Glottal Spectrum 과 화자식별 Parameter와의 상관 관계에 관한 연구

이윤주, 신동성, 배명진
숭실대학교 정보통신공학과

On a Study of Relation Between Glottal Spectrum and Speaker Identification Parameter

Yoonjoo Lee, Dongsung Shin, Mỳungjin Bae
Dept. of Telecommunication, Soongsil University
E-mail : mjbae@saint.soongsil.ac.kr

요약문

음성인식 시스템은 인간의 의사소통 수단인 음성을 기계가 인지할 수 있게 하는 것이다. 이러한 음성인식 알고리즘 개발은 현재 활발히 진행되고 있다. 올바른 음성인식 시스템의 구현을 위해서는 높은 인식률 구현과 적은 처리시간이 요구된다. 또한 인식률 향상을 위해서는 그 구현 알고리즘이 복잡해지고 이에 따라 많은 처리 시간이 요구된다. 본 논문에서는 성문 특성에 따른 Glottal Spectrum에 적응적인 필터계수를 적용하여 인식률 향상을 도모하였다. 제안한 알고리즘을 모의 실험한 결과 전체 인식률이 2% 향상되었다.

1. 서 론

개인이나 특정 단체의 정보의 보안을 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

본 논문은 각 프레임에 성문 특성을 반영을 하여 프리엠퍼스필터 계수를 구하고 원래 음성을 프리엠퍼시스 필터에 통과시킴으로서 스펙트럼 평탄화하는 방법을 제안하였다.

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 종속여부에 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.

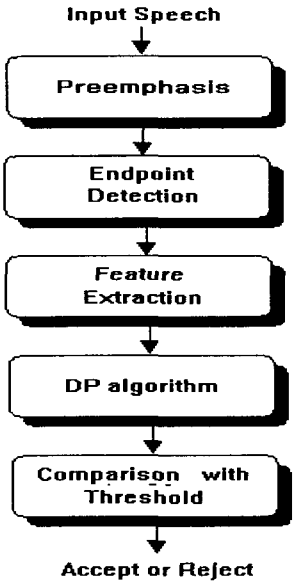

그림 2-1. 일반적인 화자인식 과정

3. 화자인식 시스템 구현

3.1 음성구간 검출

본 논문에서는 음성구간을 검출하기 전 먼저 안정된 피치구간을 찾은 뒤 무성음구간을 포함하기 위해 일정 범위 내에서 입력된 음성을 모두 저장한다.

이렇게 저장된 음성구간에 대해서만 단구간 에너지와 영 교차율을 이용하여 음성구간을 검출한다. 그리고 음절사이의 묵음구간이 존재 할 수 있기 때문에 끝점이 검출된 후에도 일정 프레임(Frame)동안

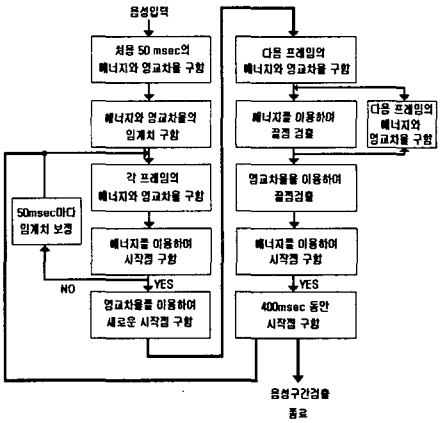

그림 3-1. 음성구간 검출

다시 음성의 시작점을 단구간 에너지를 이용하여 검출한다. 만일 또 다시 시작점이 검출되면 묵음구간이 존재하는 음성발성으로 간주하고 다시 끝점을 검출하는 과정을 반복적으로 수행한다[2].

3.2 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 이 필터에 사용되는 계수값을 각 음성구간마다 적응적으로 구한다. 예를 들어 유성음인 경우 그 계수값이 0.93~0.97으로 구해지지만 무성음인 경우 그 값이 0.2 미만으로 떨어지게 된다.

이렇게 구해진 계수값을 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.

3.3 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 택하였다[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

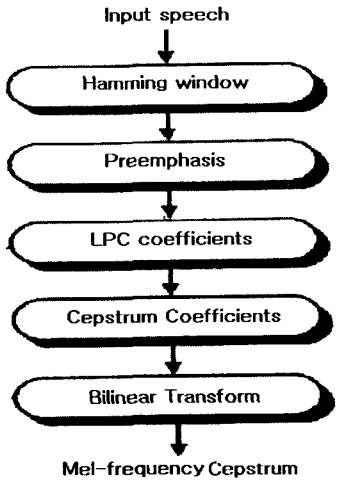

그림 3-2. 특징 벡터 추출

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.

4. 제안한 알고리즘

본 논문에서는 프리엠퍼시스필터 계수를 성문 특성을 이용하여 적응적으로 적용하고 스펙트럼을 평탄화시켜서 인식률을 향상시키는 방법을 제안하였다. 성문특성을 고려한 필터계수를 구하는 방법은 다음과 같다.

단구간 자기상관 함수는 (식5.1)로 표현 가능하다.

$$\Phi_n(i,j)=\sum_{m=0}^{N-1-(i-j)} s_n(m)s_n(m+i-j),\ 1\le i\le p, 0\le j\le p \quad (식4.1)$$

여기서 $R_n(j)=\sum_{m=0}^{N-1-j} s_n(m)s_n(m+j)$ (식4.2)

$$\sum_{j=1}^{p} a_j\Phi_n(i,j)=\Phi_n(i,0),\ \text{for } i=1,\ldots,p \quad (식4.3)$$

자기상관법(Auto-correlation Method)을 이용하여 (식5.3) 를 풀면 다음과 같이 표현된다.

$$\begin{vmatrix} R_n(0) & R_n(1) & . & R_n(p-1) \\ R_n(1) & . & . & R_n(p-2) \\ . & . & . & . \\ R_n(p-1) & . & . & R_n(0) \end{vmatrix} \begin{vmatrix} \alpha_1 \\ \alpha_2 \\ . \\ \alpha_p \end{vmatrix} = \begin{vmatrix} R_n(1) \\ R_n(2) \\ . \\ R_n(p) \end{vmatrix} \quad (식4.4)$$

p=1에 대하여 위의 식을 정리하면 다음과 같은 식으로 표현 가능하다.

$$\alpha_1=\frac{R_n(1)}{R_n(0)} \quad (식4.5)$$

그림 6-1, 그림 6-2는 음성신호와 자기상관법을 이용하여 측정한 음성신호의 기울기를 나타낸 것이다.

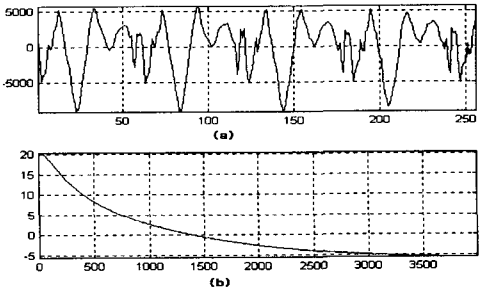

그림 4-1. 유성음 (a)음성신호, (b)측정된 기울기

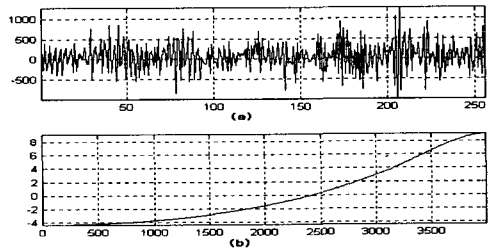

그림 4-2. 무성음 (a)음성신호, (b)측정된 기울기

위와같은 특징을 이용하여 필터 계수를 구하였고 인식 알고리즘에 적용하였다.

5. 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. 실험결과 제안한 방법이 전체 표준패턴과의 비교를 수행한 방법에 비해 전체 인식률이 2% 향상되

었고 인식시간은 5% 증가하였다.

표 5-1. 인식률

	False Accept	False Reject	Recognition Accuracy
기존의 방법	0.83	4.17	94.0%
제안한 방법	0.33	2.47	96.0%

6. 결론

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.

본 논문에서는 프리엠퍼시스필터 계수를 성문 특성을 이용하여 적응적으로 적용하고 스펙트럼을 평탄화시켜서 인식률을 향상시키는 방법을 제안하였다. 실험결과 기존의 방법에 비해 5%의 시간은 증가하였으나전체 인식률이 2%향상되었다.

7. 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.

[5] L R. Rabiner, R.W Schafer, " Digital Processing of Speech Signal", Prentice Hall, 1978.

[6] 배명진,"디지탈 음성분석", 동영출판사, 1998. 4.

[7] Oppenheim, Schafer,"Discrete Time Signal Processing", Prentice Hall, 1989.

[8] Emanuel C. Ifeachor,"Discrete Time Signal Processing", Addison Wesley,1993.

[9] 오영환, "음성언어정보처리",홍릉과학출판사,1998.

[10] Douglas O, shaughnessy,"Speech Communication",IEEE Press,1996.

[11] A. M. Kondoz,"Digital Speech", John Wiley & Sons Ltd, 1994.

[12] 배명진,"디지탈 음성합성", 동영출판사, 1998. 2.

[13] 민소연, 강은영, 배명진, "성문특성이 제거된 성도특성에 관한 연구", 대한전자공학회, 추계 종합학술대회, 2000년 11월 25일.