DTW를 이용한 향상된 문맥 제시형 화자인식

An Enhanced Text-Prompt Speaker Recognition Using DTW

신 유 식*, 서 광 석*, 김 종 교*
(You Shik Shin*, Kwang Seok Seo*, Chong Kyo Kim*)

요 약

본 연구에서는 문맥 종속 또는 문맥 독립형 화자 인식에서의 단점을 개선하는 방법으로 문맥 제시형 화자 인식 실험을 수행하였다. 화자 인식 알고리즘으로는 개선된 Dynamic Time Warping(DTW)을 사용하였고 실시간 처리를 위하여 전체 계산량을 증가시키지 않는 아주 간단한 끝점검출알고리즘을 사용하였으며, 여러 가지 다양한 특징파라미터를 이용하여 인식실험을 행한 결과 weighted cepstrum을 이용했을 때 가장 좋은 인식성 능을 얻을 수 있었다. 실험결과 세 개의 단어를 제시하였을 경우 화자식별오류는 0.02%를 보였고, 화자확인은 문턱값을 적절히 정했을 때 사용자 거부율 1.89%, 사칭자 허용률 0.77%, 총 확인 오류 0.97%를 보였다.

ABSTRACT

This paper presents the text-prompt method to overcome the weakness of text-dependent and text-independent speaker recognition. Enhanced dynamic time warping for speaker recognition algorithm is applied. For the real-time processing, we use a simple algorithm for end-point detection without increasing computational complexity. The test shows that the weighted-cepstrum is most proper for speaker recognition among various speech parameters. As the experimental results of the proposed algorithm for three prompt words, the speaker identification error rate is 0.02%, and when the threshold is set properly, false rejection rate is 1.89%, false acceptance rate is 0.77% and verification total error rate is 0.97% for speaker verification.

I. 서 론

현대 정보화 사회로의 발달로 인한 소비자의 욕구에 맞추어 다양한 서비스가 개발되고 있다. 이러한 서비스에는 은행의 자동 현금지급서비스, 전화 쇼핑서비스, 정보검색서비스 등이 널리 이용되고 있으며, 본인 또는 허가를 얻은 사람만이 접근해서 서비스를 이용하도록 되어 있다. 그러나 만약에 허가를 얻지 않은 사람이 접근해서 사용을 한다면 심각한 사회문제를 발생시킬 수 있다. 따라서 개인의 신분확인은 보안을 필요로 하는 곳에 필수적이다. 과거에는 사람의 신원을 확인하는데 신분증, ID카드, 도장, 서명 등으로 신원을 확인하였으나 이러한 경우 분실이나 복사 또는 사본을 만들어서 사용하는 문제점이 발생하였다.

따라서 인간의 특성을 이용하여 신원을 확인하고자 하는 연구가 진행되고 있다. 이러한 연구 분야 중에서 인간의 음성을 이용하여 신원을 확인하는 연구가 바로 화자인식분야이다. 이는 화자의 음성신호에서 발생되는 개인의 특성들을 추출하여 화자를 인식한다. 음성 신호로부터 말하는 사람이 누구인지를 판단하는 화자인식(speaker recognition)기술은 task의 성격에 따라 화자식별(speaker identification)과 화자확인(speaker verifica- tion)으로 나눌 수 있다. 여기서 화자 식별이란 등록된 화자들 중 발화자가 누구인가를 알아내는 것이고, 화자확인은 특정인이라고 자칭하는 인식 대상이 본인인지 여부를 알아내는 과정을 의미한다.

문맥 종속형 화자 인식은 제한된 문장이나 단어를 발성하여 화자를 인식하는 방법이며, 문맥 독립형 화자 인식은 임의의 대화 문장이나 대화를 화자가 발성하여 발성한 화자를 인식하는 방법이다. 이러한 화자인식 방법을 이용한 시스템에서는 비등록자가 등록자의 목소리를 흉

*전북대학교 전자공학과
접수일자 : 1998년 10월 27일

내내거나 등록자의 음성을 녹음한 후에 비등록자가 이를 이용하여 등록자인 것 처럼 모방할 수 있는 단점이 있다. 본 논문에서는 이러한 단점들을 해결할 수 있는 문맥 제시형 화자 인식을 수행하였다. 문맥 제시형 화자 인식은 등록된 단어나 문장을 랜덤하게 제시하여 문맥종속형이나 문맥독립형의 문제점을 보완할 수 있다.[1]

Ⅱ. 음성 구간 검출 및 특징파라미터

음성 구간의 검출은 화자인식 시스템에 큰 영향을 줄 수 있기 때문에 정확한 검출이 요구될 뿐만 아니라 실시간 시스템을 사용하기 위해서 전체 계산량을 크게 증가시키지 않는 효율적인 방법이어야 한다. 따라서 본 논문에서는 기존에 제시되었던 에너지, 영교차율, 피치 등을 사용하지 않고 프레임 절대값의 합을 이용하여 구간을 검출하였으며 구하는 식은 다음과 같다.

$$E=\sum_{n=0}^{N-1}|x(n)| \tag{2.1}$$

음성구간검출의 전체적인 과정은 다음과 같다. 먼저 입력된 음성데이터는 8kHz로 샘플링되고 16bit로 양자화된다. 이러한 음성 데이터를 프레임단위로 처리를 하기 위하여 16ms(128 sample)로 프레임의 크기를 결정하였다. 문턱값은 배경잡음에 의해 구해지게 되는데 일단 시스템이 시작되면 초기 8프레임(128ms)동안에서는 음성이 입력되지 않는다는 가정하에 한 프레임에서의 평균 절대값의 합을 구하고 이 평균 절대값의 합에 정수배인 2배를 한 값을 문턱값으로 실험을 통하여 결정하였다. 문턱값이 정해지면 다음에 입력되는 프레임이 문턱값보다 크면 잠재적으로 음성구간이 시작되었다고 가정을 한다. 이 때, 문턱값보다 큰 구간이 8프레임이상 지속되면 음성구간이라고 판정을 하고 처음으로 문턱값을 넘는 프레임보다 2프레임 앞선 프레임부터의 데이터를 출력한다. 또한, 음성구간내에서의 묵음구간을 구별하기 위해서는 음성구간이 지속되다가 문턱값보다 작은 프레임이 15프레임(240ms) 이하이면 음성구간내의 묵음구간으로 판정을 하고 15프레임 이상이면 음성구간이 끝난 것으로 판정을 하였다. 이때, 음성구간이 끝난 것으로 판정되면 문턱값보다 작은 프레임이 처음으로 시작되는 점을 기준으로 2프레임 후까지의 데이터를 출력한다.

본 연구에서 구현한 알고리즘의 흐름도는 그림 2.1과 같다.

이렇게 구한 음성을 이용하여 LPC 켑스트럼(LPC cepstrum), CMS(cepstral mean subtraction), 가중 켑스트럼(weighted cepstrum), 멜-켑스트럼(mel-cepstrum)을 구하였다.[2]

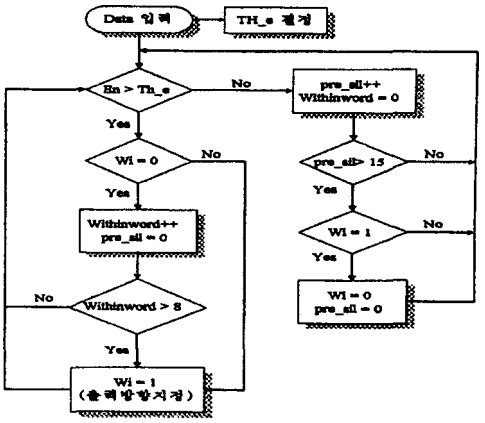

그림 2.1. 실시간 음성 구간 검출 알고리즘
Fig. 2.1. Real time end point detection algorithm.

Ⅲ. 인식 시스템

3.1. DTW 알고리즘

DTW 알고리즘은 시험 패턴과 기준 패턴을 시간축상에 최적이 되도록 배열한 후, 이 최적 변형 경로를 통한 최적 거리를 얻어내는 방법이다.[3]

DTW 알고리즘은 2차원 $d(m,n)$ 평면상에서 적절한 제한조건을 만족하는 $d(1,1)$에서 $d(M,N)$까지의 최적 경로를 구하는 방법이다. 임의의 점 $(m(j), n(j))$까지 축적된 거리를 다음과 같이 정의할 수 있다.

$$C_i(m,n)=\sum_{j=1}^{J} d_i(m(j),n(j))\cdot w(j) \tag{3.1}$$

여기에서 $(m(j),\ n(j))$, $j=1,2,\cdots, J$는 주어진 경로이고, $w(j)$는 각 경로에 따른 가중치이다. 그러면 최적 경로는 $d(M,\ N)$에서의 축적된 거리 $C_i(M,\ N)$을 최소화하는 경로로

$$D_i(T,R_i)=\min_{path} C_i(M,N) \tag{3.2}$$

로 표시된다. 최적 경로를 구하는 과정에서 제한 조건들은 다음과 같은 목적으로 주어진다. 즉, 부분적으로는 경로 기울기의 범위를 제한하며, 전체적으로는 경로의 허용 영역을 제한한다.

3.2. 개선된 DTW알고리즘

본 논문에서는 DTW의 계산량 감소를 위하여 일반적으로 사용되는 DTW알고리즘의 전역제한조건을 적용하여 계산량을 1차적으로 감소시켰다. 이 때 계산량감소비율은 가로축과 세로축이 각각 N, M일 경우에 다음과 같다. 여기에서 k는 최대 경사기울기를 나타낸다

$$R=\frac{\left(k-\frac{N}{M}\right)\left(k-\frac{M}{N}\right)}{k^2-1} \tag{3.3}$$

전역제한조건 중 최대 경사기울기를 2로 결정하였을 경우 계산량감소비율은 1/3이 되어 전체계산량이 줄게 되는 결과를 얻게된다.[4]

이 때 DTW알고리즘상에 곱셈, 덧셈, 비교의 연산량은 각각 $R \times N \times M \times 3$, $R \times N \times (M+1) \times 3$, $R \times N \times M$의 계산량이 필요하다.

2차적으로 DTW의 속도의 향상을 위해 최소누적제한조건을 이용하였다.

$$\bar{C}_{ip}(m,n)=\min_{path}\sum_{j=1}^{P} d_i(m(j),n(j))\cdot w(j) \tag{3.4}$$

최소누적제한조건이란 임의의 P점까지의 최소누적경로의 합을 식 (3.4)와 같이 정의하였을 경우 누적거리를 구하게 되는데 여기서 발생되는 최소의 C_{ip}의 값이 D_i보다 커지게 되면 계산을 중단하고 다음 참조패턴으로 넘어가서 계산을 하게 된다. 그러나 이렇게 비교부분을 첨가시켰을 경우 비교연산량은 $R \times N \times M$만큼 증가하게 된다.

그렇기 때문에 이러한 비교연산량을 첨가시켰을 경우와 첨가를 하지 않았을 경우의 전체연산량을 비교해야 한다. 세로축과 가로축의 크기가 모두 N이라고 가정하면 비교연산을 한 후 D_i보다 커진부분의 연산을 하지 않는다. 이 연산을 하지 않는 부분의 계산량의 감소를 구했을 경우에 다음과 같은 연산량의 감소를 각각 나타낼 수 있다.

곱셈, 덧셈 비교의 경우는 각각 $3\times\left(\sum_{i=1}^{N/3} i-\sum_{i=1}^{N/9} 3i\right)$,

$3\times\left(\sum_{i=1}^{N/3} i-\sum_{i=1}^{N/9} 3i\right)+\frac{N}{3}$, $\left(\sum_{i=1}^{N/3} i-\sum_{i=1}^{N/9} 3i\right)$ 씩 감소가 된다.

여기에서 i=1은 끝점을 나타낸다.

즉 증가된 비교 연산량보다 감소된 전체 연산량이 더 큰 값을 가졌을 때 알고리즘을 이용하였을 경우 이득이 나타난다. 본 실험에서는 3음절로 구성된 단어일 경우 N과 M의 값이 보통 60이 되기 때문에 계산식에 의하여 끝점으로부터 i=3인 경우 연산이 수행되지 않는다면 전체 연산량의 감소가 비교연산량의 증가보다 훨씬 작은 값을 갖게 된다. 실험을 통하여 알고리즘의 성능을 비교해 본 결과 비교연산을 첨가한 알고리즘을 적용하였을 경우 보통 끝점으로부터 i=15~35로 약 20%~70%의 연산량 감소로 인한 속도 향상을 얻을 수 있었다. 전체적인 흐름도는 다음과 같다.

3.3. 문턱값 결정원리

모든 입력 음성과 기준 패턴간의 거리를 구하여, 사용자를 주장하는 사람을 받아들일 것인지 거부할 것인지를 결정하는 문턱 거리값과 비교한다. 사용자인 화자의 거부, 즉 사용자 거부율 $P(Nls)$과 사칭자 허용율 $P(Sln)$ 등 두 가지 가능한 에러가 있다. 일반적으로 문턱치는 이들 에러에 대한 비중에 따라 선택한다. 만약 문턱 거리값이 너무 작게 설정되면 사용자 거부율은 높아지나 사칭자 허용율은 낮아진다. 반대로, 이 문턱치를 크게 설정하면 사용자 거부율은 낮아지나 사칭자 허용율은 높아진다. 따라서 필요에 따라 $P(Sln)$을 낮추어 사용할 수도 있고, $P(Nls)$을 낮추어 사용할 수도 있다. 일반적인 경우 $P(Sln)$과 $P(Nls)$에 중간에서 문턱값을 결정한다. 그러나 이런 방법은 등록자 이외에 비등록자의 음성데이터가 많이 필요하며, 또한 실제 시스템 구현에 있어서 새로운 화자의 등록시마다 문턱값이 변하게 되므로 결정을 하는데 어려움이 따르게 된다.

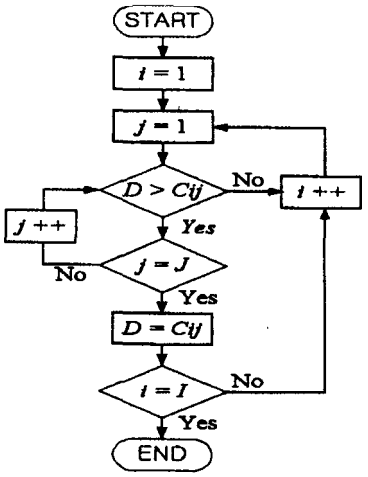

그림 3.1. 개선된 DTW알고리즘
Fig. 3.1. Improved DTW algorithm.

따라서 본 논문에서는 등록된 화자만의 데이터를 이용하여 문턱값설정을 위해 세가지 방법을 이용하여 결정하였다.

첫 번째 방법으로 평균 + 1.5σ, 두 번째는 평균 + 2σ, 세 번째는 평균 + 2.5σ로 결정하고 실험을 수행하였다.

3.4. 문맥제시형 화자인식 알고리즘

화자인식 방법을 이용한 시스템에서는 비등록자가 등록자의 목소리를 흉내내거나 등록자의 음성을 녹음한 후에 비등록자가 이를 이용하여 등록자인 것처럼 모방하였을 경우 이를 제지할 마땅한 방법이 존재하지 않는다. 따라서 본 논문에서는 이러한 단점들을 어느정도 해결할 수 있는 문맥 제시형 화자 인식 방법을 제시하였으며 이를 이용하여 화자인식을 수행하였다. 문맥 제시형 화자 인식은 정해진 화자인식용 단어를 랜덤하게 제시하는 방법이다.

본 논문에서는 close set에 대하여 화자식별을, open set

에 대하여 화자확인 실험을 하였다. 먼저 오 인식률을 감소시키기 위하여 하나의 단어를 제시하는 방법과 하나가 아닌 세 개의 단어를 랜덤하게 제시하는 방법을 사용하였다. 이때 False Acceptance(FA)와 False Rejection (FR)율의 상승을 방지하기 위하여 세 개의 단어중 2개 이상 같은 등록자로 판명이 날 경우에 등록자로 결정하였다.

Ⅳ. 실험 및 결과

본 논문에서는 화자인식 실험을 위해 등록자 6명에 대하여 10개의 단어를 20회 발성하여 앞의 4회 데이터는 참조패턴으로 나중 16회 데이터는 close set의 시험패턴으로 사용하였다. open set실험을 위하여 비등록자 14명에 대하여 10개단어를 2회씩 발성하였다. 녹음환경은 주변잡음이 존재하는 일반적인 실험실이며, 10개의 단어는 8kHz 16bit로 녹음한 후 4가지의 음성 특징파라미터를 추출하였으며 특징파라미터의 차수는 고정적으로 모두 15차로 결정하였다. 화자인식을 위한 구성도는 그림 4.1과 같다.

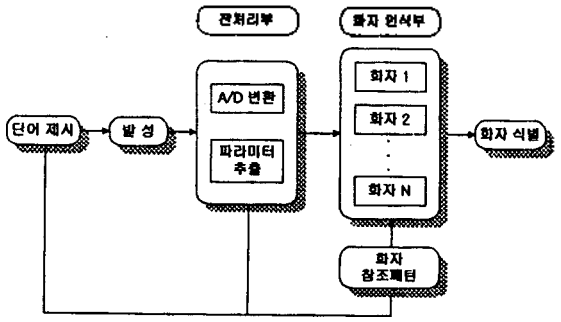

그림 4.1. 화자인식 시스템
Fig. 4.1. Speaker recognition system.

단어는 랜덤하게 시스템에서 제시되며 화자는 제시된 단어를 발성하게 된다. 정확한 화자 인식을 위해서는 화자 개개인의 특성을 잘 반영할 수 있는 특징파라미터의 선택과 문맥종속형 화자인식을 대상으로 할 경우 어휘선택이 중요하다.

화자인식을 위한 표준어휘세트를 선정하기 위해서 음소별 화자식별을 실험하여 각 음소의 화자특성 반영정도를 비교하고, 실제 단어 내에서의 영향을 고려하기 위해 단어별 화자식별 실험을 통한 단어들 중에서 표 4.1과 같이 10개를 선정하였다.[6]

표 4.1 표준 단어 세트
Table 4.1. Standard word set.

1	공예품	6	무녕왕릉
2	나그네	7	사과나무
3	다람쥐	8	아주머니
4	요사이	9	정육면체
5	발자취	10	훈민정음

4.1 실험 1 (화자식별 : close set)

전체의 실험은 먼저 4가지의 특징파라미터를 이용하여 close set일 경우에 대하여 화자식별 실험을 하였다. 다음 표 4.2는 그 결과를 나타낸 것이다.

표 4.2. 여러 가지 파라미터의 인식률
Table 4.2. The recognition rate for each parameter.

	Mel-cepstrum	Cepstrum	CMS	Weighted cepstrum
1	88.54	88.54	88.54	100
2	86.46	85.42	86.46	96.88
3	85.42	86.46	89.58	92.71
4	90.62	90.62	96.88	100
5	83.33	85.42	87.50	93.75
6	86.46	84.38	91.67	100
7	87.50	87.50	90.62	96.88
8	84.38	90.62	85.42	100
9	84.38	86.46	88.54	100
10	89.58	86.46	90.62	97.92
T	86.67	87.19	89.58	97.81

각 단어에 대하여 화자식별 실험을 행한 결과, weighted cepstrum을 이용하였을 경우 2.19%의 오인식결과를 얻을 수 있었으며 이를 단어별 인식률과 전체 인식률로 나타내면 각각 그림 4.2 및 그림 4.3과 같다. 한 개의 단어를 제시하였을 경우와 세 개의 단어를 랜덤하게 제시한 문맥 제시형 화자인식 실험을 하였을 경우 두 개 이상이 같은 화자에 대응했을 때 그 화자로 식별한 실험결과를 그림 4.4에 도시하였으며, 오인식률 0.02%의 결과를 얻어 인식률이 2.17% 향상됨을 알 수 있다.

결과적으로 가중 캡스트럼일 경우에 가장 높은 인식률을 보였다.

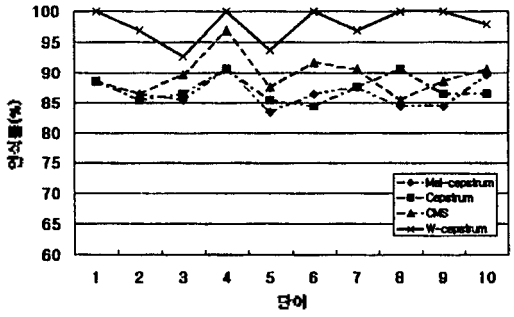

그림 4.2. 각 단어당 인식률
Fig. 4.2. The recognition rate for each word.

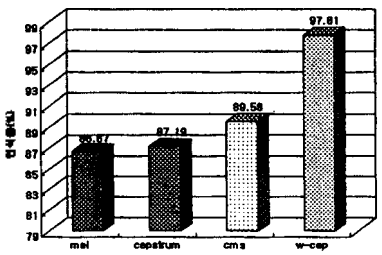

그림 4.3. 전체 인식률
Fig. 4.3. Total recognition rate.

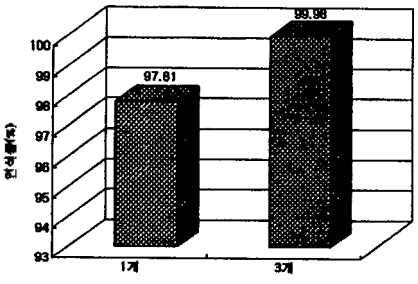

그림 4.4. 문맥제시단어 개수에 대한 인식률
Fig. 4.4. Recognition rate for the number of prompted word.

4.2 실험 2 (화자확인 : open set)

화자확인 실험시에는 화자식별 실험에서 가중 캡스트럼이 가장 높은 인식률을 보였기 때문에 이를 음성 특징 파라미터로 결정하였다.

한개의 단어를 랜덤하게 제시하여 화자확인 실험을 하였을 경우 문턱값을 평균에 표준편차의 1.5배, 2배, 2.5배를 더하여 실험을 하였으며 이 때 각각 FA(false acceptance)는 1.61%, 3.63%, 6.19%로 FR (false rejec- tion)은 14.17%, 6.33%, 3.83%를 얻을 수 있었다. 세 개의 단어를 랜덤하게 제시하였을 경우 두개 이상이 같은 화자에 대응하면 그 화자로 인식하며 FA는 각각 0.20%, 0.77%, 1.80%, FR은 4.53%, 1.89%, 0.97%의 결과를 얻을 수 있었다.

이러한 결과로는 정확한 문턱값을 선정하기는 어렵기 때문에 새로운 문턱값 결정 단위를 선정하여 이를 VM (verification mean error), VT(verification total error)로 하여 다음과 같이 구하였다.

$$VM = \frac{FA + FR}{2} \times 100(\%)$$

$$VT = \frac{AT + RT}{TA + TC} \times 100(\%)$$

여기서,

AT : falsely Accepted Trial

RT : falsely Rejected Trial

TA : Total number of Autotrial

TC : Total number of Crosstrial

이며 문턱값이 표준편차의 2배인 경우에 VT, VM값이 가장 낮았다.

표 4.3은 한 개의 단어를 제시하였을 경우, 표 4.4는 단어 3개를 제시하였을 경우의 FA, FR, VM, VT를 나타낸 것이며, 그림 4.5, 4.6 및 4.7은 각각 세가지의 문턱값에 대한 오차율을 나타낸 것이다.

표 4.3. 단어 1개 제시형 인식률
Table 4.3. Recognition rate for prompted 1-word.

문턱값	FR	FA	VM	VT
1.5배	14.17	1.61	7.89	4.91
2배	6.33	3.63	4.98	4.17
2.5배	3.83	6.19	5.01	5.57

표 4.4. 단어 3개 제시형 인식률
Table 4.4. Recognition rate for prompted 3-word.

문턱값	FR	FA	VM	VT
1.5배	4.53	0.20	2.37	0.97
2배	1.89	0.77	1.33	0.97
2.5배	0.97	1.80	1.39	1.66

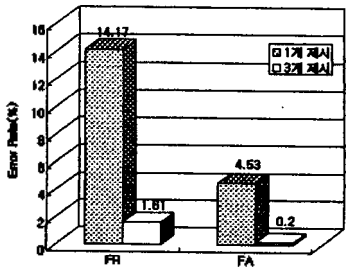

그림 4.5. 1.5σ일 경우 error rate
Fig. 4.5. Error rate of 1.5σ.

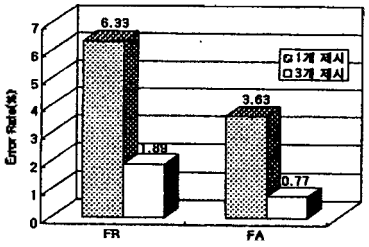

그림 4.6. 2.0σ일 경우 error rate
Fig. 4.6. Error rate of 2.0σ.

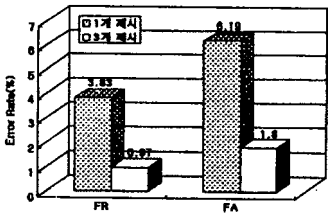

그림 4.7. 2.5σ일 경우 error rate
Fig. 4.7. Error rate of 2.5σ.

V. 결 론

본 논문에서는 문맥을 미리 제시하여 화자가 제시된 단어를 발성하는 형태의 화자인식을 수행하였다. 음성 구간 검출을 실시간으로 처리하기 위해 프레임의 절대값의 합을 이용하여 음성구간을 검출하였으며, 인식 알고리즘으로는 속도가 개선된 DTW를 사용하였다. 인식률을 향상시키기 위해 3개의 단어를 제시하여 2개 이상 일치하는 화자가 있으면 그 화자로 인식하도록 하였다. 또한 화자의 특성을 잘 나타낼 수 있는 특징 파라미터를 선정하기 위해 4가지의 특징 파라미터에 대해 실험을 하였다. 또한, 화자확인시 필요한 문턱값 결정은 등록된 화자의 훈련 데이터를 이용하여 결정하는 방법을 제시하였다. 실험결과

화자식별은 특징 파라미터로 가중 캡스트럼을 이용하였을 때 99.98%로 가장 높은 인식률을 얻었고, 화자확인은 문턱값을 적절히 정했을 때 사용자 거부율 1.89%, 사칭자 허용률 0.77%, 총 확인 오류 0.97%를 보였다.

앞으로 더 많은 연구를 위해서는 한국어의 표준 데이터베이스가 구축이 되어야 한다고 생각한다.

참 고 문 헌

1. Chi Wei Chi, "An HMM Approach to Text- Prompted Speaker Verification," Proc. of the IEEE, vol. 2. pp. 673-676. 1996.
2. Furui & Sondhi, Advances in Speech Signal Processing, Dekker, 1991.
3. L. R. Rabiner & R. W. Schafer, Digital Processing of Speech Signal, Prentice-Hall, Englewood Cliffs, N. J., U.S.A., 1978.
4. L. R. Rabiner & Biing-Hwang Juang, Fundamentals of Speech Recognition, Prentice-Hall, AT&T, U.S.A, 1993.
5. D. A. Reynolds, R. C. Rose, "Robust Text- Independent Speaker Identification Using Gaussian Mixture Speaker Models," IEEE Trans. Speech and Audio Processing, vol. 3, no. 1. pp. 72-83, 1995.
6. 오영환, "음성합성시스템의 평가 및 화자확인시스템을 위한 표준어휘세트의 설계", 한국과학기술원, 1994.

▲신 유 식 (You-Shik Shin)
한국음향학회지 vol. 17 No. 1 참조
현재 : 전북대학교 전자공학과 박사과정

▲서 광 석 (Kwang-Seok Seo)
한국음향학회지 vol. 17 No. 1 참조
현재 : 전북대학교 전자공학과 석사과정

▲김 종 교 (Chong-Kyo Kim)
한국음향학회지 vol.,17 No. 1 참조
현재 : 전북대학교 전기 · 전자 · 제어공학부 교수