확률신경망에 의한 숫자음성열로부터의 화자확인

엄익태*, 강권일**, 김문현**
*동원대학 전자계산과
**성균관대학교 전기전자 및 컴퓨터공학부
iteum@tongwon.ac.kr, kangki@nature.skku.ac.kr, mhkim@ece.skku.ac.kr

Speaker Verification for Spoken Digit Sequence by Probabilistic Neural Network

Ig-Tae Um*, Kwon-Il Kang**, Moon-Hyn Kim**
*Dept. of Computer Engineering, Tongwon College
**Dept. of Information Engineering, Sungkyunkwan Univ.

요약

화자확인은 기본적으로 각 입력 음성에 대해 하나의 임계치를 기준으로 수락과 거부의 두 가지 결정을 내리나, 본 논문은 네 자리의 비밀번호를 음성으로 입력하였을 때 각 숫자음성에 대한 지역적인 결정을 두 개의 임계치를 이용하여 수락, 거부, 결정유보의 세 가지로 구분하고, 비밀번호 전체에 대한 판단 규칙을 제안하였다. 지역적 결정에 필요한 화자에 대한 신뢰척도의 측정치는 확률신경망을 통해 구하였다. 다섯 명의 화자를 대상으로 수행한 실험에서 하나의 임계치를 이용한 기존의 방식은 5.3%의 오류를 나타냈고, 본 논문에서 제안한 방식은 2.1%의 오류를 보였다.

1. 서론

화자확인은 임의의 화자로부터 음성데이터와 화자의 신원을 제공받았을 때, 그 음성데이터의 화자의 신원이 주장되는 신원과 일치하는 지를 판단하는 과제이다. 화자확인은 음성을 기반으로 한 출입통제 및 보안시스템 등의 응용분야에 적용될 수 있다.

일반적인 화자확인 시스템은 등록된 화자로부터 훈련데이터를 제공받아 통계적인 방법이나 신경망에 의해 화자모델을 만들고, 화자확인을 위한 음성이 입력될 때 화자모델에 대한 신뢰척도의 측정값을 구하여, 사전에 결정된 임계치와의 비교를 통해 수락 또는 거부의 결정을 내린다.

화자확인 시스템은 입력음성에 대한 제약에 따라 문장고정형과 문장독립형으로 분류할 수 있다. 문장고정형 화자확인 시스템은 화자모델을 만들기 위한 훈련데이터를 수집할 때 화자에게 지정된 문장을 발화하도록 요구하고, 후에 화자확인을 위한 입력에 대해서도 동일한 문장을 사용하도록 한다. 이 유형의 시스템들은 기본적으로 템플릿 정합을 수행하게 되는데, 이를 위해 DTW(Dynamic Time Warping)이나 HMM(Hidden Markov Model)등이 많이 이용된다.[1]

문장독립형 화자확인 시스템은 입력되는 문장에 제약을 두지 않는다. 이 유형의 대표적인 접근방법으로는 많은 문장을 통계적으로 처리하는 방식[2]과 음소단위의 화자모델을 구축하여 입력음성으로부터 추출한 음소들에 대해 신뢰척도를 측정하여 판단하는 방식[3]이 있다. 이 방식들은 HMM(Hidden Markov Model), GMM(Gaussian Mixture Model), 신경망등을 채택하고 있다.

화자확인에 신경망을 적용할 경우 출력층 노드의 활성화 수준을 신뢰척도의 값으로 이용한다. 그러나 이에 대한 이론적 근거는 명확하지 않다.

본 논문은 비밀번호와 같이 여러 개의 숫자들로 구성된 데이터를 음성을 통해 발화하였을 때, 발화자에 대한 화자확인의 방법을 제안하였다. 개개의 숫자음절은 확률신경망에 의해 처리되고, 각 숫자음절에 대한 신뢰척도의 측정치는 임계치와 비교되는데, 이때 하나의 임계치를 이용하여 수락 또는 거부를 결정하는 방법과 두 개의 임계치를 이용하여 수락, 거부, 판단유보의 세 가지로 결정하는 방법을 검토하였고, 이에 따라 단위 숫자음성에 대한 지역적 결정을 통합하는 두 가지 방식을 제안하였다.

2. 화자확인 시스템

2.1 신뢰척도

화자확인은 N 개의 클래스 중 선택된 하나의 클래스에 대해 주어진 입력이 그 클래스에 속하는지를 결정하는 문제이다. 일반적인 화자확인 시스템은 사전에 등록된 화자에 대하여 훈련데이터로부터 화자모델을 만들고, 입력음성을 주장된 화자모델과 비교하여 그 화자에 대한 신뢰척도를 구한다. 신뢰척도로 베이즈 분류기의 사후확률을 이용할 수 있다. 이때 베이즈의 정리에 의해 사후확률 $P(\lambda|X)$는

$$P(\lambda|X)=\frac{p(X|\lambda)P(\lambda)}{p(X)} \quad (1)$$

로 표현된다. 여기서 X는 입력음성이고, λ는 화자모델을 나타낸다. 그런데 이 사후확률을 신뢰척도로 직접 사용할 경우, $P(\lambda)$는 시스템에 새로운 화자가 등록될 때마다 값이 변하며, $p(X|\lambda)$는 입력채널의 상태, 주변의 잡음, 발화시 화자의 상태 등에 따라 화자내 변이가 크게 나타나므로 시스템 성능을 저하시킬 수 있다. 이에 대한 대책으로 사후확률의 비율에 의한 정규화가 많이 채택된다. 즉, 정규화된 신뢰척도는

$$\theta=\frac{P(\lambda_i|X)}{P(\lambda_j|X)}=\frac{p(X|\lambda_i)}{p(X|\lambda_j)} \quad (2)$$

로 나타낼 수 있다. 여기서 λ_i는 주장된 신원의 화자를 나타내고, λ_j는 정규화를 위해 선택된 화자집단을 나타낸다. 정규화를 위한 화자집단의 선택에는 여러 가지 방식이 있으며 대표적인 방법으로는 주장된 신원의 화자와 음향적으로 가까운 화자들의 집단을 이용하는 것이다. 본 논문에서는 소규모의 화자집단을 대상으로 실험을 하기 때문에 정규화 화자집단에 주장된 신원의 화자를 제외한 모든 화자들을 포함시켰다. 그리고 $p(X|\lambda_j)$는 정규화 화자집단에 포함된 모든 화자들에 대한 $p(X|\lambda)$들을 평균하여 구하였다.

2.2 확률신경망

식 (2)에서 알 수 있듯이 정확한 신뢰척도를 얻기 위하여 정확한 확률밀도함수 $p(X|\lambda)$의 추정이 요구된다. 확률신경망은 커널(Kernel) 기반의 확률밀도추정 방법을 신경망의 구조로 표현한 것으로서 구해지는 확률밀도함수가 이론적으로 잘 규

명되어 있다는 특징이 있다.

그림1.은 확률신경망의 원형을 보여주고 있다.[4]

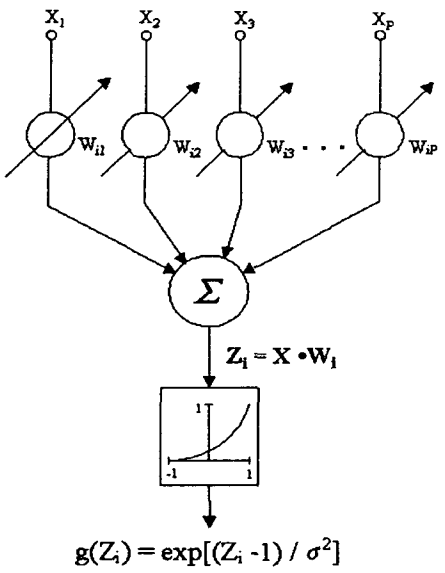

그림 1. 확률신경망의 구조

그림1.에서 X와 Wi는 p차원의 벡터이며 X는 입력, Wi는 훈련데이터중의 하나를 나타낸다. σ는 너비 파라미터로서 확률밀도함수 추정의 정확성을 조절하는 기능을 한다. 본 논문에서는 확률밀도함수 추정자로서

$$g(X)=\frac{1}{n}\sum_{i=1}^{n} e^{-\frac{\|X-X_i\|^2}{2\sigma^2}} \qquad (3)$$

를 사용한다. 식(3)에서 X_i는 훈련데이터에 포함된 하나의 데이터를 나타낸다. 따라서 확률밀도함수는 n개의 가우션(Gaussian) 함수의 합의 평균으로 구할 수 있다. 확률신경망은 다층퍼셉트론에서와 같은 긴 학습시간을 필요로 하지 않으므로, 새로운 화자를 등록할 때 장점이 있다. 반면에 화자확인 실행 시 훈련데이터의 크기와 화자의 수에 따라 확률밀도함수 값의 계산이 길어질 수 있다.

2.3 숫자음성열에 대한 화자확인

본 논문은 네 개의 숫자음을 포함하는 숫자음성열을 입력으로 한다. 이를 위해, 화자 k의 숫자 j에 대한 임계치 $\theta_0^{k,j}$를 설정한다. 입력되는 숫자음성열이 $X=X_1X_2\ldots X_N$로 표시되고, 주장된 화자가 k일 때, X에 대한 수락/거부의 판단 d^k는 i번째 숫자음성 X_i에 대한 지역적인 판단 d_i^k를 종합하여 구할 수 있다. 위의 X_i가 '영'에서 '구'까지의 숫자 중 j를 나타낼 때 d_i^k는 X_i에 대한 신뢰척도 θ_i와 임계치 $\theta_0^{k,j}$의 비교에 의하여 구한다. 임계치 $\theta_0^{k,j}$는 그림 2.와 같은 오류분포도에 의해 구해진다.

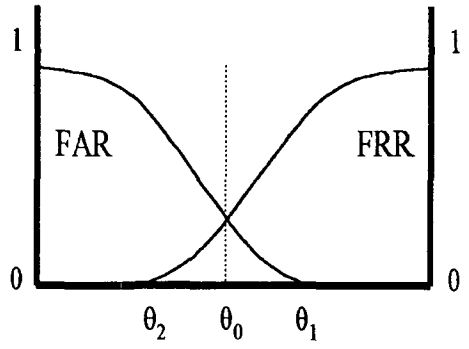

그림2. 임계치 결정을 위한 화자확인 오류분포도

그림2.는 특정 화자의 특정 숫자음성에 대한 화자확인 오류분포를 나타낸다. 따라서 k명의 등록된 화자에 대해 10 k개의 오류분포도를 구해야한다.

그림2.에서 FAR(False Acceptance Rate)은 사칭자를 수락하는 오류율을 나타내며, FRR(False Reject Rate)은 참 화자를 거부하는 오류율을 나타낸다. FAR은 사칭자로부터의 데이터에 대한 신뢰척도 측정값을 누적시킨 결과이고, FRR은 참화

자에 대한 신뢰척도 측정값의 누적치를 나타낸다.

θ_0는 FAR과 FRR이 같아지는 지점을 나타내며, 기존의 많은 연구들은 이 임계치를 기준으로 수락과 거부의 두 가지 결정을 내린다. θ_1은 FAR이 영이 되는 임계치를 나타내고, θ_2는 FRR이 영이 되는 임계치이다. θ_0를 기준으로 한 판단을 할 경우, 임의의 단위음성에 대한 신뢰척도 측정값이 θ_1과 θ_2 사이의 값을 취하게 된다면 판단에 오류를 포함할 가능성이 높아진다. 즉, 단위음성에 대한 판단에서부터 오류를 허용하는 것으로 볼 수 있다. 따라서 본 논문에서는 두 개의 임계치 θ_1과 θ_2를 이용하여 결정구간을 세 영역으로 나눈다. 즉 임의의 단위음성에 대한 신뢰척도의 측정값이 θ_1보다 클 경우 수락, θ_2보다 작게 되면 거부, θ_1과 θ_2의 사이의 값을 갖게되면 판단유보의 결정을 내린다. 여기서 판단유보는 나중에 단위음성에 대한 지역적인 판단을 종합하여 최종판단을 내릴 때 중요한 결정권을 갖지 않는 것을 나타낸다.

본 논문에서는 단위음성에 대해 두 가지 결정을 내릴 경우와 세 가지 결정을 내릴 경우의 전체 시스템 성능에 미치는 영향을 알아보기 위하여 다음과 같이 두 가지 방법을 고안하였다.

방법1.

X_i가 숫자음성 j를 나타내고, 주장되는 화자가 k일 때, X_i에 대한 판단 d_i^k는 $\theta_0^{k,j}$를 기준으로 한다. 이때 θ_i는 X_i에 대한 신뢰척도의 측정값이다.

$$d_i^k = 1 \quad 단,\ \theta_i \geq \theta_0^{k,j} \qquad (4)$$
$$= 0 \quad 단,\ \theta_i < \theta_0^{k,j}$$

$$d^k = \prod_{i=1}^{N} d_i^k \quad (d^k = 1 \text{ 일 때 수락}) \qquad (5)$$

식(4)에서 구한 d_i^k를 종합하는 방법으로 식(5)의 공식을 채택하였다. 식(5)가 의미하는 바는 하나라도 단위음성에 대한 판단이 거부로 나타나면 전체 숫자음성열에 대한 판단을 거부로 결정하는 것이다. 이러한 공식은 특정 화자에 대한 음성을 사칭할 때, 사칭자가 흉내낼 수 없는 음소들이 존재한다는 사실에 근거를 두고 있다. 식(5)에서 N은 숫자음성열 X에 포함된 단위수자음성들의 개수를 나타낸다. 본 논문의 실험에서는 N=4를 이용하였다.

방법2.

X_i에 대하여 수락, 거부, 판단유보의 세 가지 결정을 사용하는 것이다. 이를 위해 $\theta_1^{k,j}$과 $\theta_2^{k,j}$를 사용한다.

$$d_i^k = 1 \quad 단,\ \theta_i > \theta_1^{k,j} \qquad (6)$$
$$= 0 \quad 단,\ \theta_i < \theta_2^{k,j}$$
$$= \alpha \quad 단,\ \theta_2^{k,j} \leq \theta_i \leq \theta_1^{k,j},\ 0 < \alpha < 1$$

$$d^k = \prod_{i=1}^{N} d_i^k \quad (d^k > \alpha^N \text{ 일 때 수락}) \qquad (7)$$

식(7)에서 하나라도 단위음성에 대한 결정이 거부로 나타나면 전체 숫자음성열에 대한 최종판단이 거부로 결정되지만, 판단유보로 결정이 난 단위음성은 최종판단에 중요한 역할을 하지 않게 된다.

3. 실험 및 결과

본 논문의 실험은 5명의 화자를 대상으로 하였다. 실험에 참여한 화자는 모두 남자이며 연령적

으로는 20대가 3명, 30대가 2명이다. 각각의 화자로부터 '영'에서 '구'까지의 숫자음성별로 20개씩의 데이터를 얻었다. 이중 숫자별 화자모델을 만드는데 15개의 데이터가 이용되었고, 나머지 5개의 데이터가 테스트용으로 사용되었다. 그리고 특정화자 1명을 선택하였을 때 나머지 4명의 화자가 사칭자의 역할을 맡았다. 음성데이터는 컴퓨터용 마이크를 통해 16KHz 16비트로 수집되었다. 수집된 데이터들은 음성의 길이가 각각 달라 강제적으로 길이정규화를 수행하였으며, 길이정규화 후에 숫자음성을 프레임 단위로 나누고 이에 대해 MFCC(Mel Frequency Cepstral Coefficient)를 구했다. 각 프레임에 대한 12차 MFCC를 하나의 열 벡터로 합성하여 단위숫자음성에 대한 특징벡터로 이용하였다. 각 화자별 숫자별로 클래스를 나누고 각 클래스별로 확률신경망을 이용하여 확률밀도함수를 추정하였다. 표1.은 단위숫자음성에 대한 실험결과로서 각 항은 화자별 숫자별로 참화자의 데이터에 대한 신뢰척도를 측정한 값과 사칭자의 데이터에 대한 측정값이 그림2.에서의 θ_1과 θ_2사이에 놓이게 된 회수를 백분율로 나타낸다. 즉 (FAR(θ_2)+FRR(θ_1))×100(%)를 나타낸다.

표1.일 보면 '일'과 '육'은 화자간 식별이 잘 되고 있으나, '삼','사','오'등은 화자간 신뢰척도의 측정값이 많이 중첩됨을 보여준다. 36% 까지 높은 비율이 나타나게된 원인으로서는 실험에 참여한 화자들의 특성과 실험에 사용한 방법상의 문제점을 고려해 볼 수 있다. 화자들이 비슷한 연령층의 남성들이기 때문에 원천적으로 화자확인이 어려워질 수도 있다. 보다 객관적인 원인으로서는 강제적인 길이 정규화에 의한 음성데이터의 왜곡과 각 클래스에 대한 훈련데이터의 부족에 따른 부정확한 확

표1. 단위숫자음성에 대한 오류율(%)

	화자1	화자2	화자3	화자4	화자5
영	0	24	24	0	8
일	0	0	0	4	8
이	4	0	0	2	24
삼	16	24	24	0	20
사	0	24	24	36	4
오	16	0	20	16	20
육	0	0	0	0	4
칠	16	0	0	20	0
팔	0	8	8	20	0
구	0	0	0	20	24

률밀도함수 추정 등을 생각할 수 있다. 또한 확률신경망을 사용하는데 따르는 σ의 최적치 선정의 문제 등도 실험결과에 영향을 미치는 것으로 생각된다. 비밀번호 네 자리 숫자를 음성으로 발화했을 경우에 대한 최종실험에서 방법1은 5.3%의 오류를 보인 반면 방법2는 2.1%의 오류율을 보였다. 따라서 지역적 판단을 효과적으로 종합하는 방법의 모색이 중요한 문제임을 알 수 있다.

실험은 Pentium 컴퓨터 상에서 Matlab을 이용하여 만든 시뮬레이션 프로그램을 통해 이루어졌다.

4. 결론

화자확인은 등록된 화자들 중에서 음성입력의 발화자를 찾는 화자식별과 달리 적합한 신뢰척도를 요구하고 정확한 측정값을 필요로 한다. 본 논문은 신경망을 이용한 화자확인 시스템의 개발을 위해 확률신경망을 채택하였다. 확률신경망을 이용한 확률밀도함수의 추정은 이론적 근거가 명확

하므로 화자확인에 적합하다고 생각된다. 숫자음성열에 의한 화자확인은 단일 숫자음성에 의한 화자확인 보다 효율적이라는 것을 실험결과를 통해 알 수 있다. 여러 개의 단위음성으로부터 얻어진 정보를 종합하는 방법은 음소기반의 문장독립형 화자확인에서도 중요한 문제가 되고 있는데 본 논문에서 제안한 방법이 잘 적용될 수 있다고 판단된다.

참고문헌

[1] Furui, S.,"Recent advances in speaker recognition", Pattern Recognition Letters, Vol. 18 , PP. 859-872, 1997

[2] Reynolds, D. A., Rose, C. R.,"Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Trans. on Speech and Audio Processing, Vol. 3, No. 1, PP.72-83, January 1995

[3] Olsen, J., " A two-stage procedure for phone based speaker verification", Pattern Recognition Letters, Vol. 18, PP 889-897, 1997

[4] Specht, D. F., " Probabilistic Neural Networks", Vol. 3, PP.109-118, 1990

[5] 오영환, 음성언어정보처리, 홍릉과학출판사, 1998

[6] Bishop, C. M., Neural Networks for Pattern Recognition, Oxford University Press, 1995