2002년도 대한전자공학회/대한전기학회 시스템 및 제어분야 합동 추계학술대회 논문집(2002.11.30)

강인한 정합과정을 이용한 텍스트 종속 화자인식에 관한 연구

이 한 구, 이 기 성
홍익대학교 전기정보제어공학과

A study on the text-dependent speaker recognition system Using a robust matching process.

Hanku Lee, Keeseong Lee
Dept. of Electrical, Information and Control, Hongik University
E-mail : ta219@hanmail.net

Abstract

A text-dependent speaker recognition system using a robust matching process is studied. The feature histogram of LPC cepstral coefficients for matching is used. The matching process uses mixture network with penalty scores. Using probability and shape comparison of two feature histograms, similarity values are obtained. The experiment results will be shown to show the effectiveness of the proposed algorithm.

1. 서론

화자인식 시스템은 현대사회에서 정보를 관리, 운용하는데 있어서, 중요한 문제로 떠오르고 있는 보안문제의 한 해결책이 될 것이다. 음성의 장점은 사용자가 친근감을 느낀다는 점이다. 그러나 사용자의 기분, 환경 등에 따라서 많은 차이가 나기 때문에 신뢰성이 비교적 낮다는 문제점이 있다[1]. 이러한 문제점을 해결하기 위해서 화자에서 좀 더 강인한 특징을 추출하는 방법, 효율적인 정합 방법에 대한 연구가 활발히 진행되고 있고 다른 인식 시스템과 결합해서 신뢰성을 높이는 연구도 많이 진행되고 있다.

본 논문은 화자인식 시스템에서 가장 중요한 인식률의 향상을 목적으로 하고 있다. 화자의 음성에서 추출한 특징값들의 분포형태를 특징 히스토그램으로 나타내고, 이것을 화자의 혼합 확률(Mixture)이 포함되어 있는 혼합망(Mixture Network)의 입력으로 사용한다. 혼합망에서는 특징 히스토그램과 혼합망의 각 노드의 평균값으로 이루어진 등록된 특징 히스토그램과의 형태를 비교해서 다른 정도에 따라서 벌점(Penalty Score)을 주고, 확률을 이용한 비교를 통해 높은 인식률을 보여 준다..

2. 화자인식 시스템

화자인식은 화자의 음성에 의해 화자를 인식하는 것을 말한다. 화자인식은 크게 화자확인과 화자식별로 나누어진다. 화자확인은 발성된 음성이 원하는 화자인지 아닌지를 구분해내는 것으로 의뢰인에 의해 초기 등록이 요구된다. 이 방법은 발성된 음성과 확인을 바라는 의뢰인 이름의 입력으로 시작되어 확인과정을 거친 후 의뢰인으로서 수락할 것인지, 거부할 것인지를 결정하게 된다. 화자식별은 등록된 N명의 사람들 중 가장 비슷한 사람을 찾아내는 과정이다. 화자확인은 출입통제 시스템이나 원격지 데이터베이스 검색에 이용할 수 있고, 화자식별 기술은 범인식별, 자동 회의록 작성 등에 이용될 수 있다.

화자인식은 인식대상이 되는 음성의 발성방법에 따라 문장 종속형과 문장 독립형으로 나뉘어 진다. 화자가 발성한 문장이 고정되어 사용하는 화자인식 시스템을

문장 종속형이라 하고, 문장이 정해져 있지 않고, 자유롭게 발성하는 경우를 문장 독립형이라 한다. 문장 종속형인 경우 음운에 기반을 둔 공통적 특징의 개인 차이를 평가하게 되므로 미리 저장된 표준패턴과 비교를 수행하는 음성인식과 거의 동일한 방법을 사용한다. 또한 선정 어휘에 따라 인식률이 영향을 받으므로 화자의 특성이 잘 나타나는 모음, 비음 등의 음운이 균형을 이룬 어휘 집합의 선택이 필요하다. 이에 비해 문장 독립형의 경우 임의로 자유롭게 발성된 음성 신호로부터 음운정보를 제거한 화자정보만을 사용해야 하므로 전자보다 어려운 문제이다. 이 방법은 많은 음성 자료가 필요하며, 음성으로부터 음운정보를 제거하기 위해 통계분석을 하거나, 음소단위로 분리하여 각 음소와 관계되는 화자정보를 얻어내는 방법을 사용한다. 본 논문에서는 문장종속형인 화자확인 시스템을 구축하였다.

3. 음성의 처리와 화자인식 연구

음성신호는 긴 시간으로 보았을 때, 시변신호이지만 음성 발성기관의 특성상 시간에 따른 변화는 매우 느리기 때문에, 짧은 음성신호도 안정된 신호로 간주할 수 있고, 따라서 음성신호처리는 단구간(short-term) 해석을 기본단위로 한다[2].

3.1 음성부분의 추출

음성인식을 위해서는 먼저 해당음성이 발음된 부분과 발음되지 않은 부분을 구별해내야 한다. 음성신호는 5~30ms내의 시간에서는 안정된 신호특성을 지니고 있다. 따라서 단구간별로 해석이 가능한데, 음성부분의 추출을 위해서는 평균에너지, 영교차율 등의 방법이 사용된다. 음성부분의 추출은 먼저 평균 에너지법을 사용하여 유성음 부분을 찾아내고 영교차율을 이용하여 무성음 부분을 찾아냄으로써, 최종적으로 발음된 부분을 추출한다.

3.2 음성의 특징추출

그 동안 일반적인 음성인식 분야에서 널리 쓰이는 음성 샘플의 특징으로는 음성의 피치(pitch), 필터뱅크별 FFT 계수, 웨이블릿 계수, 단구간 에너지, 편자기(PARCOR) 계수, LPC(Linear Predictive Coding) 계수, LPC cepstrum 계수 등의 방법이 있다. 이 중에서 LPC cepstrum 계수가 가장 강인한 성능을 나타내는 것으로 최근의 많은 연구에서 각광을 받아왔다. 본 논문에서 사용되는 LPC 과정과 이에 따른 LPC cepstrum 계수 추출에 대해서 설명한다.

음성에서 p차 LPC cepstrum 계수를 추출하는 방법은 다음과 같다. 11.025kHz로 저장한 음성신호를 한 프레임 당 300샘플(N=300)로 나누어 200샘플(M=100)씩 오버랩을 시키면서 해밍 윈도우(Hamming window)를 곱해준다. l번째 프레임에 대하여 해밍 윈도우로 처리된 음성 정보는 p차 만큼의 자기상관계수(autocorrelation coefficients)를 구해서 Durbin 알고리즘을 통해 LPC 계수를 추출한다. 이러한 과정을 통해서 구한 p차의 LPC 계수는 일정치 않은 값의 범위를 갖는다. 따라서 음성인식에 안정적인 LPC Cepstrum 계수로 변환해서 사용한다[3].

개인에 따라 특징값을 추출한 후 신분 확인을 위하여 이전에 등록된 특징값과 새로 입력된 특징값과의 유사도를 측정한다. 이러한 과정을 정합(Matching)이라 하며 기존의 연구에서도 정확성 및 효율성을 동시에 만족시키기 위해 노력해왔다.

3.3 신경회로망을 이용한 방법

(1)SOFM(Self Organizing Feature Map)과 LVQ (Learning Vector Quantization)을 이용하는 방법

SOFM으로 새로운 2차원 패턴을 생성한 후, LVQ로 인식하는 방법이다. 비교적 간단한 방법이지만 노이즈에 매우 약한 단점이 있다. 따라서 사용자의 수가 적고 비교적 인식률이 낮아도 되는 간략한 정합을 시행하는 목적에 국한하여 이용되고 있다[4].

(2) FSCL(Frequency Sensitive Competitive Learning) 또는 SOFM을 이용한 방법

입력으로 특징벡터가 신경회로망에 들어가면 승자 뉴런들이 발생한다. 그 다음 각 뉴런이 승자 뉴런이 된 횟수를 구하고 가중치를 곱해 준다. 이 값들을 이용해서 소속도를 구한다. 소속도 값에 의해서 인식 결과가 정해진다[5].

(3) HMM(Hidden Markov Model)을 이용한 정합

확률적인 접근 방법으로서, 음성의 동적인 특성까지도 잘 나타내기 때문에 매우 인식률이 높은 정합법이다. 각 화자에 대한 HMM파라메터를 구해서 저장해 놓은 후, 입력 음성의 코드워드 벡터가 발생할 최대 확률값을 구해서 결정하는 방법이다. 현재 가장 일반적으로 사용되는 정합 방법이다[4].

(4) 동적 시간 정합(Dynamic Time Warping)을 이용한 방법

동일인이 동일 시간 단어를 발성해도 특별히 훈련받은 사람이 아니면 발성시 마다 단어의 시간적 길이가

변화한다. 이를 표준 패턴과 단순히 비교하면 시간축이 고르지 않기 때문에 오류나 인식 불능(reject)이 생기게 된다. 이 영향을 제거하기 위한 방법이 동적 시간 정합법이다. 길이가 서로 다른 두 개의 자료에서 최적의 정합 경로를 찾아 두 자료를 서로 비교할 수 있는 방법이다.

특징벡터를 코드북을 이용해서 코드워드 벡터로 양자화 한다. 저장되어 있는 음성 코드워드 벡터와 입력 음성 코드워드 벡터의 길이를 동적 시간 정합 방법을 이용해서 맞춘 다음 유사도를 비교해서 인식 결과를 정한다.

4. 제안하는 화자인식 정합 알고리즘

본 논문에서 제안하는 화자 인식 정합 방법은 기존의 고립단어의 경우와 연속음의 경우 각각 크기 32, 64의 코드북을 사용한 것과는 다르게 비지도 벡터 양자화(Unsupervised Vector Quantization)를 이용하여 화자 마다 각각 크기가 30~60사이인 최적의 코드북을 작성한다. 그것을 이용해서 특징 벡터를 코드워드 벡터로 바꾸고, 코드워드 벡터는 다시 특징 히스토그램으로 나타낸다. 이 특징 히스토그램은 혼합망의 입력으로 들어가고 출력으로 유사도 값이 나온다[6]. 이 유사도 값은 임계치를 이용한 판단논리로 수락 또는 거부를 결정한다.

화자마다 코드북을 가지고 있으면 새로운 화자 등록시 다시 전체 코드북을 작성할 필요없이 새로운 화자의 코드북만을 작성하면 되기 때문에 확장성이 좋다.

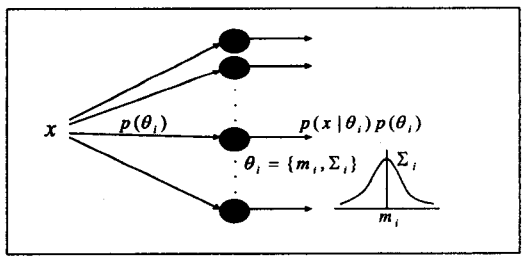

그림 1. 기존의 혼합망

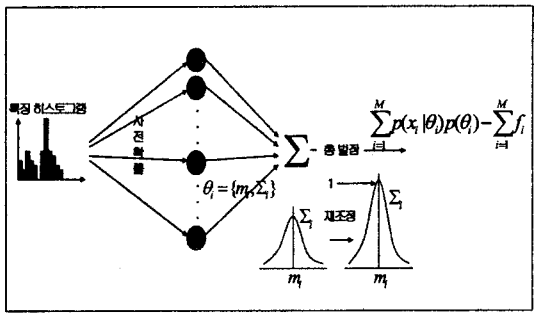

그림 2. 제안하는 혼합망

10회 등록한 음성 샘플을 이용해서 히스토그램의 평균과 분산을 구하고, 각 노드의 가우시안 분포를 구한다.

$p(\theta_i)$: 사전확률(노드 i 가 발생할 확률)

$$p(x|\theta_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} e^{-(1/2)(x-m)^T \Sigma_i^{-1}(x-m_i)} \quad (1)$$

$$p(X|\Theta) = \sum_{i=1}^{M} p(x_i|\theta_i)p(\theta_i) \quad (2)$$

기존의 혼합망의 경우 입력으로 하나의 값만이 들어가지만 제안한 화자인식 시스템은 특징 히스토그램이 입력으로 들어간다. 즉, 특징 히스토그램의 각 값들이 입력이고, 모든 노드에 입력이 동시에 들어간다. 그림 2 에서는 새로운 혼합망 구조를 보여준다. 각 노드의 확률은 가우시안 분포를 이용하고, 크기는 평균값에서 확률값이 1을 갖게 모두 재조정 하고, 나머지 값도 그에 맞추어 크기를 재조정한다. 특징 히스토그램의 각각의 값과 혼합망의 각 노드의 평균값이 완전히 똑같을 때는 1의 값을 얻게 된다. 입력으로 특징 히스토그램이 들어오면, 히스토그램의 각 값들은 혼합망의 각 노드로 분배되어 들어간다. 노드에서 나오는 결과값은 노드가 발생할 확률. 즉, 사전확률과 발생한 노드가 가지는 가우시안 분포에서 분배되어 들어온 값이 발생할 확률을 곱한 것이다. 혼합망의 결과값은 각 노드에서 발생한 확률을 모두 더한 것이 된다.

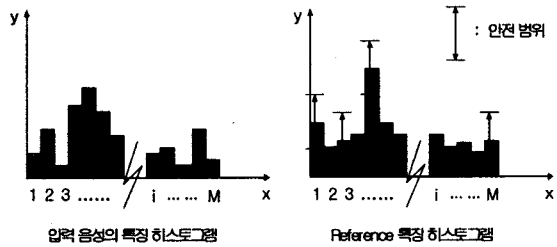

그림 3. 특징 히스토그램간의 형태 비교

특징 히스토그램 x축의 1에서부터 M은 mixture

network의 노드(node)번호를 나타낸다. 벌점은 reference와 입력음성의 차이가 안전 범위 내에 있을 경우 벌점은 0이다. 음성은 발음할 때마다 변화하는 특성이 있기 때문에 히스토그램의 비교에서도 변화의 정도를 인정하는 안전범위를 정하였다. 이 범위내의 변화는 잡음이나 개인의 특징으로 간주한다. 그러나 안전 범위를 벗어날 경우 잡음이나 동일한 화자의 특징이 아닌 타인의 특징으로 간주하고 정도에 따라서 벌점이 주어진다. mixture network에서 노드 i 의 값과 입력 음성에서 i 의 값의 차이를 D_i 라고 할 때, 벌점 f_i 는 다음과 같이 구할 수 있다.

$|D_i| \leq 4$(안전 범위) : $f_i = 0.0$ (3)

$4 < |D_i| \leq 8$: $f_i = 0.1$ (4)

$8 < |D_i| \leq 10$: $f_i = 0.2$ (5)

$10 < |D_i| \leq 11$: $f_i = 0.5$ (6)

$11 < |D_i|$: $f_i = 1.0$ (7)

f_1부터 f_M까지의 값을 모두 더한 후, 그 값을 이용해서 식 (8)로 최종적인 유사도 값, S_{total}을 구한다.

$$S_{total} = p(X|\Theta) - \sum_{i=1}^{M} f_i \quad (8)$$

S_{total}은 마지막으로 임계값을 이용해서 수락 또는 거부를 결정한다. 임계값은 실험을 거쳐 최적의 값으로 정하게 된다. 실험으로부터 임계값은 0.85로 설정하였다.

화자	FRR	FAR
A	0/13	0/117
B	7/13	0/117
C	1/13	0/117
D	0/13	0/117
E	3/13	0/117
F	7/13	0/117
G	4/13	0/117
H	3/13	0/117
I	2/13	0/117
J	3/13	5/117
결과	30/130	5/1170

표 1. 제안한 정합과정을 이용한 실험

여기서 FRR는 False Reject Rate과 FAR은 False Accept Rate을 나타낸다.

5. 결론

본 연구는 화자 인식 시스템에서 인식률을 떨어뜨리는 가장 큰 요인인 발음할 때마다 달라지는 음성의 특성에 강인한 정합 알고리즘 개발을 목적으로 하고 있다.

음성에서 LPC cepstrum 계수를 추출한 후, 비지도 VQ로 만들어진 코드북을 이용해서 코드워드 열을 구하고 특징 히스토그램을 만든다. 노이즈가 포함되고, 음성의 길이나 발음이 조금만 달라져도 전체적으로는 큰 영향을 끼치지 않는 음성의 전체적인 특징을 나타내는 특징 히스토그램을 입력으로 사용한다. 정합 방법으로는 형태비교에 적합하게 변형된 mixture network를 이용함으로써 높은 인식률을 유지할 수 있다. 음성의 동적인 특성을 이용해서 개인의 독특한 특성을 추출해 내고, 또한 변화가 심한 음성을 인식하기 위해서 지금보다 좀 더 정확한 클러스터링을 할 수 있다면 더욱 향상된 인식률을 기대할 수 있을 것이다.

참고문헌

[1] Benjamin Miller, "Vital Signs of Identity," *IEEE Spectrum*, Feb. 1994.

[2] John R. Deller, Discrete-Time Processing of Speech Signals, IEEE Press, New York, 1993.

[3] L. Rabiner, Biing-Hwang Juang, Fundamentals of speech recognition, Prentice Hall, New Jersey, 1993.

[4] 박준하, "HMM과 유전 알고리즘을 이용한 한국어 음성의 음소단위 인식," 석사학위청구논문, 홍익대학교 전기제어공학과, 1997년 12월.

[5] 정준원, "경쟁학습 신경회로망과 유전자 알고리즘을 이용한 텍스트종속 화자검증," 석사학위청구논문, 홍익대학교 전기제어공학과, 1996년 12월.

[6] Hujun Yin and Nigel M.Allinson, "Self-Organizing Mixture Networks for Probability Density Estimation," *IEEE Trans. on Neural Net.* Vol. 12, No. 2, March 2001.