음성의 청각특성을 이용한 화자식별시스템의 성능향상에 관한 연구

이윤주, 오세영, 배재옥, 배명진
숭실대학교 정보통신공학과
mjbae@saint.soongsil.ac.kr

On a Performance Improvement of Speaker Recognition by using the Auditory Characteristics of Speech

Yoonjoo LEE, Seyoung OH, Jaeok BAE, Myungjin BAE
Dept. of Information and Telecommunication Engr., SoongSil University
1-1 Sangdo-5Dong, Dongjak-Ku, Seoul 156-743, KOREA

Abstract

The pre-emephasis filter as the conventional method emphasizes all components of high frequency that reflects the speaker characteristics. However this filter don't show the auditory characteristics of speaker's speech. In order to emphasize the perceptual characteristics, we propose the speaker recognition system that uses the perceptual weighting as the preprocessor because the Auditory characteristic of human is sensitive to the formant peaks. This filter has the characteristcs that both deemphasizes the low-formants and emphasizes the high formants. As a result of the proposed method, we improve the total recognition rate 1.7% better than the conventional method.

1. 서 론

개인이나 특정 단체의 정보의 보완을 위해 사용자의 확인과정이 필요하다. 이 때 확인 절차는 사용자에게 용이하여야하며 확인 내용은 정확해야 한다. 이러한 점을 고려하여 근래에 들어 사용자의 음성특성을 이용한 사용자 확인 방법이 고안되었다. 즉, 사용자가 특정 Password 또는 임의의 말을 발성한 뒤 발성된 음성을 바탕으로 사용자를 확인하는 방법이다. 이러한 방법에는 화자가 발성한 음성으로부터 스펙트럼의 특성을 나타내는 특징벡터를 추출하여 패턴매칭을 통해 화자를 인식하는 방법이 있다. 화자의 특징 벡터 추출시 사람의 인지적인 특성이 부여된 특징 벡터를 추출한다면 사용자 확인의 정확성은 증가된다. 따라서 본 논문은 이러한 특징을 이용하여 음성신호를 인지 가중 필터에 통과 시켜서 사람의 귀가 잘 인지할 수 있는 고주파 포만트의 봉우리 특성을 강조시킴으로써 화자 인식 시스템의 성능을 향상시키는데 그 목적을 둔 것이다.

2. 일반적인 화자 인식 시스템

2.1 화자 인식의 분류

화자 인식은 크게 두 가지로 나누어 처리되고 있다. 첫째로 화자식별(Speaker Identification)은 등록된 화자 집단에 지금 요청중인 화자의 발성이 등록되어 있는지를 결정하는 과정이다. 둘째로 화자확인(Speaker Verification)은 지금 발성중인 화자가 요청한 그 사람인지 아닌지(yes-no task)를 결정하는 과정이다. 또한 화

자인식은 인식 방법에 따라서 다음과 같이 4가지로 구분할 수 있다. 첫째로 패턴정합법(Pattern matching)에 의한 동적 정합법(Dynamic time warping, DTW)은 입력패턴을 미리 정해진 기준 패턴과 비교하여 최적화된 유사성을 판단하는 방법이다. 둘째로 신경회로망은 각 화자별로 신경회로망을 구성하고 화자간의 변별력을 갖도록 학습을 수행하여 인식하는 방법이다. 그러나 이 방법은 새로운 화자의 추가시 다시 학습시켜야 한다는 단점과 고도의 병렬계산 능력이 요구되기 때문에 실제 응용시에서는 적합하지 않다. 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 음성신호를 확률적으로 모델링하여 처리하는 HMM(Hidden Markov Model)은 학습기능을 이용하여 화자내의 변이를 흡수 할 수 있으며, 입력패턴의 비선형 정합을 수행하는 특성이 있다. 마지막으로 화자 인식 시스템은 인식에 사용하는 문장의 종속여부에 따라 정해지지 않는 어휘로 인식을 수행하는 텍스트 독립형과 정해진 어휘만을 발성해야하는 텍스트 종속형으로 나눌 수 있다[1].

2.2 화자 인식 과정

전체적인 화자 인식은 그림 2-1과 같은 과정을 통해 처리한다. 먼저 발성된 음성신호로부터 음성구간을 검출한다. 검출된 음성신호를 단구간으로 나누어 화자의 특징벡터를 추출하여 시험패턴으로 사용한다. 마지막으로 시험패턴과 참조패턴과의 패턴 정합을 수행하여 화자를 인식한다.

3. 인지 가중 필터를 이용한 성능향상

인간의 귀에 대한 주파수 응답은 평탄(Flat)하지 않기 때문에 특징벡터 추출시 오차가 발생한다. 또한 화자인식에 유용한 특징벡터는 고차 포만트 성분에 있다. 따라서 이러한 특징을 잘 반영하기하고 또한 특징벡터 추출시 발생하는 오차를 최소화하기 위해서 인지적인 가중필터를 사용함으로써 특징벡터를 최적화할 수 있다. 즉, 인지 가중 필터를 사용하여 인지적으로 덜 중요한 주파수들의 영향을 상쇄하고 인지적으로 중요한 주파수의 영향을 극대화한다. 인지 가중 필터는 식 (3.1)과 같이 나타낼 수 있다.

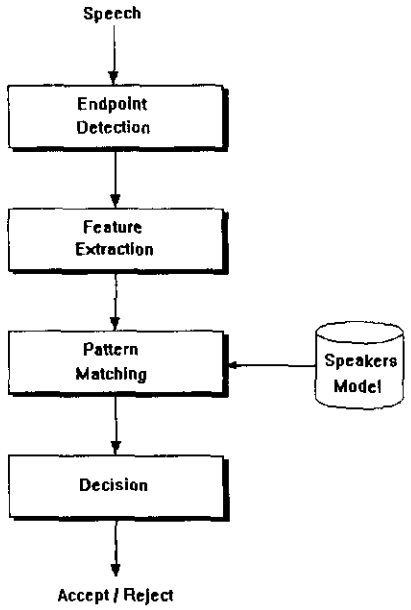

그림 2-1. 화자 인식 과정

그리고 이러한 인지 가중 필터의 특징은 그림 3-1과 같다[3].

$$W(z) = \frac{A(z)}{A(z/\gamma)} = \frac{1-\sum_{i=1}^{p} a_i z^{-i}}{1-\sum_{i=1}^{p} a_i \gamma^i z^{-i}}, \quad 0 \le \gamma \le 1 \qquad \text{식 (3.1)}$$

where, $A(z) = 1-\sum_{i=1}^{p} a_i z^{-i}$

a_i = LPC Coefficient

γ = Weighting Parameter

일반적으로 화자인식을 위해 화자가 발성한 음성신호에서 특징 파라미터를 추출하기 전 고주파성분의 영향을 증가시키기 위해 프리엠퍼시스 필터를 사용한다. 하지만 이 경우 인지적으로 중요하지 않은 성분들까지 그 영향이 증가된다. 또한 저주파 성분과 고주파 성분의 포만트 봉우리 부분의 차이가 심하여 특징벡터 추출 시

오차가 발생한다. 따라서 본 논문에서는 프리엠퍼시스 필터를 통과한 음성신호에 인지 가중 필터를 적용하여 사람의 청각 특성이 반영된 특징벡터를 추출하는 것이다.

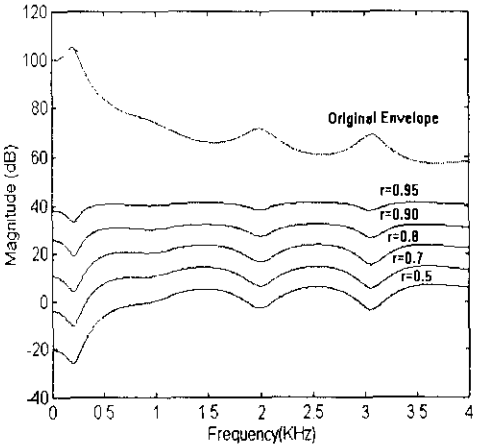

그림 3-1. 인지 가중 필터의 특성.

4. 전체 시스템 구성

4.1 음성 구간 검출

본 논문에서는 음성구간을 검출하기전 안정된 피치 구간을 먼저 찾은 뒤 무성음구간을 포함하기 위해서 일정 범위내에서 입력된 음성을 모두 저장하게 된다. 이렇게 저장된 음성구간에 대해서만 단구간 에너지와 영교차율을 이용하여 음성구간을 검출한다. 그리고 음절사이의 묵음구간이 존재 할 수 있기 때문에 끝점이 검출된 후에도 일정 프레임 동안 다시 음성의 시작점을 단구간 에너지를 이용하여 검출한다. 만일 또다시 시작점이 검출되면 묵음구간이 존재하는 음성으로 간주하고 다시 끝점을 검출하는 과정을 반복한다[2].

4.2 특징 벡터 추출

본 논문에서는 화자의 특징 벡터로 14차 LPC Cepstrum을 사용하였다. 그림 4-1과 같이 먼저 Hamming Window를 사용하여 단구간으로 음성을 나눈다. 음성신호의 고주파항의 영향을 강조시키기 위해 프리엠퍼시스 필터를 사용하였고 인지적인 특성을 증가하기 위해 인지가중필터를 사용하였다. 이렇게 필터를 통과하여 나온 신호로부터 LPC를 구하고 LPC- Cepstrum 변환식을 이용하여 14차 LPC-Cepstrum을 구하였다[2]. 그리고 이렇게 구해진 계수를 귀의 특성을 고려한 mel-frequency scale로 왜곡시켜 특징 파라미터인 14차 mel-cepstrum을 구한다.

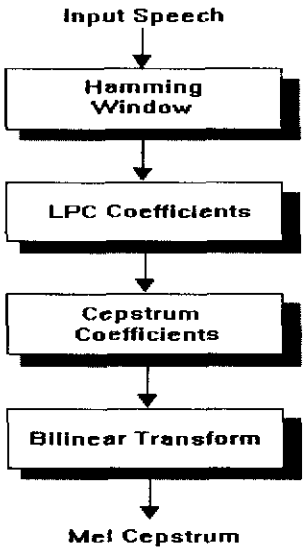

그림 4-1. 특징벡터 추출.

4.3 패턴 정합

본 논문에서 사용한 패턴정합 방법은 Dynamic Program방법인 DTW(Dynamic Time Warping)이다. 이 방법은 시간축을 비선형적으로 왜곡시켜 참조패턴과 시험패턴을 정합하는 방법으로 특징벡터의 시간적 변화를 수용할 수 있고 그 알고리즘은 다음과 같다[2].

단계 1)

$$g_1(c(1)) = d(c(1)) \cdot w(1)$$

단계 2)

$$g_k(c(k)) = \min_{c(k-1)}[g_{k-1}(c(k-1)) + d(c(k)) \cdot w(k)]$$

단계 3)

$$D(A,B) = \frac{1}{N} g_K(c(K))$$

5. 실험 및 결과

본 논문의 알고리즘을 시뮬레이션하기 위해 사용된 실험 장비는 IBM PC에 마이크가 장치된 16-bit A/D변환기를 인터페이스 시켰다. 실험은 일반 실험실 환경에서 10명의 남녀 화자가 각각 본인의 이름을 발성한 음성 시료를 11.025KHz로 샘플링하고 16bit로 양자화하여 사용하였다. 한 프레임의 길이는 300샘플이며, 150샘플씩 Overlap시켜 특징벡터를 추출하였다. 인식을 위한 특징벡터로는 14차 mel-cepstrum을 사용하였다. 기준패턴으로는 20대 남녀 10명이 일주일동안 발성한 음성 4번 발성하게 하여 구성하였고, 200개의 음성데이타를 사용하여 실험하였다. 그리고 사칭자의 효과를 알아보기 위해서 4명으로 하여금 등록된 화자의 음성을 일주일동안 4번씩 발성하게 하였다. 그림 5-1은 본 실험에서 사용한 화자인식 시스템의 전체 블록도이다. 실험결과 인지 가중 필터를 사용하지 않은 경우에 비해 전체 인식률이 1.7%증가하였다.

표 1. 인식율

	FA	FR	전체 인식율
기존의 방법	0.83	4.17	95.0%
제안한 방법	0.56	2.77	96.7%

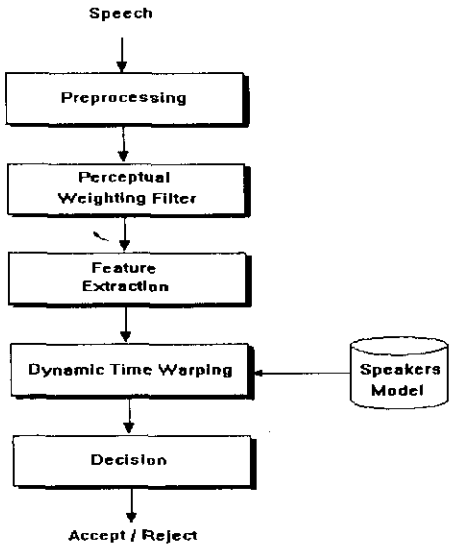

그림 5-1. 인지가중필터를 이용한 화자인식.

6. 결론

일반적으로 화자의 특성은 고차 포만트에서 많이 나타나기 때문에 이를 강조시키기 위해서 프리엠퍼시스 필터를 사용한다. 그러나 이 방법은 모든 고주파 성분을 강조시키기 때문에 인지적인 특성에 불필요한 특성까지도 강조하는 특징을 수반한다. 따라서 본 논문은 이러한 인지적인 측면을 강조하기 위한 인지 가중 필터를 사용함으로써 화자 인식 시스템의 성능 향상에 관한 연구이다. 제안한 방법으로 인식실험을 수행한 결과로는 프리엠퍼시스만을 사용한 기존의 방법보다 1.7%의 향상을 얻을 수 있었다.

7. 참고문헌

[1] S. Furui, *Digital Speech Processing, Synthesis and Recognition*, Marcel Dedder, Inc. ,1992

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

[3] A.M. Kondiz, *Digital Speech*, John wiley & Sons, 1994