Cepstrum 계수를 이용한 화자인식

*박 남 규 강 철 호
광운대학교 전자통신공학과

Speaker verification using cepstrum coefficients

Nam Gyu Park Chool Ho Kang
Dept. of Electronic Comm., Kwang Woon Univ.

Abstract

In this paper, zero crossing rate, energy, pitch gain, first formant and 10 cepstrum coefficients have been used to extract the features of the speakers.

Fifteen parameters are sorted into three kinds of group. First one includes all of them, second one with only cepstrum coefficients and the third one with five parameters except cepstrum coefficients. FR(False Reject) and FA(False Accept) ratios are calculated for each group.

1. 서 론

사람이 발음하는 음성에는 의사 전달의 정보 뿐만 아니라 말하는 사람이 누구인가하는 화자의 고유한 특성을 포함하고 있다[1][2]. 이러한 고유한 특성을 추출하여 화자인식에 사용하였다.

사람이 발성하는 음성은 개개인의 발성 지속시간과 발성시간의 차이때문에 시간축이 비선형적으로 변동하며 이러한 시간축의 변동제거와 시간축의 정규화는 DP(Dynamic Programing) 기법에 의해 거의 완벽한 인식 수행을 얻을 수 있다. 그러나 본 논문에서는 각 화자의 발음때에 발성시간의 차를 고려함에 있어서 발성시간도 일정한 습관에 의해 결정된다는 점을 고려하여 비교적 간단한 LTW(Linear Time Warping)을 사용하였다.

LPC는 각기 다른 음성이 성도의 주파수특성을 변화시켜서 발생되는 점을 고려하여 디지털 부호화한 것으로서 이것은 사람의 성도를 시간에 따라 변화하는 계수를 갖는 선형 여파기로 모델화한다. 본 논문에서는 기준 화자 3명에 대해서 자기상관법(autocorrelation method)을 사용하여 10개의 LPC 계수를 추출했으며 이를 변환한 cepstrum계수가 다른 파라미터보다 화자 인식에 더 유용함을 보여준다.

2. 실험 방법

(1) 음성 신호의 분석

본 실험에 사용된 데이타는 각 3명의 기준 화자가 15번 발음한 15개의 음성중 3개는 표준패턴의 작성에 이용하였고 표준패턴 작성에 이용한 3개의 데이타를 포함한 15개의 데이타를 FR에 사용하였다. 또한 기준화자가 아닌 화자가 발음한 72개의 데이타를 FA에 사용하였다.

각 화자가 1.28초 동안에 발음한 음성을 4KHz LPF를 거친후 10KHz로 샘플링 하였다. 384점을 한 프레임으로 하여 98프레임을 분석했다. cepstrum 계수의 추출시 preemphasis 처리로 $\beta = 0.9$ 로 하였다.

(2) 파라미터 추출

본 실험에서는 우선 유성음과 무성음을 구별하여 유성음에 한해서만 위에서 언급한 15개의 파라미터를 추출하였다. 유성음과 무성음의 검출은 영교차율에 의해 유성음과 무성음을 구별하였다.

다음의 그림 1은 화자 확인의 블록도를 나타내고 있다.

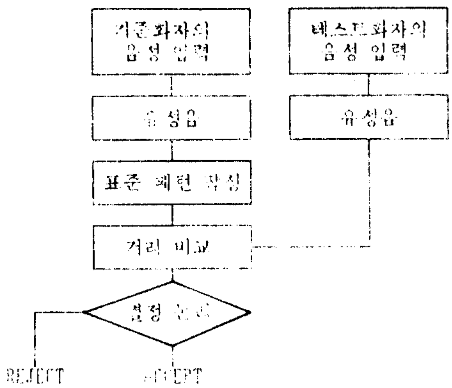


그림 1. 화자 확인 블럭도.

영교차율 (Zero Crossing Rate)은 다음과 같이 정의된다.

$$Zn = \sum_{m} | sgn(X(m)) - sgn(X(m-1)) | \qquad (1)$$

한 프레임에서의 영교차율이 0 ~ 98 회이면 유성음으로, 98회 이상이면 무성음으로 간주하였다[3].

LPC계수를 변환한 cepstrum계수가 화자인식에 더 유용하며[4], LPC계수는 autocorrelation 방법으로 10차까지 구했다. cepstrum은 주파수 영역으로 변환한 후 대수를 취하여 역변환한 것을 의미한다. cepstrum 계수 c_n는 LPC계수 a_k로부터 식(2)에 의한 반복적인 방법으로 구할 수 있다.

$$c_1 = a_1$$

$$c_n = \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) a_k c_{n-k} + a_n, \quad 1 < n \le p \qquad (2)$$

(3) 표준 패턴 작성

r인 기준 화자가 표준 패턴의 작성에 기여한 발음을 Rr1, Rr2, ..., RrN 이라고 하면 각 발음마다 검출되어진 유성음의 프레임수를 f(Rr1), f(Rr2), ..., f(RrN)이라고 할 수 있다. 화자 r의 표준 [illegible] 다음과 같다.

$$Fr = med\{ f(Rr1), f(Rr2), \ldots, f(RrN) \} \qquad (3)$$

여기서 med 는 [illegible] 중간값을 나타낸다.

기준화자 r의 표준 패턴의 유성음 구간 프레임에서 15개의 파라미터중 k 번째 파라미터를 구하는 경우 r화자의 표준 패턴의 j번째 유성음 프레임에서의 값 $R_r^{(k)}(j)$ 는 각 발음 $R_{ri}^{(k)}$를 linear warping 하여 식 (4)로 구할 수 있으며 warping함수 w(j,i)는 식 (5)로 구할 수 있다.

$$R_r^{(k)}(j) = \frac{1}{N}\sum_{i=1}^{N} R_{ri}^{(k)}(w(j,i)) \qquad (4)$$

$$w(j,i) = \frac{f(R_{ri}) - 1}{F_r - 1} \times (j-1) + 1 \qquad (5)$$

위 식에 의해서 작성된 표준 패턴과 이에 기여한 음성들과의 평균거리 $\bar{D}_r^{(k)}$는 다음의 식(6)로 구해지며 표준 편차는 식 (7)로 구할 수 있다.

$$\bar{D}_r^{(k)} = \left[\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{F_r}\sum_{j=1}^{F_r}\left(R_r^{(k)}(j) - R_{ri}^{(k)}(w(j,i))\right)^2\right]\right]^{1/2} \qquad (6)$$

$$\sigma_r^{(k)} = \left[\frac{1}{N}\sum_{i=1}^{N}\left(\bar{D}_r^{(k)} - D_{ri}^{(k)}\right)^2\right]^{1/2} \qquad (7)$$

(4) 테스트 패턴과의 거리 비교

입력된 테스트 화자의 음성이 T 일때 표준 패턴과 테스트 패턴과의 유클리드 거리 $D_r^{(k)}$는 아래의 식 (8)로 구할 수 있다.

$$D_r^{(k)} = \left[\frac{1}{F_r}\sum_{j=1}^{F_r}\left(R_r^{(k)}(j) - T(w(j))\right)\right]^{1/2} \qquad (8)$$

$$w(j) = \frac{f(T) - 1}{F_r - 1} \times (j-1) + 1 \qquad (9)$$

(5) 결정 논리

이 실험에서 사용한 결정 논리는 기준 화자의 표준패턴 평균거리와 테스트 패턴과의 거리 $D_r^{(k)}$가 아래의 식(10)보다 클 경우에 기준화자와 테스트 화자가 서로 다른 것으로 판단하였다.

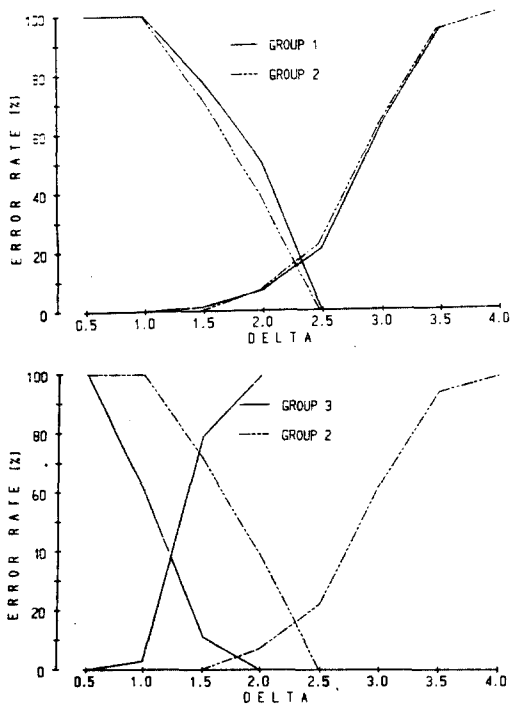
$$\text{IF } D_r^{(k)} > \bar{D}_r^{(k)} + \Delta \cdot \sigma_r^{(k)} \text{ THEN REJECT} \qquad (10)$$
$$\text{ELSE ACCEPT}$$

여기에서 Δ는 화자내 변동의 범위로 설정하기 위한 계수이다.

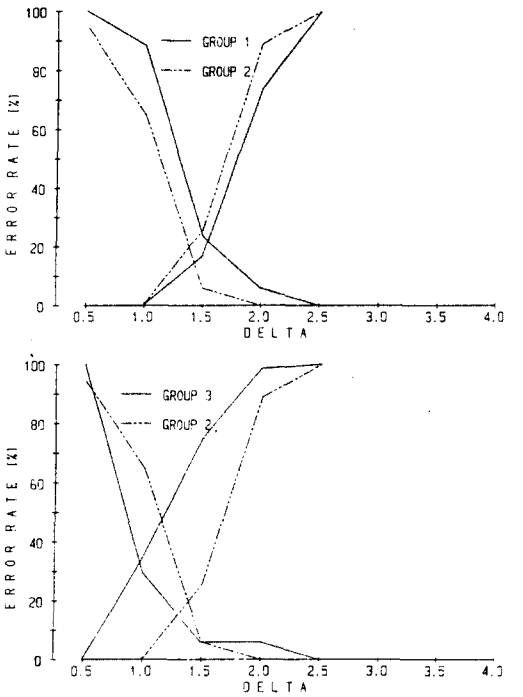
3. 실험 결과

본 논문에서 행한 실험은 "통신 연구실" 이라는 유성음과 무성음이 포함된 일반적인 음성에 대해서 우선 cepstrum계수를 제외한 나머지 5개의 계수 (영교차율, 에너지, 피치, 이득, 제 1 포만트)를 포함한 그룹 3과 cepstrum 계수가 포함된 그룹 2, 그리고 15개의 계수 (영교차율, 에너지, 피치, 이득, 제 1 포만트, cepstrum 계수)를 모두 포함한 그룹 1으로 구분하여 각각에 대한 동일성을 요구한 화자가 기준 화자일때 이 기준 화자를 거부하는 FA(False Reject) 와 기준 화자로 잘못 인식하는 FR(False Accept)로 성능을 나타내었다.

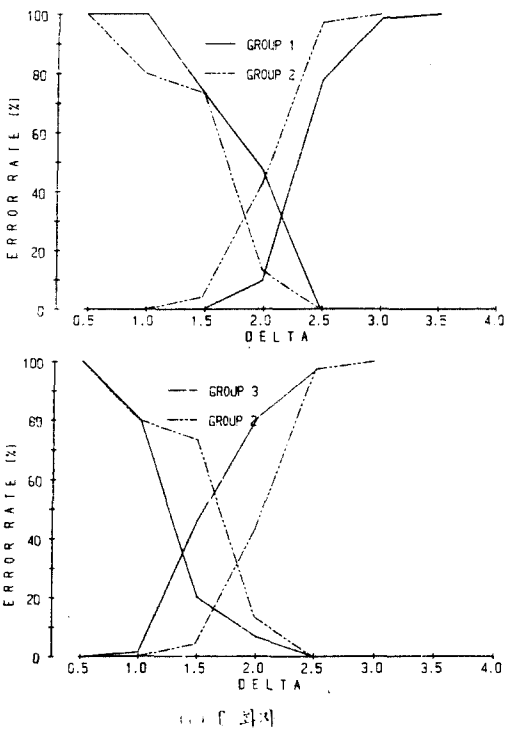
실험 결과 cepstrum 계수만을 사용한 것과 cepstrum 계수를 포함하는 15 개의 파라미터를 다 사용한 것과는 성능의 차이를 거의 볼 수 없었으며 cepstrum 계수만을 사용했을때의 FR과 FA가 교차하는 점에서의 오류율이 A화자의 경우 16% , B화자의 경우 31% ,C화자의 경우 20% 이다. 다음의 그림 2은 각화자에 대한 FR과 FA의 오류율을 나타낸 그림이다.



(a) B 화자



(b) E 화자



(c) F 화자

그림 2. 각 화자의 그룹별 FR과 FA의 오류율 비교.

4. 결론

FR과 FR이 교차하는 점에서의 오류율은 cepstrum계수만 사용할 경우 15개의 계수(영교차율, 에너지, 피치, 아독, 제 1 포만트, cepstrum계수)를 전부 사용하는 그룹과는 오류율이 비슷하나, cepstrum 계수를 제외한 5개의 파라미터를 사용한 그룹보다 10개의 cepstrum 계수만을 사용한 그룹의 오류율이 더 낮다는 점을 확인하였다. 즉 15개의 파라미터 전부를 사용하지않고 cepstrum 계수만을 사용하여도 같은 인식율을 보였다. 이는 cepstrum 계수로 변환하기 전의 LPC계수가 음성의 주파수 특성을 잘 나타내므로 유사한 음성의 패턴에 잘 맞으며 계수의 수가 입력 음성 데이타에 비하여 비교적 적으므로 다루기가 용이하기 때문인 것으로 생각된다. 또한 이 cepstrum 계수에서 인식의 정도에 비례한 중요도에 따라 각 계수마다 다른 가중치를 준다면 더 좋은 결과를 얻으리라 생각되어지며 이에 대한 연구가 진행중이다. 또한 화자인식에 사용되는 파라미터의 갯수를 줄인다면 실시간으로 음성 인식을 하는데 소요되는 시간을 줄일 수 있을 것이라 생각된다.

5. 참고 문헌

1. J.J. Wolf, " Efficient acoustic parameters for speaker recognition", JASA, vol.51, pt.2 , pp.2044 -2056, June, 1972.
2. B.S. atal, "automatic speaker recognition based on pitch contours", JASA, vol.52, pp.1687 1697, Dec, 1972.
3. L.R.Rabiner and R.W.Schafer, Digital Processing of Speech Signals, Prentice Hall, New Jersey, 1978.
4. S. Furui, " Cepstral analysis technique for automatic speaker verification", IEEE Trans. on ASSP, vol.29, pp.254-272, 1981.