$\bar{\chi}(n) = \chi(n) - \alpha\, \bar{\chi}(n-1),\ 0.9 \leq \alpha \leq 1.0 \tag{4}$$

3.2.3 The LPC analysis

LPC analysis is a method that forecasts the present sample using past P samples. The relation is defined by equation (5).

$$y(n) = \chi(n) + \sum_{k=1}^{P} \alpha_k \bullet y(n-k) \tag{5}$$

The autocorrelation method, which has good safety, is used for the LPC analysis. The autocorrelation method, as defined by equation (6), represents the grade of the similarity between the samples of two signals, which are separated with any uniform distance, as the function of the relation between the distance and the similarity. Durbin's method is used in the process of yielding the LPC coefficient. This method has the advantage that can know the present pattern information by knowing the information of just previous pattern, so it has an advantage of calculation.

$$R(i) = \sum_{k=1}^{P} \alpha_k R(| i-k |) \tag{6}$$

3.3 The cepstrum [6]

Because the cepstrum is the reverse transformation of the function of frequency domain, it may be the function of time domain. The most powerful characteristic is having the capability that separates the voice information into the detailed structured information and the spectrum envelope information. The obtained LPC coefficient can get the LPC cepstrum coefficient by equation (7).

$$\hat{C}_1 = -\alpha_1$$
$$\hat{C}_n = -\alpha_n - \sum_{m-1}^{n-1} \frac{m}{n} \alpha_m \hat{C}_{n-m} (1 < n \leq p)$$
$$\hat{C}_n = \sum_{1}^{p} \frac{m}{n} \alpha_m \hat{C}_{n-m} (p < n) \tag{7}$$

We use the tempered window technique to relieve the sensitivity of the obtained cepstrum coefficient as defined by equation (8).

$$W_m = [1 + \frac{Q}{2} \sin(\frac{\pi m}{Q})], (1 \leq m \leq Q) \tag{8}$$

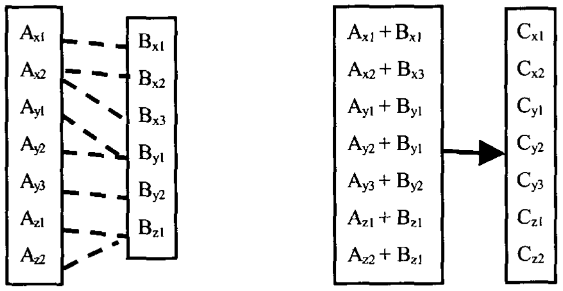
And we use the delta cepstrum, which applies the time difference, for the representation of the more improved and extended spectrum as defined by equation (9).

$$\frac{\partial}{\partial w}[\log \ |S(e^{jw,t})|] = \sum_{m=-\infty}^{\infty} \frac{\partial C_m(t)}{\partial t} e^{-jwm}$$

$$\frac{\partial C_m(t)}{\partial t} = \Delta C_m(t) \approx \mu \sum_{k=-K}^{K} kC_m(t+k) \tag{9}$$

3.4 DTW method

Even when the identical man speaks the same word, the time length of the word is changed at each time of the utterance. Therefore, if the inputted word is simply compared with the standard pattern, since the time axis is not equal, the error on the incapability of the recognition will be occurred. To resolve this, the change pattern of the non-linear time axis is linearly normalized to classify the patterns. The DTW method matches the standard pattern and the inputted pattern, and so makes the other new pattern and recognizes it as shown in Fig. 3-4.



(a) Patterns with different frame length (b) New reference patte

Fig. 3-4. Pattern matching

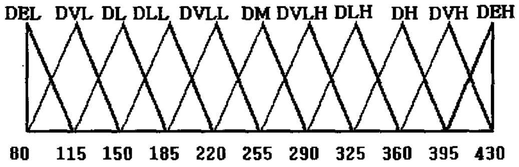
As the method of measuring the similarity, the method of the normalized distance is used as known in equation (10). The method of the normalized distance is calculated by the error between the standard pattern and the inputted pattern. From the result of calculation, the smaller error distance is, the more similar frame is.

$$D(T_x, T_y) = \min \sum_{k=1}^{T} d(\varphi_x(k), \varphi_y(k)) m(k) \tag{10}$$

Since the DTW method for an isolated word has higher recognition rate than a serial word, this method is much used for an isolated word. Since the command, which operates the cursor, is recognized as the isolated word, we use the DTW method in this paper. Any pattern, which has smallest error distance by comparison of the 6 standard pattern with the inputted word pattern among the total 6 words, is recognized as the correspond word.

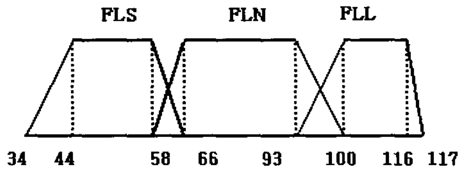
4. Fuzzy inference algorithm to the critical value of the DTW[8]

The membership function of the DTW distance error and the membership function of the frame length for a subject is shown in Fig.4-1 and Fig.4-2.



DEL : Distance error value of DTW is Extremely Low.
DVL : Distance error value of DTW is Very Low.
DL : Distance error value of DTW is Low.
DLL : Distance error value of DTW Little is Low.
DVLL : Distance error value of DTW is Very Little Low.
DM : Distance error value of DTW is the Middle.
DVLH : Distance error value of DTW is Very Little High.
DLH : Distance error value of DTW is Little High.
DH : Distance error value of DTW is High.
DVH : Distance error value of DTW is Very High.
DEH : Distance error value of DTW is Extremelv High.

Fig. 4-1.The membership function of the DTW distance error



FLS : Frame Length is Short
FLN : Frame Length is Normal
FLL : Frame Length is Long

Fig.4-2. the membership function of the frame length

Throughout the several experiments, the allowable range of the DTW error value by the frame length is constructed as shown in TABLE 1.

Table 1. The fuzzy inference rule

DTW / frame	DEL	DVL	DL	DLL	DVLL	DM	DVLH	DLH	DH	DVH	DEH
FLS	CD	CA	CA	CA	CA	CD	CD	CR	CR	CR	CR
FLN	CR	CD	CA	CA	CA	CA	CA	CD	CR	CR	CR
FLL	CR	CR	CD	CA	CA	CA	CA	CA	CA	CD	CD

The membership function of the acceptance and rejection as the user's voice is inferred by the calculated distance error value and the frame length of the inputted pattern as shown in Fig.4-3.

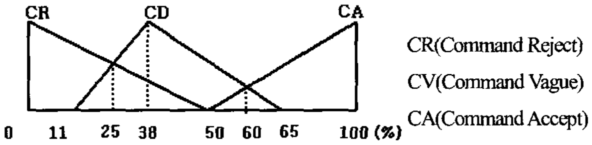


Fig. 4-3. The membership function of the acceptance and rejection as the user's voice

5. The recorded voice interception algorithm

There is a method that registers previously and presents randomly the short words, as the existing recorded voice interception algorithm.[9] But if one refers to the presented words and plays them by using the newest electronic record equipments, the method has no meaning.[6] In this paper we propose the recorded voice interception algorithm that uses ACC(Average of Cepstrum Coefficient) to intercept the recorded voice with the digital recorder, which has the good quality of sound.

5.1 The characteristic extraction for the recorded voice interception

Because of the LPC analysis is based on the all-pole model that considers only the spectrum pole, the LPC cepstrum follows the spectrum pole. We use cepstrum coefficient, which represents the spectrum envelope information, to protect the recorded voice.[4,7].

We classified the characteristic between the actual voice and the recorded voice with obtaining the ACC by each degree as the equation (11), because even when the identical man speaks the same word, the frame length of the word is changed at each time of the utterance.

$$ACC_k = \sum_{i=1}^{N} (C_k)_i / N \qquad (11)$$

C_k : cepstrum coefficient of the k-th degree
N : The frame length of a word

According to the characteristic of the speaker and the word, ACC has the different distribution between the actual voice and the recorded voice. As an example, Fig.5-1 shows the ACC characteristic comparison between the actual voice and the recorded voice in case that the subject speaks 'naver'.

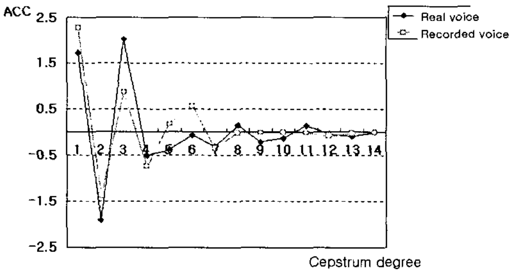


Fig.5-1. ACC characteristic comparison between the actual voice and the recorded voice.

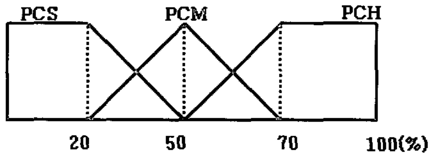
5.2 The fuzzy inference for the interception of the recorded voice

The subjects speak 6 words 30times respectively and we get the ACC membership function for the subjects and words. We obtain the value of the ACC membership function by each degree as the equation (12).

$$\tilde{A} = \sum_{i=1}^{C} \mu_{\tilde{A}} (ACC_i) / ACC_i \qquad (12)$$

C : The degree of the cepstrum
Ã : the value of the ACC membership function

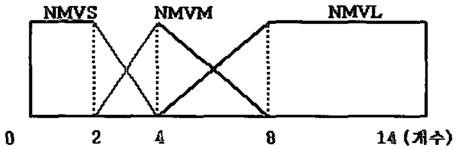
The membership function for the percentage of content of the actual voice, which is obtained from the average value of ACC membership function, is shown in FIg.5-2.



PCS : Percentage of Content is Small
PCM : Percentage of Content is Middle
PCH : Percentage of Content is High

Fig. 5-2. The membership function for the percentage of content of the actual voice

If the percentage of content of the actual voice is high, then inputted voice is actual voice, if low, then it is recorded voice. We count the number of the zero value of the ACC membership function by each degree to reduce the mistaken recognition of the middle range of the percentage of content. The zero value of the ACC membership function means that it is not included in the range of the actual voice. The membership function of the zero value of the ACC membership function is shown in Fig.5-3.



NMVS : Number of the zero Membership Value is Small
NMVM : Number of the zero Membership Value is Middle
NMVL : Number of the zero Membership Value is Large

Fig.5-3. The membership function of the zero value of the ACC membership function

In case of the middle range of the percentage of content, if the number of the zero value of the ACC membership function is few, then it is decided as actual voice, if many, then it is decided as recorded voice. In case of the number of the zero value of the ACC membership function is in the middle range, if the percentage of content of the actual voice is high, then it is decided as actual voice, if low, then it is decided as recorded voice. The actual voice decision, from the percentage of content of the actual voice and the number of the zero value of the ACC membership function, is constructed as the TABLE 2 and the membership function of that is shown in Fig.5-4.

In case that the result is under 55%, the system decides the inputted voice as the actual voice and logs-in the Web site. If the result is 75% or more, the system decides the inputted voice as the recorded voice and ends the system. When the value of the result is between 55% and 75%, the system

Table 2. fuzzy inference rule table

PERCENTAGE OF CONTENT / NUMBER OF ZERO MEMBERSHIP VALUE	PCS	PCM	PCH
NMVS	OV	AV	AV
NMVM	RV	OV	AV
NMVL	RV	RV	OV

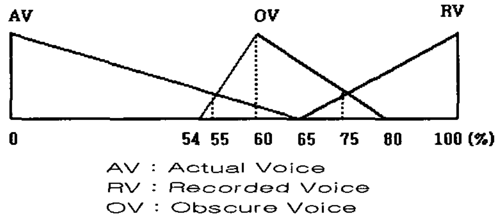


Fig. 5-4. The membership function of the existence of the actual voice

decides that the inputted voice is ambiguous so the system receives the reinputted voice. If the number of times of the reinput is 3 times or more, then the system decides the inputted voice as the recorded voice and close the system.

6. The Web-site automatic log-in system

In the already established system, firstly the user enters the site after inputs the URL in the address window, next inputs the ID and the password, and then clicks the transmission button. As shown in Fig.6, if the user's voice is recognized, our system links to the appropriate site by using the stored site address. If the appropriate site uses the frame, iframe and script according to the windows presentation method, the user's system gets the appropriate data. The system extracts the address of log-in Web site, ID, password and hidden name from the obtained data, and mixes the password and the user's account. The compounded information is transmitted with the post method that conquests the size limitation of the transmitted data and prevents the information outflow of the user's ID and the password. According to the above mentioned procedure, the window, which the user can use the logged-in Web site, appears on the monitor of the user's system.

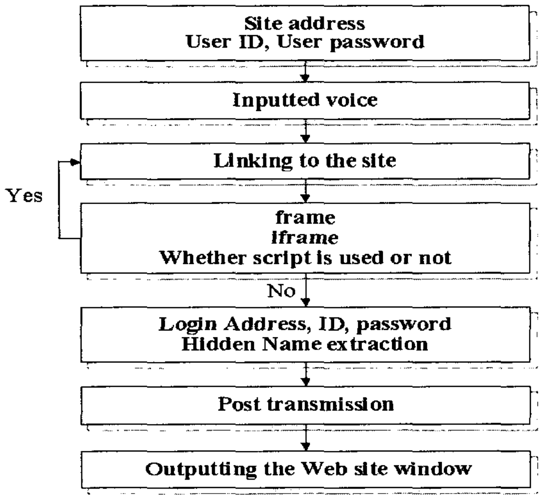


Fig. 6. The system conception of the automatic Web site login

7. The experiment and the result

7.1 The experiment

For the speaker recognition experiment, we accumulate the voice patterns of 6 site name 10times utterance respectively to the 3 boy-students and the 3 girl-students of the Tongmyong University of information technology. For the recorded voice experiment, we recognize the speaker and store it as wav file simultaneously. The standard membership function is made by the 30 actual voice pattern for the each word of site name. We experiments the recognition with the 30 actual voice pattern and 120 digital recorded voice pattern. Fig.7 shows the recorded voice generation process for the experiment of the recognition rate.

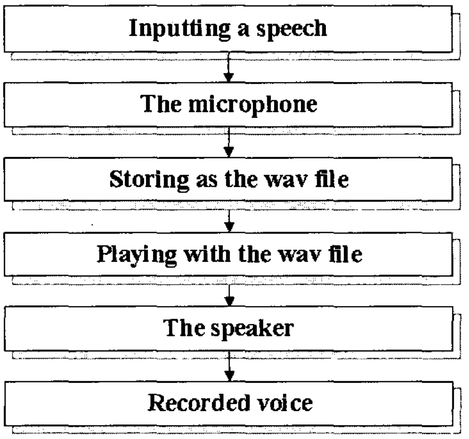


Fig. 7. The recorded voice generation process

Recording place is the general laboratory that there exist surrounding noise, and we record the voice with PCM1125Hz, 16bit mono.

The voice pattern is changeable according to the microphone characteristic, the transmission line and the background noise, therefore we experiment with two kinds of the microphone, which have the characteristic as shown in TABLE 3.

Table 3. The microphone characteristic

mike / characteristic	mike1	mike2
Manufactured good	Ceron 500	ACC-M3
type	Dynamic mike	Dynamic mike
Fixed direction	Single direction	undirectional
sensitivity	78dB	70dB
Frequency range	70-15,000Hz	50-17,000Hz
Regular impedance	600 ohms	600 ohms

We use microphone1 to make the standard pattern and to recognize the actual voice. TABLE 4 shows the experiment table of the 120 recorded voice according to the microphone characteristic.

Table 4. The experiment table according to the microphone

experiments cases	record	recognition
case1	mike1	mike1
case2	mike1	mike2
case3	mike2	mike1
case4	mike2	mike2

7.2 The result of the experiment

TABLE 5 shows the result of the recorded voice experiment according to the each cases. As known in TABLE 5, when the recorded voice is inputted to the system, this intercepts that with average 90%, the mistaken recognition as the actual voice is 2% and the ambiguous decision, if the inputted voice is real or recorded, is 8%.

Table 5. The interception rate of the recorded voice

result(%) cases	interception	ambiguous	recognition
case1	91	7	2
case2	92	7	1
case3	88	10	2
case4	89	8	3

8. conclusion and the problem in future

We propose the system, which decides if the inputted voice is recorded voice or not by using the cepstrum coefficient and the fuzzy inference to improve the problem that user's recorded voice is mistaken recognized as the actual voice. According to the characteristic of the microphone voice pattern is changed, therefore we generate the recorded voice using two kinds of the microphone and get high recognition rate, protecting the recorded voice with 98%, from the result of the experiment.

As the problem in future, recorded voice interception experiment, using the microphone and the speaker of various kinds, is considered. Searching the generalized method by the experiment of the various age is considered, too.

References

[1] Jeong-Gak Lyu, Gun-Hee Lee, Tae-Shik Shon, Song-Hwa Chae, Dong-Kyoo Kim, "Modified Implementation of Authentication using Message Digest in WWW", Proceedings of the Korean Information Science Society Fall Conference, Vol.28, No.2, PP.691-693, 2001.

[2] Hyoun-Joo Go, Sang-Won Lee, Myung-Geun Chun, "Iris Pattern Recognition for Personal Identification and Authentication Algorithm", The KIPS Transactions : PartC, Vol.8-C No.5, PP.499-506, 2001.

[3] Phil-Joong Lee, Ju-yeon Cho, "A Study of Identity Verification by Biometres", Journal of Korea Institute of Information Security & Cryptology, Vol.2, No.4, PP.67-74, 1992.

[4] Younh-Hwan Oh, "Information Processing of Voice Language ", Hongrung Press, 1998.

[5] Hyeong-Soon Kim, "Technical tendency of the voice recognition", The Magazine of the IEEK, Vol.22, No.5, PP.529-540, 1995.

[6] Hyun-Yeol Jeong, "The present condition and the prospect of the technique of the speaker recognition system using voice", Communications of the Korea Information Science Society, Vol.19, No.7, PP.32-44, 2001.

[7] L.R.Rabiner, B.H.Juang, "Fundamentals of speech recognition", Prentice Hall, 1993.

[8] Myung-Kyung Chu, Young-Sun Sohn, "Cursor Moving by Voice Command using DTW method", Journal of Fuzzy Logic and Intelligent Systems, Vol.11, No.1, PP.3-8, 2000.

[9] Kwang-Seok Seo, You-Shik Shin, Chong-Kyo Kim, "The Text-Prompt Speaker Recognition for Customer Discrimination", Proceedings of the Acoustical Society of Korea Conference, Vol.17, No.1, PP.127-130, 1998.

Young-Sun Sohn
received B.S., M.S. degree in electronics from Dong-A University in 1981, 1983. From 1990 to 1998, he has been senior researcher in ETRI. He received Ph.D. degree from Tsukuba University in Japan. From 1998 to present, he is an Associate Professor, Dept. of Information & Communication Engineering college of Engineering, Tongmyong University in Korea. His research interests include Human Interface, Fuzzy Measures and Fuzzy Integrals, and Evaluation.