고차 반사계수 특성을 이용한 화자인식의 성능 향상에 관한 연구

이윤주, 오세영, 함명규, 배명진
숭실대학교 정보통신학과

On a Study of the Improvement of Speaker Recognition with Characteristics of High Order Reflection Coefficients

Yoonjoo LEE, Seyoung OH, Myungkyu HAM, Myungjin BAE
Dept. of Telecommunication, Soongsil University
E-mail : mjbae@saint.soongsil.ac.kr

ABSTRACT

As the number of reference patterns increase in the text dependant speaker recognition, the recognition performance of the system degrades. So, if reference patterns were decreased the high recognition rate can be obtained. It's because the speaker recognition can obtain the high discrimination. In this paper, to decrease the number of reference patterns, we choose candidate reference patterns to perform pattern matching with test pattern by high order component of the reflection coefficients of the uttered speech signal

Consequently the total recognition rate of the proposed method is about 2% higher than that of the conventional method.

1. 서 론

개인이나 특정 단체의 정보의 보완을 위해서는 사용자의 확인 과정이 필요하다. 이때 확인 절차는 사용자에게 사용이 용이해야 하며 확인 내용은 정확해야 한다. 이러한 점을 고려하여 근래에 들어 사용자의 음성특성을 이용한 사용자 확인 방법이 고안되었다. 즉, 사용자가 특정 패스워드(Password) 또는 임의의 음성을 발성한 뒤 발성된 음성을 바탕으로 사용자를 확인하는 방법이다. 이러한 방법에는 화자가 발성한 음성으로부터 스펙트럼의 특성을 나타내는 특징벡터를 추출하여 저장된 각각의 기준패턴과 패턴매칭(Pattern Matching)을 통해 화자를 인식하는 방법이 있다. 음성신호의 패턴매칭을 이용한 화자인식 방법에는 동적패턴정합(Dynamic Time Warping-DTW)법이 있다. 그러나 이 방법은 특정 시스템에 등록된 사용자의 수가 증가함으로써 비교패턴과 패턴매칭을 수행할 기준패턴의 수가 증가하게 되므로 데이터량이 많아지게 된다. 따라서 이로 인해 사용자 확인 시간과 오인식률이 늘어나게 된다 [1].

그러므로 이러한 단점을 보완하기 위해서 본 논문은 화자들의 특성중 화자간변이(Inter-Speaker Variance)를 이용하여 비교패턴과 패턴매칭을 수행할 몇 개의 기준패턴을 후보자로 선별함으로써 인식률을 향상시키는 방법에 관한 것으로 다음과 같은 특성을 이용한다.

일반적으로 음성신호의 반사 계수 특성은 1차 반사계수는 +1에 근접해 있고, 2차 계수는 -1에 근접해 있다. 또한 고차로 갈수록 0에 근접한 값으로 감소하는 특징을 가지고는 있지만 화자마다 그 값의 범위는 크게 다르다[2]. 그러므로 본 논문에서는 이러한 반사계수의 고차특성을 이용하여 화자인식시 비교하는 대상을 줄이는 방법에 관한 것이다.

2.일반적인 화자 인식 시스템

2.1 화자 인식의 분류

일반적으로 화자 인식은 크게 두 가지로 나누어 처리되고 있다. 첫째로 화자식별(Speaker Identification)은 등록된 화자집단에 지금 요청중인 화자의 발성이 등록되어 있는지를 결정하는 과정이다. 둘째로 화자확인(Speaker Verification)은 지금 발성중인 화자가 인식시스템이 요청한 그 사람인지 아닌지

(Yes-no task)를 결정하는 과정이다.

또한 화자인식은 인식 방법에 따라 4가지로 구분할 수 있다. 첫째로 패턴정합법(Pattern Matching)에 의한 동적 정합(Dynamic Time Warping)은 입력패턴을 미리 정해진 기준 패턴과 비교하여 최적화된 유사성을 판단하는 방법이다. 둘째로 신경회로망을 이용한 방법은 각 화자별로 신경회로망을 구성하고 화자간의 변별력을 갖도록 학습을 수행하는 인식 방법이다. 그러나 이 방법은 새로운 화자의 추가시 인식 시스템을 다시 학습시켜야 하고 고도의 병렬계산 능력이 요구되기 때문에 실제 응용시에는 적합하지 않다는 단점이 있다. 세 번째 방법인 벡터양자화 방법은 입력 패턴과 양자화 코드북(Codebook) 사이의 거리로 유사성을 판단하는 방법이지만 많은 학습자료가 필요하고 화자간의 동적인 변화 특성을 이용하지 못하기 때문에 인식률에 한계가 있다. 마지막으로 은닉마코프모델(Hidden Markov Model-HMM)은 학습기능을 이용하여 화자내의 변이를 흡수 할 수 있으며, 입력패턴의 비선형 정합을 수행하는 특성이 있다.

화자인식 시스템은 인식에 사용하는 문장의 종속 여부에 따라 정해지지 않는 어휘로 인식을 수행하는 텍스트 독립형(Text Independant)과 정해진 어휘만을 발성해야 하는 텍스트 종속형(Text Dependant)으로 나눌 수 있다.

2.2 화자 인식 과정

일반적으로 패턴매칭을 이용한 화자 인식 과정은 다음과 같다. 먼저 발성된 음성신호로부터 음성구간을 검출한다. 검출된 음성신호를 창함수를 이용하여 단구간으로 나눈다. 이렇게 단구간으로 나누어진 음성 데이터에서 화자의 특징벡터를 추출하여 기준패턴으로 사용한다.

이러한 방법으로 저장된 기준패턴들과 음성입력단에서 들어온 비교패턴을 DTW 방법을 이용하여 화자인식을 수행한다.

3. 화자내 변이와 화자간 변이

대부분의 화자 인식은 음성 분석을 통해 화자의 특징을 음향 파라미터 형태로 추출하여 각 화자의 기준패턴을 만든 후, 시스템에 입력된 미지의 음성 패턴(비교패턴)과의 차이를 계산하여 식별여부를 최종적으로 결정한다. 화자 인식에 사용하는 파라미터는 화자의 특징을 충분히 표현함과 동시에 변동이 작은 것이 바람직하다. 즉, 특징 파라미터의 화자내의 변이(Intra-speaker variance)보다 화자간의 변이(Inter-speaker variance)가 큰 특성을 가져야 화자간의 구분이 용이하다.

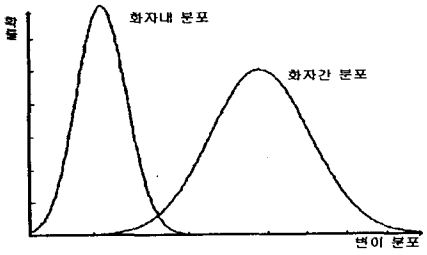

그림 3-1. 화자특징의 변이 분포

그림 3-1은 이러한 화자특징 변이 분포의 한 예를 나타낸 것이다. 그림에서는 두 분포가 서로 겹치는 부분이 있어 화자를 잘못 판정할 가능성이 존재함으로 화자간 오류를 최소화하기 위해 화자간의 구별이 뚜렷한 특징 파라미터나 분별력이 뛰어난 인식 방법이 필요하다.

4.반사계수

기본적으로 반사계수(Reflection Coefficient)를 구하기 위해서는 격자방법(Lattice Method)을 사용한다. 그러나 본 논문에서는 반사계수을 용이하게 추출하고 또한 화자인식을 적은 계산량으로 수행하기 위해서 선형예측계수(LPC Coefficient)를 통해 구하였다. 그 추출 과정은 아래 식과 같다[3].

$$E^{(0)} = r(0)$$

$$k_i = r(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} r(|i-j|)/E^{(i-1)}$$

$$a_i^{(i)} = k_i$$

$$a_j(i) = a_j(i-1) - k_i a_{i-j}^{(i-1)}$$

$$E^{(i)} = (1-k_i^2)E^{(i-1)} \text{ ,where } 1 \le i \le p$$

여기서 $a_m = a_m^{(p)}$는 선형예측 계수이고, k_m은 반사계수를 나타내며, $r(i)$는 음성 신호의 자기 상관 계수를 나타낸다.

위의 과정을 거쳐 반사계수를 구한 후 그 특성을 알아보기 위해 그림 4-1에 3명의 화자를 대상으로 분포특성을 나타내었다. 그림에서 보는바와 같이 1차 계수는 +1에 밀집되어 있고, 2차 계수는 -1의 값에 밀집되어 있는 특성을 가지고 있다. 특히 8차이하의 반사계수를 보면 화자간의 변이는 거의 나타낼 수 없다.

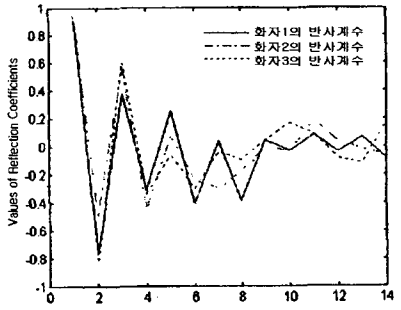

그림 4-1. 3명의 화자에 대한 반사계수 특성

그러나 8차 이상의 고차 계수로 갈수록 화자간의 변이는 커지는 것을 알 수 있다.

또한 그림 4-2와 같이 동일 화자내의 변이는 계수의 차수에 무관하게 매우 작음을 알 수 있다. 따라서 고차 반사계수를 이용하여 비교패턴과 패턴정합을 수행할 몇 개의 특정 기준패턴을 후보로 선정할 수 있다. 이렇게 함으로써 DTW를 이용한 화자 인식 방법의 단점인 처리시간 감소를 극복할 수 있다.

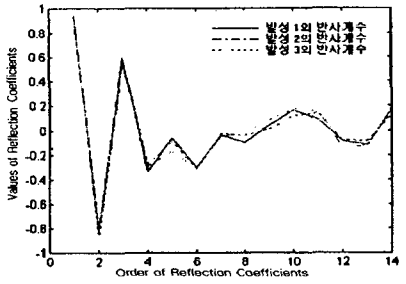

그림 4-2. 동일화자에 대한 반사계수

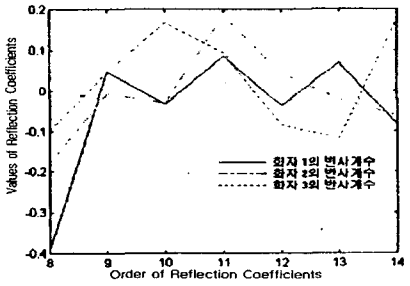

그림 4-3. 3명의 화자에 대한 고차 반사계수특성

본 논문은 화자들의 8차이상의 반사계수를 이용하여 인식시 비교할 후보자를 선정하는 것에 관한 것이다.

3명의 화자간 고차 반사계수의 특징은 그림 4-3과 같다. 그림에서 보는 바와 같이 화자간의 고차반사계수는 그 분포가 큰 차이를 나타낸다. 그러므로 이러한 특성을 이용하기 위해 음성특징인 선형예측계수(LPC)를 추출할 때 고차항의 반사계수를 구한다. 구해진 반사계수의 평균과 표준편차를 이용하여 DTW를 수행하기전 미리 저장된 기준패턴의 반사계수와 값을 비교한 뒤 등록된 각 화자의 기준패턴에서 후보자를 선정한다. 이때, 후보자는 그 값의 크기 순으로 하여 3명을 선정한다. 그리고 선정된 후보자에 한해서 DTW를 수행함으로써 시스템의 전체 인식 시간 단축뿐만 아니라 인식률 향상을 얻을 수 있다.

5. 화자인식 시스템 구현

5.1 음성구간 검출

본 논문에서는 음성구간을 검출하기전 먼저 안정된 피치구간을 찾은 뒤 무성음구간을 포함하기 위해 일정 범위내에서 입력된 음성을 모두 저장한다. 이렇게 저장된 음성구간에 대해서만 단구간 에너지와 영교차율을 이용하여 음성구간을 검출한다. 그리고 음절사이의 묵음구간이 존재 할 수 있기 때문에 끝점이 검출된 후에도 일정 프레임(Frame)동안 다시 음성의 시작점을 단구간 에너지를 이용하여 검출한다. 만일 또 다시 시작점이 검출되면 묵음구간이 존재하는 음성발성으로 간주하고 다시 끝점을 검출하는 과정을 반복적으로 수행한다[2].

5.2 특징 벡터 추출

본 논문에서는 화자의 특징 벡터로 14차 Mel-Cepstrum을 사용하였다. 그림 5-1과 같이 먼저 해밍윈도우(Hamming Window)를 사용하여 단구간으로 음성을 나눈다. 음성신호의 고주파항의 영향을 강조시키기 위해 프리엠퍼시스(Preemphasis) 필터를 사용하였고 이렇게 필터를 통과하여 나온 신호로부터 반사계수를 추출하고 평균값과 표준편차를 구한다.

구해진 반사계수를 이용하여 화자의 LPC 특성을 추출한다. 그리고 LPC-Cepstrum 변환식을 이용하여 14차 LPC-Cepstrum을 구한다. 이렇게 구해진 계수를 청각 특성을 고려한 Mel-Frequency 율로 왜곡시켜 특징 파라미터인 14차 Mel-Cepstrum을 구한다.

5.3 패턴 정합

본 논문에서 사용한 패턴정합 방법은 DTW방법이다. 이는 사용할 화자의 수가 20명으로 그 비교수가 작고 등록자가 사용하고 있는 텍스트의 시간적인 음운의 변화 특성이 중요하기 때문에 이 방법을 택

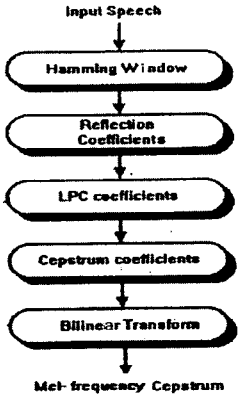

그림 5-1. 특징 벡터 추출

하였다[4]. 이 방법은 시간축을 비선형적으로 왜곡시켜 기준패턴과 비교패턴을 정합하는 방법으로 특징벡터의 시간적 변화를 수용할 수 있다.

이러한 DTW 방법을 이용하여 입력된 비교패턴과 후보자로 선정된 3개의 기준패턴간의 정합을 수행하여 최종적으로 화자를 인식한다.

6. 실험 및 결과

본 논문의 알고리즘을 모의실험하기 위해 IBM PC에 마이크가 장치된 16비트 A/D변환기를 인터페이스 시켰다. 실험은 일반 실험실 환경에서 20명의 남녀 화자가 각각 본인의 이름을 발성한 음성 시료를 11khz로 샘플링하고 16비트로 양자화하여 사용하였다. 한 프레임의 길이는 300샘플이며, 150샘플씩 오버랩(Overlab)시켜 특징벡터를 추출하였다. 인식을 위한 특징벡터로는 14차 Mel-Cepsturm을 사용하였다. 기준패턴으로는 20대 남녀 20명이 4번 발성하여 일주일 동안 발성된 음성을 사용하였다. 그리고 사칭자의 효과를 알아보기 위해서 4명으로 하여금 등록된 화자의 음성을 일주일 동안 4번씩 발성하게 하였다. 그림 6-1은 본 실험에서 사용한 화자인식시스템의 전체 블록도이다. 실험결과 제안한 방법이 전체 표준패턴과의 비교를 수행한 방법에 비해 전체 인식률이 2% 향상되었고 인식시간은 10.5% 감소하였다.

표 6-1. 인식률

	False Accept	False Reject	Recognition Accuracy
기존의 방법	0.83	4.17	95.0%
제안한 방법	0.33	2.47	97.0%

7. 결론

본 논문은 DP알고리즘을 사용한 텍스트 종속 화자인식시스템에서 패턴정합과정에서 발생하는 처리시간 증가와 오인식률 증가의 단점을 보완하기 위해 패턴정합을 수행할 데이터량을 줄이는 방법에 대한 것이다.

즉, 화자가 발성한 음성신호의 고차항의 반사계수는 화자마다 다른 분포도를 나타낸다는 특성이 있으므로 패턴정합을 수행하기전 이 반사계수의 분포도를 통해 미리 후보자를 선정하여 기존의 비선형 패턴매칭 알고리즘의 단점을 보안하였다.

실험결과 전체 패턴을 패턴정합하는 기존의 방법에 비해 10.5%의 시간단축을 얻을 수 있었고 전체 인식률이 2%향상되었다.

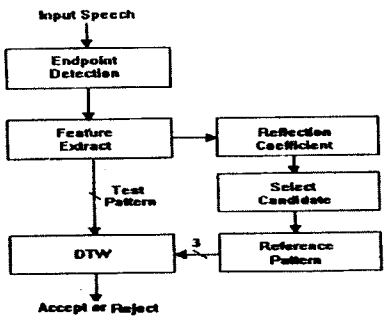

그림 6-1. 전체 처리 블럭도

8. 참 고 문 헌

[1] L. R. Rabiner & Biing-Hwang Juang, *Fundamentals Of Speech Recognition*, Prentice-Hall AT&T, U.S.A, 1993

[2] L. R. Rabiner & R.W.Schafer, *Digital Processing of Speech Signal*, Prentice-Hall, Englewood Cliffs, N.J., U.S.A., 1978

[3] A.M. Kondoz, *Digital Speech*, Jhon wiley & Sons, 1994

[4] Hiroaki Sakoe & Seibi Chiba, *"Dynamic Programming Algorithm Optimization for Spoken Word Recognition"*, IEEE Trans. on ASSP, vol.26, No.1, pp.43-49, Feb.1978.