윈도우즈 환경에서 F_1/F_0율을 이용한 화자인식의 기준패턴 형성에 관한 연구

°정종순, 이윤주, 배재옥, 배명진(*)
숭실대학교 정보통신공학과

On a Updating Reference Pattern of Speaker Recognition using F1/F0 in the WINDOWS Environment

JongSoon Jung, YoonJoo Lee, Jaeok Bae and Myungjin Bae
Dept. of Telecommunication, Soongsil University
E-mail(*) : mjbae@saint.soongsil.ac.kr

Abstract

윈도우즈 95와 같은 멀티미디어 환경 하에서 개인 신분 확인을 위한 방법은 비밀번호를 키보드로 입력받는 것이었나, 본 논문에서는 음성을 이용하는 방법으로 기존의 방법이 기준패턴의 시간에 따라 변하는 특성을 보상하지 못한다는 단점을 보완하는 방법이다. 즉, 이를 위해 음성신호의 특징인 기본주파수와 제 1 포만트의 비율을 이용하여 기준패턴을 형성화하는 방법에 관한 것이다. 제안한 방법으로 실험한 결과, 98%의 전체 인식율을 얻게 되었고 윈도우즈 환경에서 비밀번호 사용 대신 음성 사용에 대한 가능성을 보여 주었다.

Ⅰ. 서론

현대의 모든 정보 미디어가 멀티미디어 환경에서 사용될 수 있게 변함에 따라 기존의 DOS환경보다는 Windows 95환경에서 거의 모든 작업이 이루어지고 있다. 윈도우즈 95환경에서는 개인의 신분확인을 중요한 문제로 다루어, 어떤 작업이나 디렉토리에 모두 사용자의 신분확인 기능을 부여하고 있다. 예를 들면, 공유 폴더의 비밀번호, 윈도우즈 95시작할 때, 네트워크의 사용 여부를 위한 비밀번호 그리고 스크린 세이버에서 다시 원래 상태로 돌아오기 위한 비밀번호 등이 있다. 그러나 이러한 방법은 타인에게 쉽게 도용될 수 있기 때문에 보다 확실한 개인확인 수단이 필요하다. 특히 정보의 접근이 전화나 통신망 등을 이용하여 원격지에서 이루어지는 경우, 개인확인은 더욱 어렵다. 이에 비해 화자인식은 음성에 포함되어 있는 화자정보를 추출하여 개인을 확인하는 기술로서 사칭자에 대한 대처, 처리시간, 원격지 확인 등 여러 측면에서 가장 효과적인 방법중 하나이다 [1][6]. 또한, 음성을 확인 매체로 사용하기 때문에 편리하다. 그러나 개인확인을 위해 등록되어 있는 사용자의 음성은 시간이 경과함에 따라 사칭자의 음성을 허용하는 에러율이 증가하게 된다. 이러한 문제점을 해결하기 위해 벡터 양자화나 대표 평균 패턴과 같은 학습기능을 사용하여 시간적 학습기능을 수행하였다. 그러나 이와 같은 방법은 화자마다의 특별한 특성이 줄고 평균화되는 특성을 가지고 있기 때문에 사칭자의 거부능력을 저하시키는 요인이 되고 있다. 따라서 본 논문에서는 화자의 데이터를 이전 데이터와 평균화를 시키지 않고, 새로 들어온 음성 신호를 양자화한 뒤 얻어진 오차 신호가 음성신호의 저역성분을 가진다는 특징을 이용하여 새로운 기준패턴을 형성하는 방법에 관한 것이다.

Ⅱ. 패턴정합을 이용한 화자인식 시스템

화자인식 시스템은 인식기술에 따라, 패턴정합, 신경회로망, 벡터 양자화 그리고 HMM(Hidden Markov Model)등으로 구분한다. 패턴정합 시스템은 비교적 간단한 알고리즘과 최소의 하드웨어를 요구하므로 간단한 응용분야에 성공적으로 이용할 수 있기 때문에 본 논문에서는 이 방법을 이용하였다.

화자인식은 인식대상에 따라 화자확인과 화자식별로 크게 나눈다. 화자확인은 입력된 음성이 본인의 것인 지의 여부를 판정하는데 비해, 화자식별의 경우는 입력된 미지의 음성이 이미 등록된 여러 명의 화자 중 어떤 화자에 의해 발성된 음성인지를 판정하는 것을 말한다.

화자가 발성할 문장이 고정되어 있는 화자인식 시스템을 문장 의존형이라 하고, 문장이 정해져 있지 않고 자유롭게 발성하는 경우를 문장독립형이라 한다. 본 시스템은 지정된 단어나 문장을 이용하므로 문장종속 화자확인이다. 그림2.1은 패턴정합을 이용한 화자확인 시스템의 블록도 이다. 화자확인을 위한

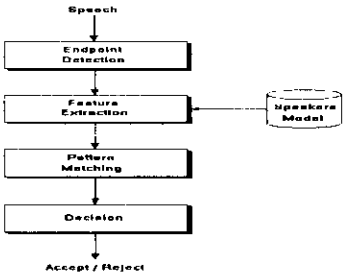

그림2.1 화자확인 시스템의 블록도

입력은 등록자의 음성과 사칭자의 음성이다. 입력된 음성은 끝점검출을 하고 특징추출을 한 다음, 패턴정합을 거치게 된다. 즉, 기준(등록자) 패턴과 입력패턴의 거리를 계산하여 사용자인지 사칭자인지를 구분하게 된다.

2.1 끝점 검출

음성구간 검출의 정확성 여부는 음성인식의 성능에 큰 영향을 미치기 때문에 정확한 검출이 요구된다. 따라서 본 논문에서는 단구간 에너지와 영교차율의 문턱치를 adaptive thresholding method를 사용하여 조절하는 방법을 사용하였다. 즉 단구간 에너지는 입력된 데이터만을 이용하여 이 값이 큰 부분은 음성구간으로 작은 구간은 묵음구간으로 결정하는 방법이다. 그러나, 이러한 방법은 무성자음과 같이 에너지가 작은 부분이나 배경잡음이 큰 경우에는 음성과 묵음을 구별하기가 어렵다. 따라서 이를 보완하기 위하여 영교차율을 사용하였다. 그리고 어떤 단어에는 pause가 존재하는데 위와 같은 방법을 사용하면 pause 전에서 끝점이 검출 될 수 있다. 따라서 본 논문에서는 이러한 오류를 개선하기 위하여 끝점이 검출된 이후에 40 프레임내에서 다시 음성의 시작점이 검출되면 pause가 있는 음성으로 간주하고 다시 끝점을 검출한다. 이때 pause 검색구간을 0.4초로 하였다[5].

2.2 특징 추출

한 프레임의 길이는 30ms로 한 프레임의 구간은 10ms로 하였다. 윈도우는 해밍 윈도우를 이용하였다. 전처리 과정을 거친 음성 데이터로부터 16차 LPC 계수를 구하였다. 본 논문에서는 LPC 계수를 이용해서 켑스트럼을 구하고, 귀의 선형특성을 고려해 켑스트럼을 mel-scale로 warping시킨 mel-cepstrum을 특징 파라미터로 사용하였다. 따라서 특징추출 과정은 음성 데이터로부터 LPC계수를 구하는 과정 그리고 이를 mel-scale로 변환시키는 과정으로 구성한다. 그림2.2는 특징추출 과정을 나타낸다.

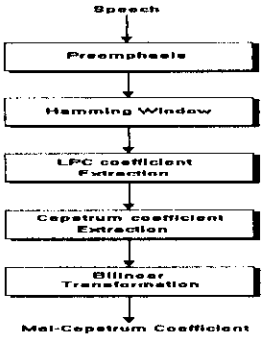

그림2.2 음성 특징 추출 과정

2.3 DTW를 사용한 패턴정합

화자가 일정한 발성을 할 경우라도 발성의 시간적인 변이가 존재한다. 이렇게 테스트발성과 기준발성은 일반적으로 다른 지속시간을 가진다. 대부분 양질의 화자인식기는 테스트와 기준 templates에서 유사한 음성 segments를 배열시키기 위해서 한 template에서 다른 template로 비선형적으로 왜곡시킨다. 이런 과정을 Dynamic Time Warping(DTW)라 한다. 프레임 대 프레임의 선형적인 편차는 한쌍의 프레임 거리가 다른 곳에 위치한 것과 비교하여 작을 때만 허락한다[2].

2.4. F-ratio를 이용한 가중 켑스트럼

화자간의 변별력을 극대화하기 위하여 가중치로 F-ratio 값을 사용하였다. F-ratio 는 특징파라미터의 유용성 척도로 주로 사용되는 것으로 화자 내의 변이로 화자간의 변이를 나눈 값이다. F-ratio의 정의는 식(2.1)과 같다.

$$F-ratio = \frac{variance\ of\ speaker\ means}{mean\ intraspeaker\ variance} \qquad (2.1)$$

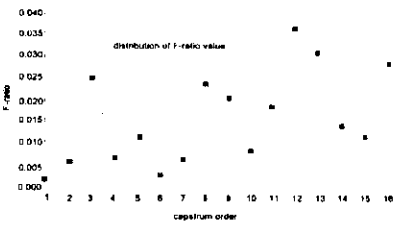

그림2.3 켑스트럼 차수에 대한 F-ratio 값의 분포

좋은 특징 파라미터는 화자간의 변이는 크고 화자내의 변이는 작은 것이다. 즉, F-ratio 값이 클수록 좋은 파라미터라 볼 수 있다. 그림2.3은 본 논문에서

사용한 F-ratio분포도를 나타내었다[3].

Ⅲ. F_1/F_0율을 적용한 기준패턴

3.1 양자화 오차를 이용한 F_1/F_0율

음성신호를 M비트로 선형 양자화한 신호 s(n)은 다음과 같이 나타낼 수 있다.

$$s(n) = \sum_{i=0}^{M-1} a_i 2^i = \sum_{i=0}^{N-1} a_i 2^i + \sum_{i=N}^{M-1} a_i 2^i = Q_L + Q_H \quad (3.1)$$

여기서 Q_L은 음성신호를 (M-N)비트로 부호화할 때 발생하는 양자화 오차이다.

유성음 파형의 경우에 낮은 쪽 포만트는 높은 쪽의 포만트에 비해 에너지가 아주 높기 때문에 그림 3.1(b)와 같이 에너지가 우세한 기본주파수는 Q_L의 최대진폭을 유지하게 된다. 그림3.1(c)는 양자화 오차 Q_L의 정규화된 파형을 나타낸 것이다. 이렇게 정규화된 파형에서 식(3.2)에서와 같이 기본 주파수(F_0)를 구하고 한 피치구간에서의 Zero Crossing Rate (ZCR)의 역수를 $2F_1$의 주파수와 같게 하여 제1포만트를 구함으로써 F_1/F_0율을 추출하였다[4].

$$PITCH(fr) = \frac{RTCR}{N_S - N_E} \quad (3.2)$$

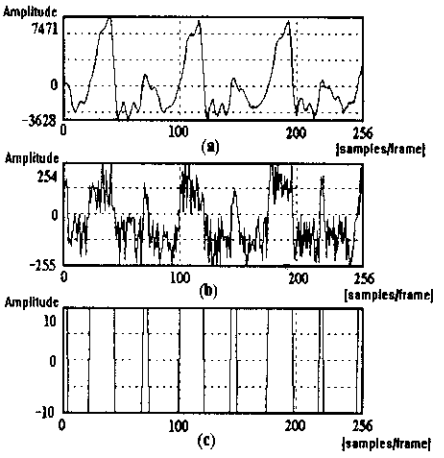

그림3.1 F_1/F_0율 검출 예
(a) 음성파형, (b) 양자화 오차 신호
(c) 정규화된 양자화 오차 신호

3.2 F_1/F_0율을 적용한 기준패턴

기존의 화자확인 시스템은 시간에 따른 화자내 변이 때문에 여러 개의 기준패턴을 사용한다. 예를 들어, 기준패턴의 수가 N개인 경우 테스트패턴과 N번을 비교하여 최소의 거리 값을 구해야 하므로 처리해야할 데이터 수가 많게 된다. 그리고 이 경우, 시간에 따라 음성을 받을 때는 시스템에 따라 일정한 기간을 두어야 한다. 만약 어떤 한 사용자에 오랜 기간 동안에 걸쳐 받은 음성으로 기준패턴을 사용하면, false acceptance 확률이 증가하게 되어 화자 인식율을 저하시키는 원인이 된다. 반대로 짧은 기간 동안 발성된 음성을 사용하면 false rejection확률이 높아져 이것 또한 인식율을 저하시키는 원인이 된다. 따라서 본 시스템에서는 이러한 모든 것을 고려하기 위해서 7개월에 걸쳐 사용자에 해당하는 음성을 받았다. 받은 음성에서 기준패턴이 될 하나의 패턴을 찾기 위해서 그림3.3에서 보듯, 음성의 끝점을 검출하고 검출된 구간에서 정규화된 양자화 오차신호를 구한다. 그리고 구해진 양자화 오차 신호로부터 F_1/F_0율을 구한다. 이렇게 구해진 F_1/F_0율을 모두 다해서 N개로 나누어 해당 사용자에 기준패턴으로 사용할 1차 문턱값을 찾아 낸다.

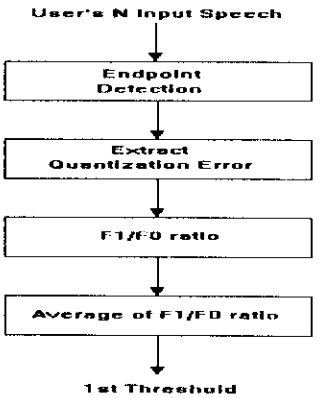

그림3.3 F_1/F_0율 이용한 기준패턴 형성과정

Ⅳ. 제안한 시스템의 전체적인 구조

그림4.1은 본 시스템의 전체적인 화자확인 시스템을 보여주고 있다. 본 시스템의 입력은 사용자의 음성과 사칭자의 음성이다. 우선, 사용자에 해당하는 각 사람에 대한 음성의 특징 추출은 앞 절에서 설명한 알고리즘을 사용하여 구한다. 구해진 특징 파라미터를 이용하여 각 사용자에 해당하는 F_1/F_0율을 구하여 기준패턴으로 사용한다. 그 다음 이 패턴과 테스트 패턴간의 거리를 계산하는데 화자간의 변별력을 분명히 하기 위해서 사용자에 해당하는 가중켑스트럼을 구해 적용한다. 그리고 마지막으로 사용자의 문턱 거리값에 따라 가부를 결정한다.

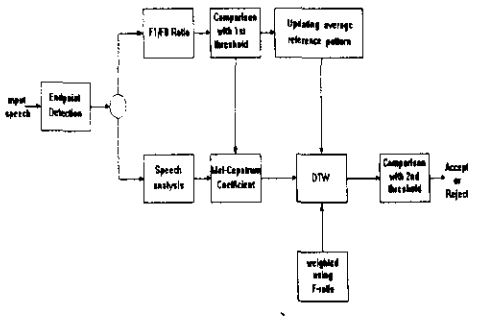

그림4.1 제안한 시스템의 전체적인 블록도

V. 실험 및 결과

본 논문을 구현하기 위해 사용한 음성데이타는 주변잡음이 있는 일반 실험실 환경에서 성인 남녀 10명이 발성한 음성으로 구성하였다. 음성 시료는 3음절의 이름으로 ASPI사의 ELP보드에서 8KHz로 샘플링하고 16bit로 양자화하여 사용하였다. 사용자는 기준패턴으로 사용하기 위해서 7개월간에 걸쳐서 본인의 이름과 다른 사용자의 이름을 발성하였고, 사칭자는 각 사용자의 이름을 발성하였다. 테스트 패턴은 8개월간 같은 방법으로 사용자와 사칭자가 발성한 음성을 가지고 임의로 선택하여 사용하였다. 그림5.1은 본 논문에서 제안한 전체 블록도로 다음과 같다. 비교 1에서는 음성이 들어오면 패턴매칭법으로 비교한 뒤, 계산된 거리가 미리 설정된 문턱값과 비교하여 문턱값보다 작으면 수락, 크면 거절하게 된다. 수락된 경우는 비교2를 거치게 하여 미리 설정된 또 다른 문턱값으로 여과시킨다. 이렇게 해서 구해진 데이터를 F_1/F_0비율로 설정된 문턱값과 다시 한번 더 비교해 완벽한 기준패턴으로 갱신한다. 이렇게 갱신을 수행하였을 경우 사칭자 에러율을 더 낮게 할 수가 있고 새로운 패턴으로 기준패턴을 갱신하기 때문에 더 높은 인식률을 얻을 수 있다.

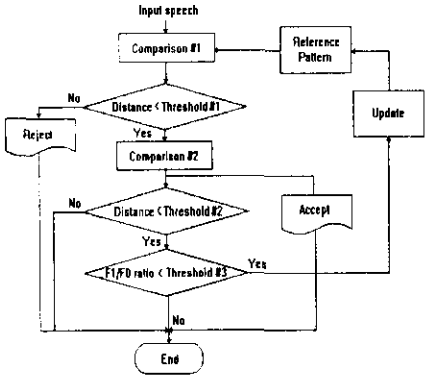

그림5.1 본 논문에서 제안한 블록도

그림5.1은 본 논문에서 실험한 결과를 보여주고 있다. 제안한 시스템의 인식율과 가중 켑스트럼을 사용하였을 때 인식율은 98%로 높은 인식율을 얻을 수 있다.

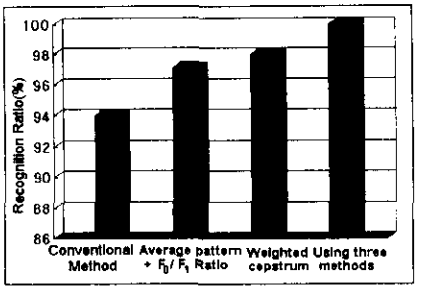

그림5.1 가중화된 켑스트럼과 F1/F0율에 의해 기준패턴 형성을 이용한 결과

VI. 결 론

본 논문은 윈도우즈 환경에서 기존의 텍스트 종속 화자인식시스템을 향상시키는 방법으로 기준패턴이 시간에 따른 변이를 보상하지 못한다는 단점을 보완하기 위한 방법으로 음성신호의 특징인 기본주파수와 제 1 포만트의 비율을 이용하여 기준패턴을 형성하였다. 제안된 방법으로 실험한 결과 인식율이 98%로 기존의 화자인식 시스템보다 4%정도의 인식율 향상을 확인하였다.

VII. 참 고 문 헌

[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] H.Ney & R.Gierloff, "Speaker Recognition Using a Feature Weighting Technique," *Proc. of ICASSP*, pp. 1645-1648, 1982

[4] Jaeok Bae, Seyoung Oh & MyungJin Bae, "On a Performance Improvement of Speaker Recognition using F_1/F_0 Ratio", Conf. *of Acous*, Korea. Vol. 16, No. 2, pp. 137-140, 1997

[5] 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

[6] D. O'Shaughnessy, "Speaker Recognition," *IEEE ASSP Magazine*, Oct. pp. 4-17, 1986.