VQ와 DTW를 이용한 문장 의존형 화자인식 시스템

정종순, 오세영, 배명진(°)
숭실대학교 정보통신공학과

Text-dependent Speaker Recognition System Using DTW & VQ

JongSoon Jung, SeYoung Oh, MyungJin Bae(°)
Dept. of Telecommunication, Soongsil University
E-mail(°) : mjbae@saint.soongsil.ac.kr

요약문

초기 DTW(Dynamic Time Warping)를 이용한 화자인식 방법은 인식률이 시간이 지남에 따라 저하된다는 단점이 있었다[1][3][4]. 따라서 이를 보완하기 위한 새로운 알고리즘이 많이 제안되었다. 본 논문에서는 DTW방법을 이용한 화자 인식 시스템의 사용자 등록시기에 화자에 대한 불충분한 음성특징을 보충하고 시간이 지남에 따라 발생하는 오인식률의 증가를 줄이기 위해 사용자 등록시 기준패턴의 정규화를 수행하고 시스템 사용시 기준패턴을 변경하는 방법이다. 본 논문에서 사용된 핵심적인 알고리즘은 VQ(Vector Quantization)와 DTW 방법이다.

본 논문의 알고리즘을 이용한 모의 실험 결과 기존의 방법에 비해 3.3% 인식률 향상되어 97.5%의 인식률을 얻을 수 있었다.

Abstract

The speaker recognition method using DTW algorithm has the problem that is reducing the performance of the speaker recognition system as the time variation. So there are many proposed algorithms to solve these problems. This paper proposes the new method to make the reference pattern that is acceptable to intra-speaker variation by reference pattern normalization. And to avoid reducing performance of speaker recognition system, we use the modified reference pattern to recognize the system user. The used methods in this paper are VQ and DTW.

As the result of simulation we can obtain the 97.5% of recognition accuracy rate.

1. 서론

음성을 이용한 통신의 기본적이고 궁극적인 목적은 그 의미의 전달에 있다 즉, 의미라는 것은 음성이 내재하고 있는 가장 큰 정보이며 중요한 정보이다. 그러나 이러한 음성은 음성이 가지고 있는 의미적인 음향학적 특징 외에도 개인의 특징이나 언어의 종류, 화자의 심리적, 육체적, 감정상의 상태에 대한 정보를 포함하고 있다. 이러한 정보들중 화자의 특징을 이용하여 화자의 신원을 파악하는 기술인 화자 인식은 전화망이 설치된 이후부터 그 연구가 활발히 진행되었다. 근래에 들어 컴퓨터를 이용한 자동 화자 인식(ARS:Automatic Speaker Recognition) 기술은 1976년 Atal의 연구를 시작으로 활발해지기 시작했으며, 부분적으로 상용화되기 시작했다.

화자인식에는 사용되는 알고리즘의 종류에 따라
여러 가지로 나눌 수 있다[1][3][4].

먼저 DTW(Dynamic Time Warping)방법은 초창기화자인식 방법으로 비선형 시간 정규화를 갖는 패턴정합 알고리즘이다 HMM(Hidden Makov Model)을 이용한 화자인식은 화자의 음성특징을 확률적으로 모델링하여 확률분포 및 상태 천이 확률을 이용하여 화자를 인식하는 방법이다.

다음으로 VQ(Vector Quantization)을 이용한 화자인식은 화자의 음성특징 패턴을 대표값으로 변환한 뒤 각각의 특징패턴과 대표값과의 거리차이를 이용하여 화자를 인식하는 방법이다. 근래에 들어서는 GMM(Gaussian Mixture Model)이나 신경회로망(Neural Network)방법을 이용한 화자인식 알고리즘의 연구가 활발히 진행되고 있다. 본 논문에서는 DTW를 이용하여 사용자를 인식하는 방법을 사용하였다. 그리고 시간이 지남에 따라 인식 시스템의 성능 저하를 막기 위해 다음과 같은 알고리즘을 사용하였다.

인식시스템의 성능이 시간에 따라 저하되는 것은 사용자의 목소리가 시간이 지남에 따라 변화한다는 문제와 기준패턴 선정시 화자의 정보가 불충분하다는 것에 문제가 있다. 따라서 본 논문에서는 정규화된 기준패턴을 사용하고 기준패턴을 시간이 지남에 따라 변경함으로써 위의 문제를 해결하려한다.

2. 사용자인식 시스템의 구조

일반적인 사용자 인식 처리는 다음과 같이 수행한다. 시스템의 사용자가 발성한 음성으로부터 사용자의 음향적인 특징을 추출하여 기준패턴으로 선정한다. 그림 2.1 은 일반적인 사용자 인식 과정의 구조를 나타낸 것이다.

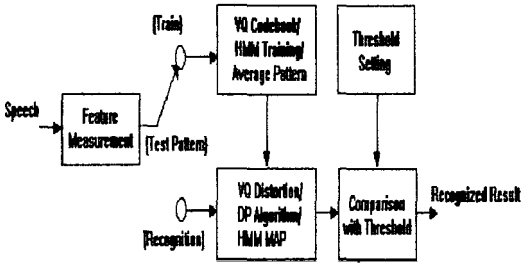

그림 2.1 일반적인 사용자 인식 시스템의 구조
fig 2.1 Generalized speaker recognition system

사용자가 시스템에 접근하기 위해 특정 키워드나 임의의 어휘를 발성하여 특징 파라메터를 추출한 뒤 이를 시험패턴으로 정한다. 이렇게 정해진 시험 패턴과 메모리에 저장 되어있

는 모든 사용자의 기준패턴과 패턴매칭을 수행한다. 시험패턴과 모든 기준패턴과의 패턴매칭이 종료되면 결정논리에 따라 사용자를 확인한다.

먼저 사용자의 음성특징 추출전에 처리해야 할 과정은 음성구간 검출이다. 음성구간 검출은 인식시스템의 성능에 큰 영향을 끼치므로 정확한 음성구간 검출이 요구된다. 본 논문에서는 다음과 같은 처리를 통해 사용자의 음성구간 검출을 수행하였다.

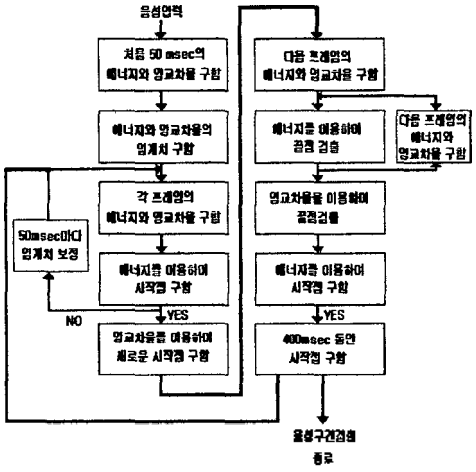

그림 2.2 음성구간검출

fig 2.2 End-point detection

그림 2.2와 같이 에너지와 영교차율(Zero Crossing Rate, ZCR)을 이용하여 음성구간 검출을 수행하였다. 즉, 에너지를 이용하여 개략적인 음성의 시작점과 끝점을 찾고 영교차율을 이용하여 보다 정확한 시작점과 끝점을 찾음으로써 보다 정확한 음성구간검출이 이루어졌다.

다음으로 본 논문에서는 사용자의 음성 특징벡터로 14차 Mel-Cepstrum을 사용하였다. 특징추출 과정은 다음 그림 2.3과 같다[1].

이 과정을 살펴보면 먼저 음성구간 검출 과정을 거친 음성신호를 Hamming Window를 이용하여 단구간으로 나눈다. 이는 음성신호는 단구간에서는 정적이므로 신호를 선형적으로 예측할 수 있기 때문에 음성구간을 단구간으로 나눈다. 음성신호의 선형예측계수(Linear Prediction Coefficients, LPC)는 자기상관함수를 이용하는 Durbin's 알고리즘으로 추출하였고 LPC와 Cepstrum 변환식을 이용하여 14차 LPC-Cepsturm을 추출하였다.

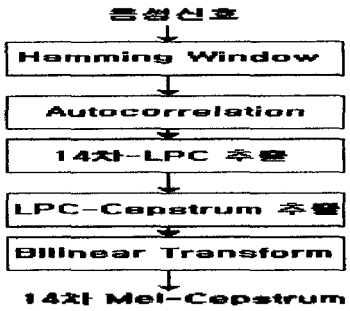

그림 2.3 특징벡터 추출

fig 2.3 Feature extraction

다음으로 사람의 청각적인 특성이 낮은 주파수에는 민감하나 높은 주파수에는 둔감하다는 특징이 있다. 따라서 이러한 청각적 특성을 반영하는 사용자의 음성특징을 추출하기 위해 Bilinear 변환을 이용하여 최종적으로 14차 Mel-Cepstrum을 구하였다.

위와 같은 과정을 거친 후 각각의 사용자에 해당하는 특징벡터를 시스템에 저장한다. 이를 기준패턴이라 한다. 사용자 인식은 기준패턴과 사용자가 시스템에 접속하기 위해 발성하는 음성에서 추출한 특징벡터인 시험패턴과 패턴매칭을 통해 사용자를 확인하는 것이다. 본 논문에서 사용한 사용자의 음성은 문장종속이며 고립단어이다. 그리고 패턴매칭 방법은 비선형 왜곡함수를 이용한 인식방법인 DTW이다.

3. 새로운 기준패턴 선정 방법

DTW을 이용한 화자인식방법은 사용자의 기준패턴 선정시 사용자가 한번 발성한 음성을 사용한다. 따라서 사용자의 화자내 변이가 클수록 인식률이 저하된다. 이러한 문제점을 해결하기 위한 방법으로는 대표평균패턴을 이용한 방법이나 가중 켑스트럼을 이용한 방법 등 여러 가지 방법이 있다[5]. 그러나 이러한 방법들은 기준패턴 선정이 긴 시간간격을 두고 기준패턴을 추출해야 한다는 단점이 있다. 따라서 본 논문에서는 사용자가 시스템을 사용하면서 자동적으로 사용자의 시간적 변화를 수용할 수 있는 기준패턴을 선정하는 방법과 초기 기준패턴 선정 시 사용자의 음성특징 정보가 부족하다는 초기의 DTW 알고리즘의 단점을 보완하기 위해 다음과 같은 방법을 제안한다.

먼저 초기 기준패턴 선정은 화자가 N번 발성한 음성을 VQ를 이용하여 정규화한다. 정규화 방법은 일반적인 VQ와는 다르게 아래와 같이 처리한다.

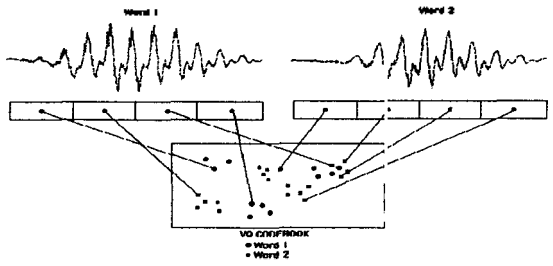

그림 3.1 백터양자화의 원리

fig 3.1 Vector Quantization

그림 3.1과 같이 한명의 화자가 N번 발성한 음성에 대해 하나의 코드북을 작성한다. 따라서 시스템에 M명의 사용자가 등록되어 있을 때 전체 코드북의 개수는 M개이며 M개의 코드북을 작성하기 위해 발성해야할 발성횟수는 M × N번이다.

사용자의 시스템 등록 과정은 다음과 같다. 먼저 시스템에 등록할 M명의 화자가 각각 N씩 발성한다. 이렇게 M개의 코드북을 작성한 뒤 코드북을 작성할 때 사용된 N개의 음성파일을 이용하여 새로운 코드북을 작성한다. 새로운 코드북의 작성방법은 다음과 같다.

예를 들어 n번째 사용자(화자)에 대한 새로운 코드북의 생성은 다음과 같은 알고리즘을 적용한다. 먼저 n번째 화자의 코드북 C_m을 작성한다.

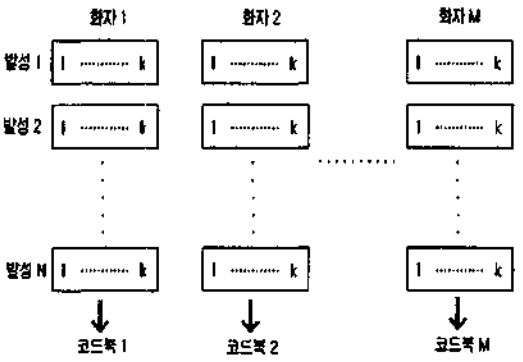

그림 3.2 새로운 기준패턴 알고리즘

fig 3.2 New algorithm of reference pattern

N : 한명의 사용자가 발성한 발성 횟수

M : 시스템에 등록할 사용자의 수

K : 한명의 화자가 등록할 음성의 음절 수

c_{nk} : n번째 발성의 k번째 음절의 코드워드 열

C_{mk} : m번째 화자의 k번째 음절에 대한 새로운 코

드 워드열

C_m : m번째 화자의 코드북

$C'_m = \{c_{m1}.....c_{mK}\}$: m번째 화자의 새로운 코드

북

$$C_{mk} = \frac{1}{N}\sum_{n=1}^{N} c_{nk}, \quad 1 \le k \le K, \; 1 \le m \le M \qquad (3.1)$$

다음으로 c_{nk}를 구하여 (식 3.1)과 같이 새로운 코드북을 작성한다. 각각의 c_{nk}을 이용하여 새로운 코드워드 열 C_{mk}를 구한다. 이때 각각의 발성에 대한 k번째 음절에 대한 프레임의 수가 다른 발성음에 대해서는 C_{mk}를 구할 때 이를 생략한다. 결과적으로 새로운 코드워드가 사용자의 기준패턴이 된다.

위와 같은 방법으로 화자내 변이와 화자간 변이를 수용하는 초기 기준패턴을 만든다.

다음으로 사용자가 시스템에 접속할 때 기준패턴을 변경하는 방법이다. 사용자의 음성은 시간이 지남에 따라 변화함으로 이러한 문제를 해결해야 한다. 본 논문에서는 사용자가 시스템을 사용할 때 사용되는 시험패턴을 이용하여 기준패턴을 변화시킨다. 그 방법은 다음과 같다. 사용자가 등록한 기준패턴을 R이라 하자. 사용자가 시스템에 접속하기 위해 사용한 시험패턴을 R'이라 하고 기존의 기준패턴을 R이라고 하면 새로운 기준패턴 $\hat{R}$은 다음과 같다. 여기서 R'은 사용자라고 옳게 인식되었을 때의 시험패턴을 나타낸다.

$$\hat{R} = \epsilon R + (1 - \epsilon)R' \qquad (3.2)$$

여기서 ϵ은 패턴의 가중치 계수를 나타내며 본 논문에서는 0.98로 정하였다.

위와 같이 기준패턴을 새롭게 함으로써 사용자의 시간적 변이를 막을 수 있다. 따라서 제안하는 초기 기준패턴 선정방법으로 사용자의 불충분한 음성특징을 보완할 수 있어 사용자의 음성특징 변이를 수용할 수 있으며 기준패턴의 변화를 주어 사용자의 시간적 변이 문제를 해결할 수 있다. 이렇게 함으로써 기존의 DTW를 이용한 사용자 인식방법의 문제점을 해결 할 수 있다.

4. 실험 및 결과

본 논문의 알고리즘을 모의 실험하기 위해 다음과 같은 처리를 수행하였다.

먼저 음성신호는 IBM-PC 220MHz에 A/D 변환기를 장치하여 11khz로 샘플링하고 16bits로 양자화하였다. 실험에 참가한 사용자는 남녀 각각 5명으로 사용된 어휘는 그들의 이름을 고립어로 발성한 음성이다. 실험은 2일 간격으로 30일간 수행하였으며 초기 기준패턴 선정은 사용자 각각 10번씩 2일 동안 발성한 음성으로 처리하였다. 특징벡터로는 30msec의 hamming window를 15msec마다 overlap하여 14차 Mel-Cepstrum을 추출하였다. VQ의 코드북 크기는 512로 하였으며 인식방법은 DTW를 택하였다. 실험은 일반 실험실에서 수행하였고 기존의 방법과 비교하여 성능을 살펴보았다. 제안한 방법의 전체적인 처리는 아래 그림 4.1과 같다.

각각의 시스템 성능을 보다 세밀히 조사하기 위해 사칭자 5명을 택하여 사용자와 사칭자에 대해 각각 False Accept와 False Reject의 비율을 구하였다. 실험 결과 제안한 방법의 인식률이 기존의 방법에 비해 3.3% 향상되었다.

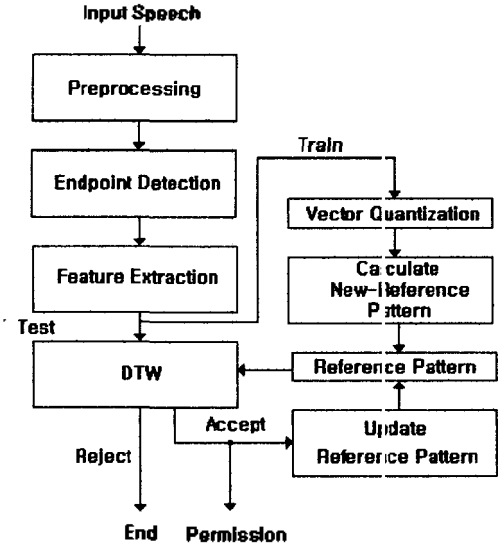

그림 4.1 제안한 알고리즘

fig 4.1 Proposed algorithm

표 4.1 실험 결과(%)

Table 4.1 Recognition accuracy

인식률 / 방법	FA	FR	전체 인식률
기존의 방법	2.6	3.2	94.2
제안한 방법	0.9	1.6	97.5

5. 결론

근래에 들어 음성을 이용한 시스템 사용자 확인에 대한 연구가 네트워크의 발달과 함께 활발히 진행되고 있다. 이렇게 음성을 이용한 사용자 확인에는 그 알고리즘에 따라 여러 가지 방법 등이 제안되고 있다[5]. 본 연구에서는 DTW를 이용한 시스템 사용자 인식을 수행하였다. 기존의 DTW방법은 초기 기준패턴 선정시 사용자 음성 특징이 불충분하다는 단점과 사용자의 음성이 시간이 지남에 따라 변화하는 문제로 인해 인식률이 저하된다는 단점이 있다.

따라서 본 연구에서는 이러한 문제들을 VQ를 이용하여 정규화 시킴으로써 초기 불충분한 음성 데이타의 문제점을 해결하고 기준패턴을 변화시킴으로써 사용자의 시간적 변이를 수용하여 해결하였다. 그러므로 시스템에 등록되어 있는 각각의 사용자의 화자내 분포를 수용할 수 있는 새로운 기준패턴을 선정함으로써 인식률 향상을 꾀하였다.

본 논문의 실험 결과 기존의 방법에 비해 3.3% 전체인식률이 증가함을 알 수 있었다.

6. 참고 문헌

[1]. LR. Rabiner, Juang., "Fundamentals of speech recognition", 1993., Prentice-Hall.

[2]. LR. Rabiner, R.W. Schafer., "Digital processing of speech signals", 1978., Prentice-Hall.

[3] P.Latace, R. DeMori., "Speech recognition and understanding recent advances trends and applications", 1990., NATO.

[4] Claudio Becchetti and Lucio prina ricotti. "Speech recognition", 1999., John Wiley & Sons

[5] 정종순, "대표평균패턴과 가중 켑스트럼을 이용한 화자인식의 성능 향상에 관한 연구", 1996., 석사학위 논문, KAIST

[6] 정종순, 배재옥, 배명진, "윈도우 환경에서 음성을 이용한 사용자 확인에 관한 연구", 한국음향학회지, Vol. 17, No. 5, 1998.

[7] 구명완외, "실시간 음성 끝점 검출 알고리즘", 제 5회 신호처리 합동학술회 논문집, 제 5권 1호, pp. 11-14, 1992

[8] 배재옥, 오세영, 배명진, "F1/F0을을 이용한 화자인식의 성능 향상에 관한연구", 한국음향학

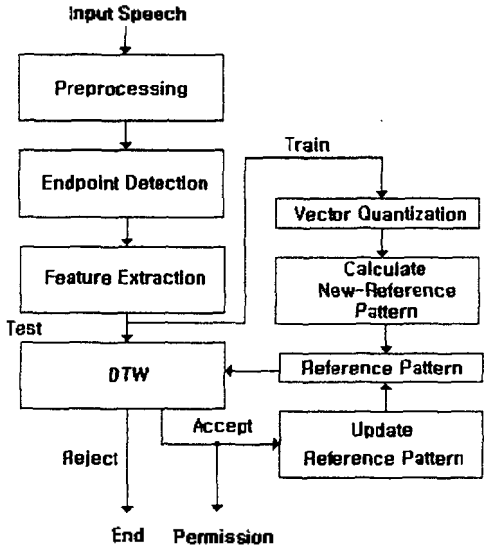

그림 4.1 제안한 알고리즘

fig 4.1 Proposed algorithm

표 4.1 실험 결과(%)

Table 4.1 Recognition accuracy

인식률 / 방법	FA	FR	전체 인식률
기존의 방법	2.6	3.2	94.2
제안한 방법	0.9	1.6	97.5

5. 결론

근래에 들어 음성을 이용한 시스템 사용자 확인에 대한 연구가 네트워크의 발달과 함께 활발히 진행되고 있다. 이렇게 음성을 이용한 사용자 확인에는 그 알고리즘에 따라 여러 가지 방법 등이 제안되고 있다[5]. 본 연구에서는 DTW를 이용한 시스템 사용자 인식을 수행하였다. 기존의 DTW방법은 초기 기준패턴 선정시 사용자 음성 특징이 불충분하다는 단점과 사용자의 음성이 시간이 지남에 따라 변화하는 문제로 인해 인식률이 저하된다는 단점이 있다.

따라서 본 연구에서는 이러한 문제들을 VQ를 이용하여 정규화 시킴으로써 초기 불충분한 음성 데이타의 문제점을 해결하고 기준패턴을 변화시킴으로써 사용자의 시간적 변이를 수용하여 해결하였다. 그러므로 시스템에 등록되어 있는 각각의 사용자의 화자내 분포를 수용할 수 있는 새로운 기준패턴을 선정함으로써 인식률 향상을 꾀하였다.

본 논문의 실험 결과 기존의 방법에 비해 3.3% 전체인식률이 증가함을 알 수 있었다.

6. 참고 문헌

[1]. LR. Rabiner, Juang., "Fundamentals of speech recognition", 1993., Prentice-Hall.

[2]. LR. Rabiner, R.W. Schafer., "Digital processing of speech signals", 1978., Prentice-Hall.

[3] P.Latace, R. DeMori., "Speech recognition and understanding recent advances trends and applications", 1990., NATO.

[4] Claudio Becchetti and Lucio prina ricotti. "Speech recognition", 1999., John Wiley & Sons

[5] 정종순, "대표평균패턴과 가중 켑스트럼을 이용한 화자인식의 성능 향상에 관한 연구", 1996., 석사학위 논문, KAIST

[6] 정종순, 배재욱, 배명진, "윈도우 환경에서 음성을 이용한 사용자 확인에 관한 연구", 한국음향학회지, Vol. 17, No. 5, 1998.

[7] 구명완외, "실시간 음성 끝점 검출 알고리즘", 제 5회 신호처리 합동학술회 논문집, 제 5권 1호, pp. 11-14, 1992