윈도우 환경에서 음성을 이용한 사용자 확인에 관한 연구

User-Identification on WINDOWS Environmemt by Using the Speech

정종순*, 배재옥*, 배명진*
(JongSoon Jung*, JaeOk Bae*, MyungJin Bae*)

요 약

본 논문은 윈도우즈 95와 같은 멀티미디어 환경 하에서 개인신분 확인 기능을 DTW 이용하여 수행하였다. 즉, 개인신분 확인을 위한 기존 방법으로는 비밀번호를 키보드로 입력받는 것이었으나, 본 논문에서는 음성을 이용하였다. 본 논문의 중요한 특징은 다음과 같다. (1) 최근의 음성패턴으로 갱신하기 위해서 F_1/F_0율을 구하여 사용하였다. 이 방법은 시간 흐름에 따른 인식율이 저하되는 것을 최소화 하기 위한 것이다. (2) 화자간의 변별력을 극대화하기 위하여 가중 켑스트럼을 사용하였다. 즉, 가중 켑스트럼은 화자별로 유용한 켑스트럼 차수를 구하여, 그 차수에 가중치를 두는 것으로 F-ratio 값을 사용하였다.

제안된 방법으로 실험한 결과, 기존의 DTW 방법을 이용한 것보다 인식율이 5%이상 개선 되었다. 따라서, 윈도우즈 환경에서 비밀번호 사용 대신 음성 사용에 대한 가능성을 보여 주었다.

ABSTRACT

In this study, we implement individual verification system for multimedia environment such as WINDOWS 95 by using DTW(Dynamic Time Warping). The conventional method for speaker recognition uses the password through the keyboard. However, this paper uses speech. The major feature of this study is summarized as follows: (1) We make completely reference pattern by updating new speech pattern with F_1/F_0 ratio. This method keeps the high recognition rate compared with the other systems whose performances degrade rapidly as time goes on. (2) We use F-ratio values as the weighted values of the cepstral coefficients. We find that the weighted cepstrum reveals an effect on intensifying the difference between the customer and the imposter. Also the speaker recognition rate is improved more 5% than the conventional DTW pattern matching with cepstrum. This shows the possibility that speech signal can be used as means of individual verification for WINDOWS environment.

I. Introduction

As most information media of today becomes the multimedia environment, many operations are made up in WINDOWS environment than DOS environment. So the individual verification in the WINDOWS is an important problem. Accordingly, all user verification function is required in any operation or directory. For example, it makes use of the password for 'share folder', for 'whether use network or not' and for 'screen saver' etc. However, this method has the risk of plagiary. Especially in case of the information access accomplished at a long distance through telephone or communication network, individual verification is a difficult problem. On the other hand, a security system using the verification of fingerprints has the demerit of requiring high cost for additional equipment. Speech signal includes "linguistic information for transfer meaning" and "speaker information - who says?". Our study in this paper aims to develop a secure individual verification system for WINDOWS environment by using speaker verification techniques.

Traditional speaker verification systems have several problems as follows: 1)They use several reference patterns for combating intra-speaker variations. This results in heavy computational requirement. 2)Their performances degrade rapidly as time goes on.

We attempt to solve the problems as follows. In our system, we use only one reference pattern by averaging several patterns using DTW and F_1/F_0 ratio for updating the new pattern. To maximize the inter-speaker variation, we use the weighted cepstrum by F-ratio value.

This paper is organized as follows. In section II, we describe the speaker recognition system by pattern matching technique. This system is our base system which is

* 숭실대학교 정보통신전자공학부

접수일자 : 1997년 12월 30일

applied with betterment of section III and section IV. In section III, we provide the construction method of average reference pattern using DTW algorithm and F_1/F_0 ratio using quantization error for updating the new reference pattern. We explain the weighted cepstrum which is the feature parameter of our recognition system in section IV. We show the overall structure of the proposed system in section V. In section VI, we discuss the experimental results and the performance of the proposed system. Conclusions are followed in section VII.

Ⅱ. The Speaker Recognition System

The speaker recognition system is divided into various classes. According to the recognition techniques, the system is classified into pattern matching, neural nets, vector quantization and hidden markov model, etc. The pattern matching system has an effect on the application part. The reason is that it needs relatively simple algorithm and minimum hardware. In this paper, we use the pattern matching technique for simplicity. According to the recognition subject, the system is classified into the speaker identification and the speaker verification. Given a candidate speaker, speaker identification system attempts to answer the question, "Who is he?". A speaker verification system addresses, "Is he whom he says he is?". Whether the text of recognition system is fixed or not, the system is classified into the text dependent system or the text independent system[1].

Our system is the text dependent speaker verification system where the used text is the customer's name. The text independent system needs the large size of speech data for speaker recognition. This is not applicable as means of individual verification at WINDOWS environment.

The Figure 2.1 shows the block diagram of the speaker verification system using pattern matching technique.

The inputs of speaker verification system are customer's speech. Input speech pass through the many procedures such as the endpoint detection, the feature extraction and the pattern matching. The system declares the speaker as the customer or imposter by the distance value between the two patterns.

2.1 End-Point Detection

The correctness of speech end-point detection has an important effect on the performance of speech/speaker recognition system. In this paper, we use the short-term energy and zero crossing rate (ZCR) parameters. We control the threshold value by adaptive threshold value method. The short-term energy parameter considers the large energy frame as speech, the small energy frame as silence. This method fails to distinguish speech from silence in case of the small energy speech data or large background noise. To make up for the weak points in this energy parameter, we use the ZCR.

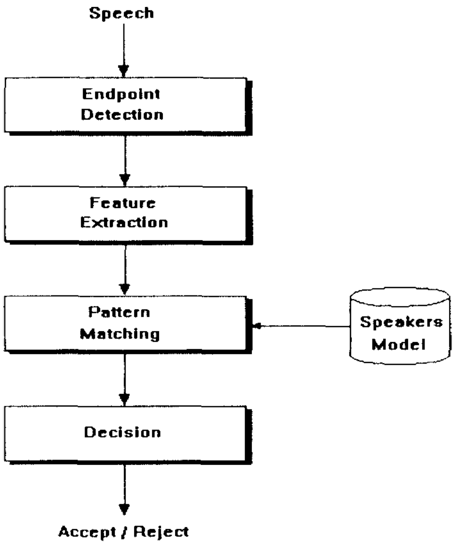


Figure 2.1 The block diagram of the speaker verification system using pattern matching technique.

We also consider the pause of interword. In case of the pause of interword, we look upon as the speech when starting point is detected within 40ms after endpoint detection.

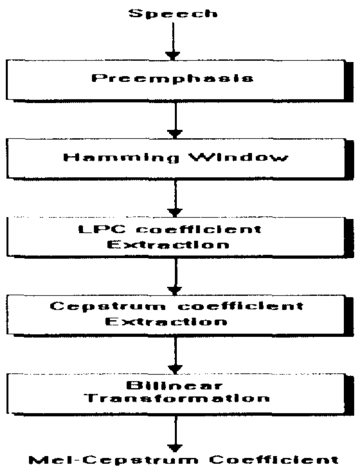


Figure 2.2 The feature extraction process.

2.2 Feature Extraction

The frame length is 30ms and the frame period is 10ms. We use the Hamming window. The 16 order LPC coefficients are extracted from speech data that pass through the pre-emphasis processing. We obtain the cepstra from LPC coefficients. The mel-scaled cepstra are extracted by warping module that transforms the LPC coefficients on the basis of mel-scale. The figure 2.2 shows the feature extraction process.

2.3 Pattern Matching Using DTW

Utterances are generally spoken at different rates, even when a single speaker repeating the same word. Thus, test and reference utterances normally have different duration. Most high-performance speaker recognizers address the problem of alignment by nonlinear warping one template onto another in an attempt to align similar acoustic segments in the test and reference templates. The procedure, called Dynamic Time Warping(DTW), combines alignment and distance computation through a dynamic programming procedure[2].

Deviations from a linear frame-by-frame comparison are allowed if the distance for such a frame pair is small compared to other local comparisons. In the absence of specified segment boundaries, DTW aligns templates by finding a time warping that minimizes the total distance measure, which sums the frame distances in the template comparison. The warping curve is derived from the solution of an optimization problem.

$$D = \min_{w(n)} [\sum_{n=1}^{T} d(T(n), R(w(n)))]$$

where each d(,) term is a frame distance between the n-th test frame and the w(n)-th reference frame. D is the minimum distance measure corresponding to the 'best path' w(n) through a grid of TR points. We select the ID of reference pattern that have the minimum D out of the whole speaker's reference pattern.

Many speaker recognition systems are published[3]-[5], which use the pattern matching.

III. Generation of Reference Pattern

3.1 F_1/F_0 ratio using quantization error

M-bit linear quantized speech signal s(n) is given in eq (3-1)

$$s(n) = \sum_{i=0}^{M-1} a_i 2^i = \sum_{i=0}^{N-1} a_i 2^i + \sum_{i=N}^{M-1} a_i 2^i$$
$$= Q_L + Q_H \qquad (3\text{-}1)$$

where Q_L is quantization error that occurred when speech signal is quantized to (M-N)bit. In case of voiced waveform, energy of lower formant is higher than higher formant. Therefore, same as fig.3.1(b), both fundamental frequency which has higher energy and the first and the second formant components maintain maximum amplitude of Q_L. Otherwise higher formants of low energy are changing rapidly in the amplitude range of Q_L. Extraction of highly correlated sector through quantization error Q_L is obtaining the normalized amplitude character which is limited in 2^n ± range. This is reducing the effect of searching pitch period according to change of speech amplitude in time domain. For an example, a normalized waveform of the first and the second formants that have the high energy is obtained by the quantization error and is shown in fig.3.1(c).

First of all, as shown in the following eq.(3-2), the interval between N_S and N_E is extracted. Here N_S is the starting point and N_E is the ending point crossing the threshold value in one frame of the normalized quantization error. And the average pitch period is obtained by the interval that is divided by RTCR(Rising Threshold level Crossing Rate) between N_S and N_E.

$$PITCH(fr) = \frac{RTCR}{N_S - N_E} \qquad (3\text{-}2)$$

The pitch value extracted on the *fr*-th frame by the eq.(3-2) must be within the pitch existence interval between 2.5ms and 25ms. If the obtained value is less than 10% of the interval between two rising threshold level crossing points in the frame, the pitch period is considered as the correct pitch value. If this condition is not satisfied, this waveform is considered as unvoiced fricative, unvoiced affricate and stop sound. Because the inverse of the Zero Crossing Interval(ZCI) over one pitch period is equal to the frequency of $2F_1$, the first formant is easily obtained.

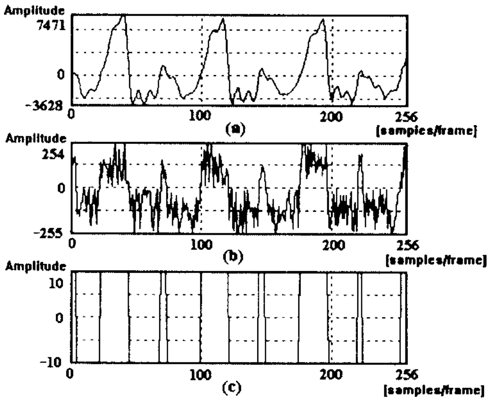


Figure 3.1 An example of extracting the normalized quantization error : (a) speech signal, (b) quantization error, (c) the normalized quantization error

3.2 Updating as new pattern by F_1/F_0 ratio

Traditional speaker verification systems use several reference patterns for combating intra-speaker variations. For example, in case that reference pattern's number is N, we compare these reference patterns with the test pattern on N times. It results in heavy computational load. Also in this case, we should reconstruct the reference pattern as time passes.

The probability of false acceptance error increases, the provided customers use the reference patterns that is made in too long time. It is due to the large allowable error of speech variation as time goes on. On the contrary, the probability of false rejection error increases, the provided customers use the reference patterns that is made in too short time. Both of them cause the increment in the recognition error rate.

In this system, we propose the method of making the reference pattern over a six-month period for the decrease in the recognition error rate. The method of constructing the reference pattern is shown in the fig. 3.2. We obtain the one reference pattern from six patterns using the DTW. That is, we average the corresponding frames by the path information obtained from DTW. The method to obtain the one reference pattern from six speech utterances is various from the viewpoint of computational methods. This paper computes the reference pattern by the fig. 3.2 for the weighting to the latest speech pattern. Also, in order to update the latest user speech pattern fig.3.3 shows this process. First, it performs endpoint detection for the input speech, and extracts quantization error signal in this period. Then F_1/F_0 ratio is obtained from the extracted quantization error signal. So the extracted F_1/F_0 ratio is summed and then divided by the number of N. For updating new reference pattern, the first threshold makes through this process. This method keeps the high recognition rate compared with other systems whose performances degrade rapidly as time goes on. This method is also easy to update rapidly as time pattern, the latest speech pattern reflects the change of speaker's voice.

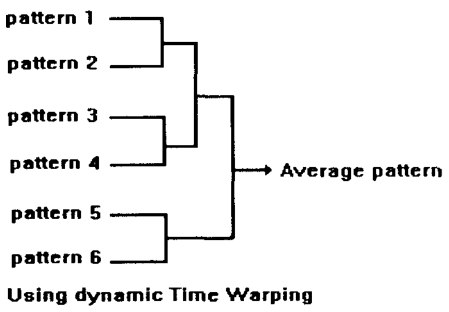


Figure 3.2 Construction of the average reference pattern.

So the latest speech pattern is weighted in comparison with the already established pattern

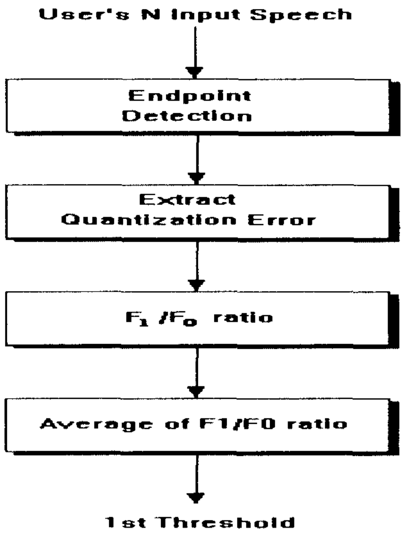


Figure 3.3 The updating process for the latest user speech pattern.

IV. Weighted Cepstrum by F-ratio

We propose the weighted cepstrum to maximize the inter-speaker discriminability. In this paper, we find the effective cepstral coefficients for each speaker and then weight the cepstral coefficient. This paper uses the F-ratio value as the cepstral coefficients as the weighting value. Experimental results show that weighted cepstrum is good parameter for speaker recognition[5].

The F-ratio value is often used as the effectiveness measurement of feature parameters. It is determined as follows: the inter-speaker variance is divided by intra-speaker variance. F-ratio values are defined as [6].

$$F-ratio = \frac{variance\ of\ speaker\ means}{mean\ intraspeaker\ variance} \qquad (4.1)$$

The numerator is large when values for the speaker averaged feature are widely spread for different speakers, and the denominator is small when feature values in the utterances of the same speaker vary small.

The feature of parameter is good when the intra-speaker variation is small and the inter-speaker variation is large.

In this paper, we compute the F-ratio value on each cepstral coefficient to measure the effectiveness of each cepstral coefficient in the inter-speaker's distinction. Our F-ratio value is computed by equation (4.2) according to the above definition. The equation (4.2) is used for the general F-ratio value and the equation (4.3) is used for each speaker's F-ratio value. The result of equation (4.3) is used as weighting function W(i).

$$F-ratio = \frac{Var(E(C_{ij})_j)_{whole\ speakers}}{E(Var(C_{ij})_j)_{whole\ speakers}}$$

$$i = 1, \cdots, I \quad j = 1, \ldots, J \qquad (4.2)$$

$$W_j(i) = \frac{Var(E(C_{ij})_j)_{whole\ speaker}}{E(Var(C_i))_{each\ speaker}}$$

$$i = 1, \cdots, I \quad j = 1, \ldots, J \qquad (4.3)$$

where I is the order of cepstral coefficients and J is the number of speakers registered. If the distribution of F-ratio values of each cepstrum order is equal or has a small variation, the speaker information is distributed in the whole cepstrum order. On the contrary, if the F-ratio values of the cepstrum order have a large variation, it shows that the specific cepstrum order has more speaker information. Figure 4.1 shows the general F-ratio values by equation (4.1). We can see the large variation of F-ratio values in each order of cepstrral coefficient in this figure. It says that some specific cepstral coefficients are more effective on speaker recognition. The results of F-ratio values on each speaker are shown in figure 4.2 by equation (4.2). The example system has 4 customers. We find the different F-ratio distributions for all the speakers in figure 4.2. This fact explains that the above weighting function is appropriate to obtain better discriminability between speakers.

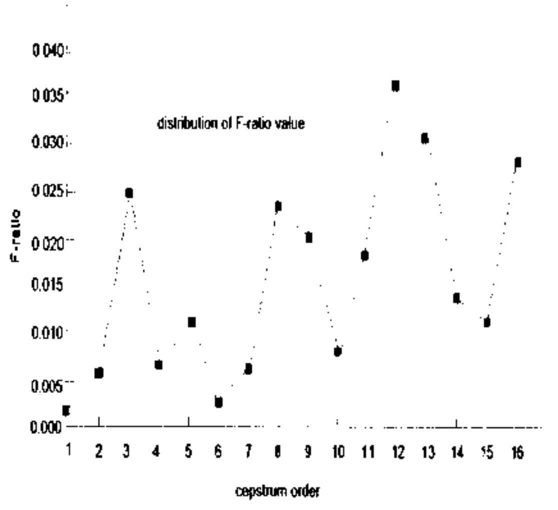
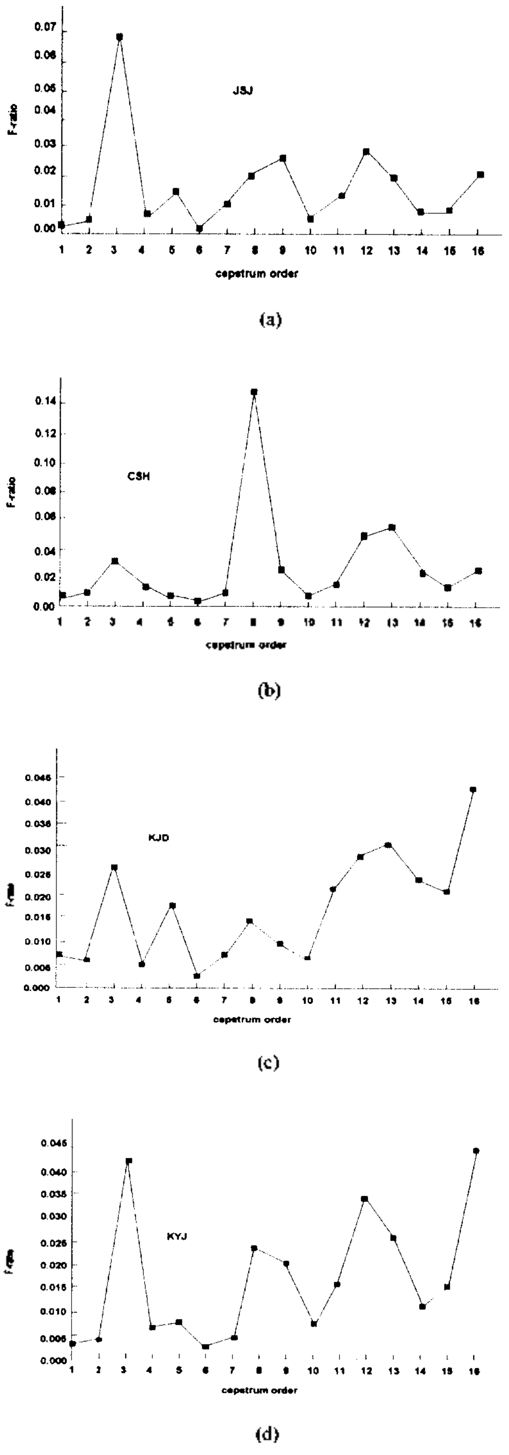


Figure 4.1 The distribution of F-ratio values of each cepstrum order.



(a)

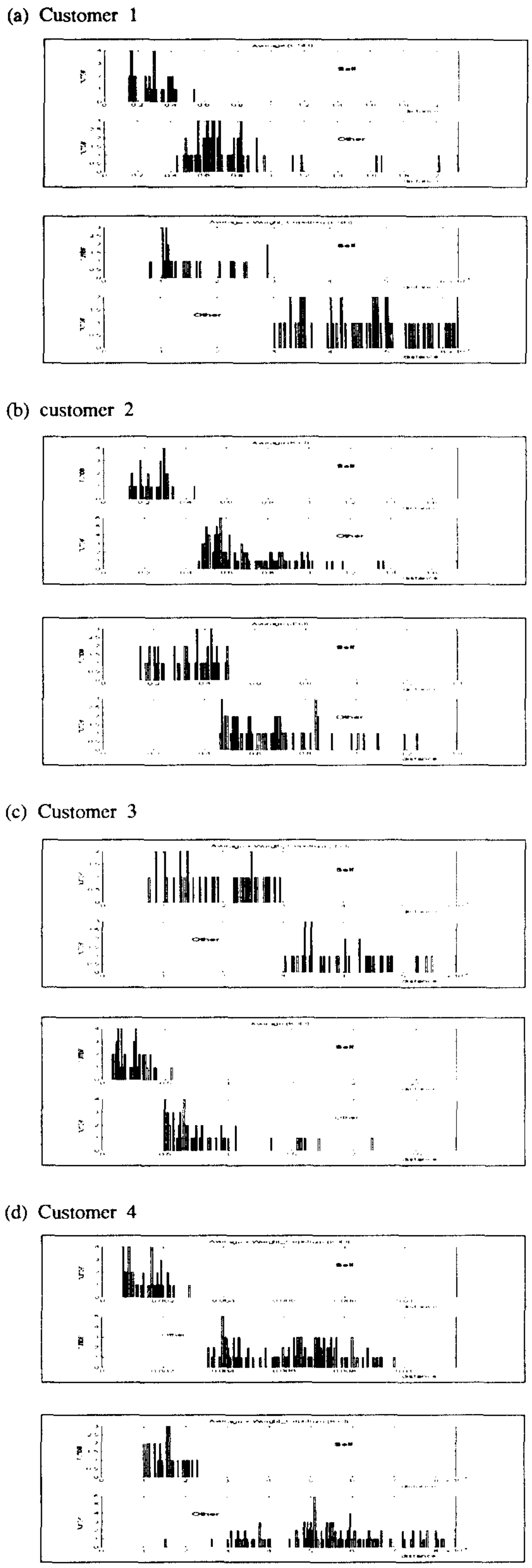
(b)

(c)

(d)

Figure 4.2 The distribution of each speaker's F-ratio values to the cepstrum order.

The result of speaker's verification is determined by comparison with the distance between the input speech pattern and the reference one and the threshold value. That is, speaker verification requires the comparison of the test pattern against one reference pattern and involves a binary decision whether the test speech matches the



(a) Customer 1

(b) customer 2

(c) Customer 3

(d) Customer 4

Figure 4.3 Establishment of the threshold value for each customer.

template of the claimed speaker. The 2 types of errors are possible, the false rejection of customer and the false acceptance of imposter. We present the probability of false rejection as P(N|s) and the probability of false acceptance as P(S|n). Generally, the threshold is selected by the ratio of both errors. In case of too high threshold value, the error rate of false acceptance decreases but the error rate of false rejection increases.

On the contrary, in case of too low threshold value, the error rate of false acceptance increases but the error rate of false rejection decreases. Consequently, it is possible to select the lower threshold value or the higher threshold value according to the application. Mostly, the threshold value is determined on the middle of P(S|n) and P(N|s).

In this paper, we determine the threshold value according to the both error rates having the same value. Figure 4.3 shows the fixed threshold value for each speaker. We can see that the weighted cepstrum using F-ratio value is very effective parameter in Figure 4.3. The figure 4.3 a) of customer 1 shows the distance values which are computed by using the general mel-scaled cepstra. The figure of b) shows the distance values which are computed by using the weighted cepstrum proposed in this paper. If we used the weighted cepstrum, the obvious value which distinguish customer from imposters appears in the results. We set up this value for the threshold.

V. The Overall Configuration of Individual Verification System for WINDOW95

We implement the individual verification system for WINDOW95 based on pattern matching technique.

The Figure 5.1 shows the configuration of our system. The system is composed of two parts : the one part to update the latest pattern and the other part for speaker recognition processing. If the new pattern is less than the first threshold, this latest pattern updates average pattern using as reference pattern.

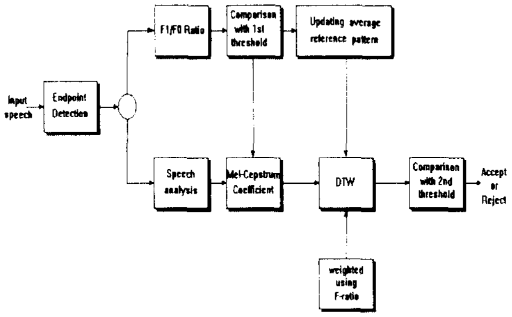


Figure 5.1 Overall individual verification system

VI. Experimental Results and Performance Tests

In this section, we test the performance of speaker verification system and analyze the experimental results. The section 6.1 explains the speech DB(database) and experimental environments. In section 6.2, we compare the proposed system with the traditional system.

6.1 DB And Experimental Environment

The speech DB consists of 10 speakers(male, female) in the common laboratory environment. The customers are 4 persons(male:2, female:2) and the imposters are 10 persons(male:5, female:5). All of the customers uttered their name over a six-month period. Also, all of the imposters uttered the customer's name over a six-month period. We choose the test pattern randomly from speech DB.

The specification of the recording environment and the speech data are as follows.

- ▶ The environment of recording and testing : The common laboratory environment
- ▶ A/D conversion : 8kHz sampling, 16 bit linear PCM, Use the ELF DSP Board of ASPI[7]
- ▶ The contents of speech DB : 3 syllables name
- ▶ The number of total test utterances : 372

In table 6.1, we show the construction of the reference pattern for each speaker. We selected one or more reference patterns randomly out of the data over six months. Six patterns are selected for each speaker. We make the one average reference pattern from these. The 'm' of customer name's parentheses means the male, the 'f' means the female. Table 6.2 shows the distribution of test patterns.

6.2 The Performance Tests of the Speaker Verification System

This section shows the results obtained from the average reference pattern updated as the latest pattern by

Table 6.1 Construction of the reference pattern

Customer	Month & No. of speech data for Average reference pattern					
	1st	2nd	3rd	4th	5th	6th
CSH(m)		2			2	2
JSJ (f)			2	1	1	2
KJD(m)	2	2				2
KYJ(f)		2			2	2

Table 6.2 Test patterns

a. Customer : CSH(82 data)

Imposter	Month & No. of data					
	1st	2nd	3rd	4th	5th	6th
LYK(m)	3		3	4	3	
LCK(m)		3			6	
SYH(m)		2		3		4
KJD(m)	2		3		5	
JSJ (f)		6	3			5
LMS1(f)		2		3		5
LMS2(f)	3			5		
JHS(f)	5		4			

b. Customer : JSJ(85 data)

Imposter	Month & No. of data					
	1st	2nd	3rd	4th	5th	6th
LYJ(m)	3		3			5
LCK(m)		3			5	
SYH(m)		2	4			
KJD(m)	3			4		5
CSH(m)	5		3		3	
LMS1(f)		3		4		5
LMS2(f)	4	4				
KYJ(f)	3		5		2	
JHS(f)	4	3				

c. Customer : KJD(97 data)

Imposter	Month & No. of data					
	1st	2nd	3rd	4th	5th	6th
LYK(m)		4			3	3
LKC(m)	3			5		s
SYH(m)	2	5	3			
CSH(m)	6		5		2	
JSJ(f)		6		5	6	3
LMS1(f)	3		2	3		
LMS2(f)	6	3				
KYJ(f)			5	4		4
JHS(f)	6					

d. Customer : KYJ(108 data)

Imposter	Month & No. of data					
	1st	2nd	3rd	4th	5th	6th
LYK(m)		4		5	2	2
LKC(m)		4			3	4
SYH(m)	3	3	5			
KJD(m)	2			5	3	
CSH(m)	6	5			3	
LMS1(f)	3		5	4		
LMS2(f)	5		5			
JSJ(f)			4		3	5
JHS(f)	5	9				

F_1/F_0 ratio, weighted cepstrum and both methods. And we compared these results with the conventional method.

1) The recognition rate of the speaker verification system using F_1/F_0 ratio

The results of the average reference pattern using F_1/F_0 is shown in figure 6.1. This figure shows the comparison results with the traditional system which has the N reference patterns. We use the test patterns in table 6.2.

In case of using the one average reference pattern updated as the latest pattern, the recognition rate improved somewhat more than that of the traditional system. And the advantage is the low computational load compared with the calculation of N patterns.

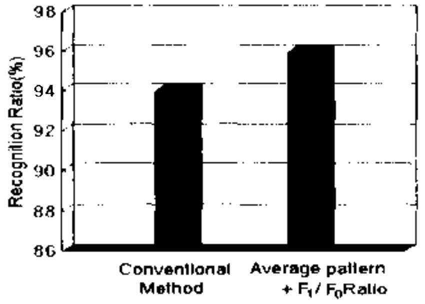


Figure 6.1 The result of use the one average pattern by F_1/F_0 ratio.

2) The results of the weighted cepstrum

We perform an experiment with the system of weighted cepstrum by F-ratio value for each speaker. The traditional system's recognition rate is about 94%. We improve the recognition rate by weighting cepstrum, which results in 99.5%. Our system keeps the high recognition rate compared with the other systems whose performances degrade rapidly as time goes on.

Figure 6.2 shows the experimental results of the weighted cepstrum.

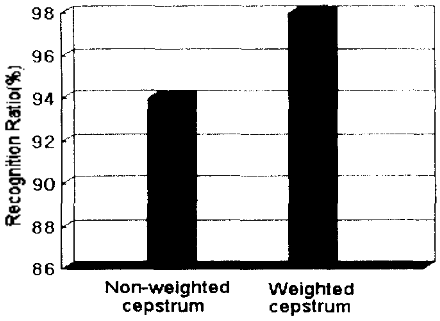


Figure 6.2 The result of the weighted cepstrum

3) The recognition rate of the speaker verification system using the above two cases

We solve the heavy computational load on the traditional system by using the one updating average reference pattern. Also the decrease problem of recognition rate as time goes on is solved by using the weighted cepstrum. Figure 6.3 shows the experimental results of using the traditional system, the average reference pattern, the weighted cepstrum and the both methods. We compare the 4 methods with each other in Figure 6.3. We confirm that the high recognition rate is obtained using the average reference pattern and the weighted cepstrum.

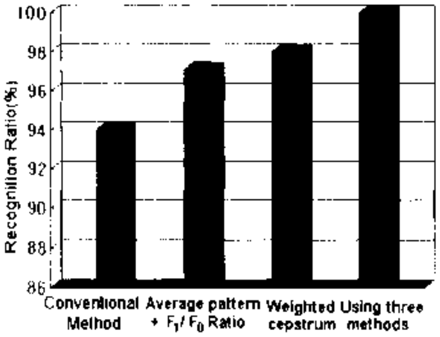


Figure 6.3 The result of the average pattern by F_1/F_0 ratio and the weighted cepstrum.

VII. Conclusion

This paper improves the traditional text-dependent speaker verification system. The conventional system has the problem of heavy computational load and the performance degrade as time goes on. First, we make the average reference pattern from the several reference patterns by DTW. This updates the latest pattern. In this case, we can update without trouble as the latest speech data. Second, we improve more the recognition rate by weighting cepstra coefficients. The weighted values are obtained by using F-ratio values. In case of using the suggested two methods, the system's recognition rate is improved by about 5%-6%.

Conclusively we implement the individual verification system for WINDOW95 according to the method suggested in this paper.

The further work includes the construction of stand-alone system and the robustness on any environment, for example, uncooperating user.

References

1. Hwang-Soo Lee, "Speaker Recognition Technique," *Proc. of*

the Speech Comm. & Signal Processing Workshop, pp. 42-46, 1995.

2. H.Sakoe & S.Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. ASSP*, Vol.26, pp. 43-49, 1978.
3. S.Furui & A.E.Rosenberg, "Experimental Studies in a New Automatic Speaker Verification System Using Telephone Speech," *Proc. of ICASSP*, pp. 1060-1062, 1980.
4. H.Ney & R.Gierloff, "Speaker Recognition Using a Feature Weighting Technique," *Proc. of ICASSP*, pp. 1645-1648, 1982.
5. J.Naik & G.Doddington, "High Performance Speaker Verification Using Principal Spectral Components," *Proc. of ICASSP*, pp. 881-884, 1986.
6. JongSoon Jung et. al., "A study on the performance improvement of speaker recognition using average pattern and weighted cepstrum," *Proc. of the Speech Comm. & Signal Processing Workshop*, pp. 179-183, 1995.
7. D. O'Shaughnessy, "Speaker Recognition," *IEEE ASSP Magazine*, Oct. pp. 4-17, 1986.
8. *Elf DSP Platform Instruction Manual*, Atlanta Signal Processors Inc., 1993.

MyungJin Bae was born in Kyungsangbukdo on May. 20 in 1956. He received the B.S. degree in electronics engineering from SoongSil University and the M.S degree in electronics engineering from Seoul National University in 1981 and 1983, respectively. From the same university, he received the Ph.D. degree in electronics engineering, too, in 1987. He has been professor for Dept. of Information and Telecommunication of SoongSil University, located at Seoul in Korea, ever since 1992. His research interests include speech coding, synthesis, and speaker recognition.

JongSoon Jung was born in Kyungkido, on feb. 9, 1966. She received the B.S. degree in Electronics Engineering from Seoul Industrial University in 1990, and the M.S. degree in Electronics Engineering from Seoul City University in 1992. Another the M.S. degree in Engineering from Kaist in 1996. And in 1998, she is a professor in Sanggi junior College and enrolled in a ph.D degree of Soongsil University. Her research interests include speaker recognition.

JaeOk Bae was born in Kyungsangbukdo on Aug. 31, 1974. He received the B.S. degree in Information and Telecommunication Engineering from SoongSil University in 1996, and enrolled the M.S. degree in Information and Telecommunication Engineering in Soong-Sil University. His research interests include speech coding, speaker/speech recognition, and speech enhancement.