1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 1(s) 호

출입자 판별을 위한 문맥 제시형 화자인식

서광석, 신유식, 김종교

전북대학교 전자공학과

The Text-Prompt Speaker Recognition for Customer Discrimination

Kwang-Seok Seo, You-Shik Shin, Chong-Kyo Kim

Dept. of Electronic Eng., Chonbuk National Univ.

요 약

본 연구에서는 문맥 종속 또는 문맥 독립형 화자 인식에서의 단점을 개선하는 방법으로 문맥 제시형 화자 인식을 수행하였다. 문맥 종속형 화자 인식은 제한된 문장이나 단어를 발성하여 출입 판별을 하는 방식으로 구현하기는 쉬우나 사칭자가 사용자의 목소리를 흉내낼 수 있으며[1], 문맥 독립형 화자 인식은 임의의 대화 문장이나 대화를 사용자에게 유도하여 일정 시간 동안 녹음한 후에 이를 이용하여 사칭자가 접근을 허가받을 수 있다는 단점이 있다. 또한 문맥 독립형 화자 인식에서는 접근 허가를 받기까지 많은 학습 시간이 필요하며 학습 시간이 적을 경우에 상당한 인식률의 저하가 발생된다. 문맥 제시형 화자 인식은 랜덤하게 제시된 단어만을 화자가 발성함으로써 특정한 문장이나 단어의 배열을 미리 녹음했다가 재생하는 방법을 배제할 수 있을 뿐만 아니라 동시에 학습을 위한 많은 시간을 소모하지 않는다는 장점이 있다. 본 논문에서는 화자로 하여금 랜덤하게 제시된 여러 개의 단어들을 순서적으로 발성하도록 하여, 발성 단어를 인식한 후에 인식된 단어를 통하여 화자를 판별하는 방법을 사용하였다.

1. 서 론

현대 정보화 사회로의 발달로 인한 소비자의 욕구에 맞추어 다양한 서비스가 개발되고 있다. 이러한 서비스에는 은행의 자동 현금지급서비스, 전화 쇼핑서비스, 정보검색서비스 등이 널리 이용되고 있으며, 본인 또는 허가를 얻은 사람만이 접근해서 서비스를 이용하도록 되어 있다. 그러나 만약에 허가를 얻지 않은 사람이 접근해서 사용을 한다면 심각한 사회문제를 발생시킬 수 있다. 따라서 개인의 신분확인은 보안을 필요로 하는 곳에 필수적이다. 과거에는 사람의 신원을 확인하는데 신분증, ID카드, 도장, 서명 등으로 신원을 확인하였으나 이러한 경우 분실이나 복사 또는 사본을 만들어서 사용하는 문제점이 발생하였다. 따라서 인간의 특성을 이용하여 신원을 확인하고자하는 연구가 진행되고 있다. 이러한 연구 분야 중에서 인간의 특성인 음성을 이용하여 신원을 확인하는 연구가 바로 화자인식분야이다. 이는 화자의 음성신호에서 발생되는 개인의 특성들을 추출하여 화자를 인식한다. 음성 신호로부터 말하는 사람이 누구인지를 판단하는 화자인식(speaker recognition)기술은 task와 성격에 따라 화자식별(speak-er identification)과 화자확인(speaker verification)으로 나눌 수 있다. 여기서 화자 식별이란 등록된 화자들 중 발화자가 누구인가를 알아내는 것이고, 화자확인은 특정인이라고 자칭하는 인식 대상이 본인인지 여부를 알아내는 과정을 의미한다.

화자확인은 화자의 확인요구를 수행하기 위한 열쇠로서 음성을 사용하는 여러 가지종류의 서비스에 응용된다. 화자인식 방법은 발화내용이 미리 정해진 경우를 문맥 종속형(text dependent) 화자인식이라 하고 임의의 단어 또는 문장을 대상으로 하는 경우를 문맥 독립형(text independent) 화자인식이라 부른다.[2] 문맥독립

형 방법은 문맥종속형에 비해 더 높은 음향학적 변이를 갖고 있기 때문에 더많은 훈련 데이터가 요구된다.

화자인식에서 가장 어려운 점은 시간상에서의 변화이다. 이는 화자의 인식정도에 중요한 영향을 준다. 비록 이것이 음성인식에서는 문제가 되지 않을지라도 화자인식에서는 이것은 두 가지 중요한 어려운 문제이다. 먼저 각각의 화자의 참조패턴이 인식을 위하여 사용되고 반복적으로 이용된다. 두 번째 개인의 특성이 음운정보보다 더 있어야한다. 화자간의 변화는 상호음소변화보다 적다.

2. 특징 추출

본 논문에서 사용한 파라메터는 LPC 캡스트럼이며 LPC 캡스트럼은 다음과 같은 과정으로 처리될 수 있다. 우선 음성신호는 해밍창(Hamming Window)에 통과되고 고주파 항을 강조시키는 프리엠퍼시스 필터에 여파된다. 이 프리엠퍼시스된 음성신호로부터 선형 예측 계수(Linear Prediction Coefficient)를 구하게 되면 식(1)을 이용하여 캡스트럼 계수를 구할 수 있다.[3][4]

$$
\begin{aligned}
c_0 &= \ln \sigma^2 \\
c_m &= a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k} \qquad 1 \le m \le p \\
c_m &= \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k} \qquad p < m \qquad (1)
\end{aligned}
$$

여기서 σ^2는 LPC계수의 이득항을 의미하고, p는 LPC계수의 차수, m은 캡스트럼 계수의 차수를 의미한다.

또한, 화자간의 변별력을 극대화하기 위하여 F-비를 이용하여 특징파라메터의 차수를 결정하였다.
F-비는 다음과 같이 화자 내의 변이로 화자간의 변이를 나눈 값이며 식(2)와 같다.[5]

$$
\begin{aligned}
F_{RATIO} &= \frac{variance\ of\ speaker\ means}{mean\ intra-speaker\ variance} \\
&= \left[\frac{1}{m}\sum_{i=1}^{m}(\mu_i-\mu)^2\right] \Big/ \left[\frac{1}{m\cdot n}\sum_{i=1}^{m}\sum_{k=1}^{n}(X_{ik}-\mu_i)^2\right] \qquad (2)
\end{aligned}
$$

여기에서 μ_i 각 화자의 평균이며, μ는 모든 화자의 평균을 나타낸다, X_{ik}는 특징파라메터를 나타낸다.

특징파라메터로 사용될 데이터는 화자간의 변이는 크고 화자내의 변이는 작은 것이 좋은 파라메터이다. 즉, F-비 값이 클수록 음성인식의 인식률을 높일 수 있다. 이를 각 차수별로 측정하여 실제 특징파라메터의 차수를 결정하였다.

3. 인식 시스템

인식 시스템은 일단 음성인식 부분과 화자인식 부분으로 구분하여 구성하였다.

3.1. 음성인식부분

음성인식방식에 있어서 크게 세부분으로 나눌 수 있다. DP(Dynamic Programming)기법을 이용하여 입력 음성을 데이터 베이스로서 저장해둔 음성과 비교하여 가장 유사한 음성을 찾아내는 DTW(Dynamic Time Warping)방식, 음성의 특징 파라메터간의 거리를 이용하여 확률적 모델간의 유사도를 이용한 HMM(Hidden Markov Model)방식, 인간의 뇌의 구조를 모델링하여 병렬처리 및 적응학습을 하는 NN(Neural Network)방식 등이 있다. 본 논문에서는 훈련을 위하여 특별히 시간이 필요하지 않는 DTW방식을 채택하였다.

또한 인식률을 높이기 위하여 2순위, 3순위의 패턴을 동시에 찾아내어 세가지 경우중에 2가지 이상 일치할 경우 인식으로 결정하였다.

3.2. 화자인식부분

화자인식부분 : 음성을 이용한 출입통제 시스템에 화자를 등록하거나 또는 확인을 하는데에는 화자가 많은 시간을 이용하여 화자의 등록을 해야하기에는 부담이 있다. 그러므로 현재 가장 많이 사용되는 화자인식모델인 GMM(gaussian mixture speaker model)에서의 학습을 위하여는 최소한 30초이상의 발성한 데이터에 대해서는 최대 85%까지 인식을 한 것으로 나타나 있다[6]. 또한 90%이상의 인식률을 갖기 위해서는 60초 이상의 발성된 데이터가 필요한 것으로 나타났으며, 그 외 여러 가지 화자모델을 사용하였을 경우에도 학습을 위한 많은 시간의 데이터가 있어야 높은 인식률이 나타나고 있다. 또한 이러한 많은 시간의 학습시간을 단축하기 위한 노력들도 나타나고 있다. 본 논문에서는 이러한 점에서 음성인식부분에서 사용되었던 알고리즘을 같이 사용하도록 하였다.

3.3. DTW 알고리즘

DTW 알고리즘은 시험 패턴과 기준 패턴을 시간축상에 최적이 되도록 배열한 후, 이 최적 변형 경로를 통한 최적 거리를 얻어내는 방법이다.

DTW 알고리즘은 2차원 $d_1(m,n)$ 평면상에서 적절한 제한 조건을 만족하는 $d(1,1)$에서 $d(M,N)$까지의 최적 경로를 구하는 방법이다. 임의의 점 $(m(j),n(j))$까지 축적된 거리를 다음과 같이 정의할 수 있다.

$$C_1(m,n)=\sum_{j=1}^{J} d_1(m(j),n(j))\cdot w(j) \qquad (3)$$

여기에서 $(m(j),n(j))$, $j=1,2,\cdots,J$ 는 주어진 경로이고, $w(j)$는 각 경로에 따른 가중치이다. 그러면 최적 경로는 $d(M,N)$에서의 축적된 거리 $C_1(M,N)$을 최소화하는 경로로

$$D_1(T,R_i)=\min_{path} C_1(M,N) \qquad (4)$$

로 표시된다. 최적 경로를 구하는 과정에서 제한 조건들은 다음과 같은 목적으로 주어진다. 즉, 부분적으로는 경로 기울기의 범위를 제한하며, 전체적으로는 경로의 허용 영역을 제한한다.

그림 1은 warping 함수에 의한 두 패턴 A, B를 정합시키는 과정을 나타낸 것이다.

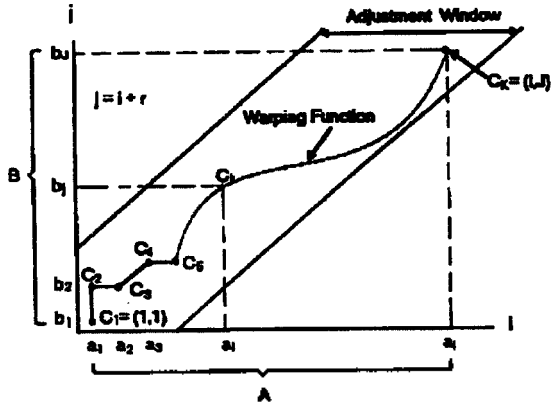

그림 1. Warping 함수와 조정 창

4. 실험 및 결과

본 논문에서는 화자인식 실험을 위해 성인 남자 5명에 대하여 5개의 단어에 대하여 6회를 발성하여 앞의 3회는 데이터는 참조패턴으로 나중 3회 데이터는 시험패턴으로 하였다. 녹음환경은 주변잡음이 존재하는 일반적인 실험실이며, 5개의 단어는 8kHz 16bit로 녹음한 후에 음성 특징벡터로 LPC-켑스트럼을 추출하였다.

화자인식을 위한 구성도는 그림 2와 같다.

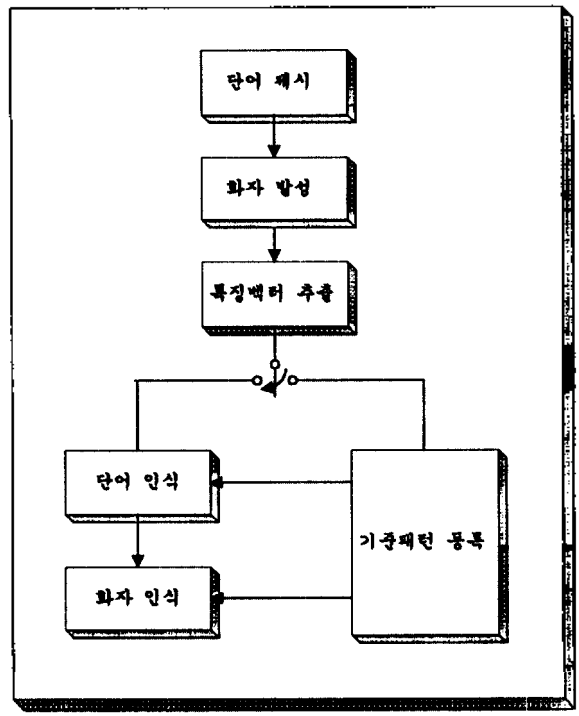

그림 2. 화자인식 시스템 구성도

문맥 독립형 과 문맥 종속형의 화자확인시스템은 매우 위험한 요소를 갖고 있다 왜냐하면 접근허가를 얻지 않은 사람도 미리 접근허가자의 목소리를 녹음한 후에 이것을 다시 재생하여 쉽게 접근허가를 받을 수 있기 때문이다. 발성해야하는 단어가 제시되는 문맥제시형 시스템은 이러한 단점을 극복할 수 있다. 단어의 제시는 랜덤하게 시스템에서 제공이 되며 이러한 제공되는 단어는 우선 단어를 인식을 수행한다. 이러한 처리절차는 시스템의 어휘가 충분히 있을 경우에는 두 단계로 나누어서 수행한다.

첫 번째 화자가 무엇이라고 말했는가를 이해해야한다, 두 번째는 인식된 어휘에 해당하는 화자를 인식하게 되는 것이다. 그림에서 보는바와 같이 화자는 제시된 단어를 발성하게된다. 정확한 화자 인식을 위해서는 화자의 발성 습관 및 발성 기관의 개인차를 잘 나타낼 수 있는 특징 파라메터의 선택과 화자의 특성을 정의할 수 있는 방법이 필요하며, 특히 문맥종속형 화자인식을 대상으로 할 경우 어휘 선택이 중요하다. 화자인식을 위한 표준어휘세트를 선정하기 위해서 음소별 화자식별을 실험하여 각 음소의 화자특성 반영정도를 비교하고, 실제 단어 내에서의 영향을 고려하기 위해 단어별 화자식별 실험을 통한 단어들 중에서 5개를 선정하였다.

표 1 화자인식용 단어

1	공예품
2	나그네
3	요사이
4	발자취
5	훈민정음

화자에 따른 단어인식률을 표 2에서 보였다.

표 2. 화자에 따른 단어 인식률

	1	2	3	4	5	인식률
	2/3	3/3	3/3	2/3	3/3	86.7%
	3/3	3/3	3/3	2/3	3/3	93.3%
	3/3	3/3	3/3	3/3	3/3	100%
	3/3	3/3	3/3	3/3	3/3	100%
	3/3	2/3	1/3	3/3	3/3	80%
	93.3%	93.3%	86.7%	86.7%	100%	

또한, 인식된 단어에 대하여 화자에 따른 화자인식률을 표 3에서 보였다.

표 3. 화자에 따른 화자 인식률

	92.3%	7.7%			
	7.1%	92.9%			
		6.7%	86.6%	6.7%	
		6.7%		93.3%	
					100%

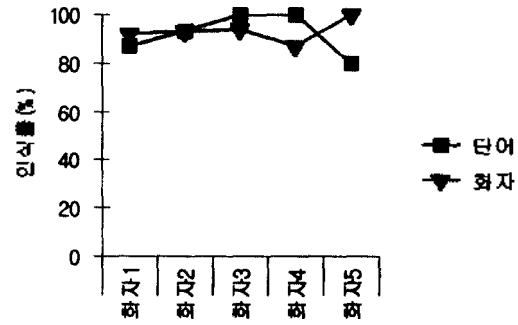

그림 3. 단어 및 화자 인식률

5. 결 론

본 논문에서는 문맥을 미리 제시하여 화자가 제시된 단어를 발성하는 형태의 화자인식을 수행하였다. 이러한 방식은 문맥독립형 화자인식의 단점이 장시간의 학습데이터의 요구 및 녹음음성의 재생의 문제를 해결할 수 있을 뿐만 아니라 문맥종속형의 단점인 등내념의 문제를 해결할 수 있다. 전체화자에 대한 단어 인식률은 92%이고 인식된 단어에 대한 화자인식률은 92.7%의 결과를 얻었다.

참고 문헌

[1] Chi Wei Chi, "An HMM Approach to Text-Prompted Speaker Verification," Proc. of the IEEE, vol.2. pp. 673-676. 1996.

[2] Furui & Sandhi, " Advances in Speech Signal Processing" Dekker. 1991.

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

[4] L. R. Rabiner & Biing-Hwang Juang, "Fundamentals of Speech Recognition," Prentice-Hall, AT&T, U.S.A, 1993

[5] 경연정 외 2인 "F-비를 가중치로 사용한 문장고정형 화자인식 시스템," 한국음향학회 학술대회 논문집 제15권, pp. 79-82. 1996년

[6] D. A. Reynolds, R. C. Rose, "Robust Text -Independent Speaker Identification Using Gaussian Mixture Speaker Models," IEEE Trans., Speech Audio Processing, vol. 3, no. 1. pp. 72-83, 1995.