論文2000-37TE-3-3

Discriminant 학습을 이용한 전화 숫자음 인식

(Telephone Digit Speech Recognition using Discriminant Learning)

韓 汶 星 *, 崔 完 洙 **, 權 賢 稷 ***

(Mun-Sung Han, Wan-Soo Choi, and Hyun-Jik, Kwon)

요 약

대부분의 음성인식 시스템이 확률 모델을 기반으로 한 HMM 방법을 가장 많이 사용하고 있다. 한국어 고립 전화 숫자음 인식인 경우에 만약 충분한 학습 데이터가 주어지면 HMM 방법을 사용해도 높은 인식률을 얻는다. 그러나 한국어 연속 전화 숫자음 인식인 경우에 비슷하게 발음되는 전화 숫자음들에 대해서는 HMM방법이 한계를 가지고 있다. 본 논문에서는 한국어 연속 전화 숫자음 인식에서 HMM 방법의 한계를 극복하기 위해 discriminant 학습 방법을 제시한다. 실험결과는 우리가 제시한 discriminant 학습 방법이 비슷하게 발음되는 전화 숫자음들에 대해서 높은 인식률을 갖는 것을 보여준다.

Abstract

Most of speech recognition systems are using Hidden Markov Model based on statistical modelling frequently. In Korean isolated telephone digit speech recognition, high recognition rate is gained by using HMM if many training data are given. But in Korean continuous telephone digit speech recognition, HMM has some limitations for similar telephone digits. In this paper we suggest a way to overcome some limitations of HMM by using discriminant learning based on minimal classification error criterion in Korean continuous telephone digit speech recognition. The experimental results show our method has high recognition rate for similar telephone digits.

Ⅰ. 서 론

HMM(Hidden Markov Model)은 음성신호를 Markov 상태열의 천이과정에서 발생하는 확률모델을 기반으로 하여 음성 데이터로부터 추출된 정보를 통계학적으로 모델링하는 알고리즘이다. 최근의 음성인식 시스템들은 연속음성을 대상으로 인식을 수행하는 경우가 대부분이다. 이러한 음성인식 시스템들은 음소단위의 학습이 요구되며 이것은 많은 작업과 많은 계산량이 요구되는 단점을 가지고 있다.

실생활에서 사용되는 네 자리 숫자 비밀번호 및 구내 전화번호등과 같은 음성인식 시스템은 HMM을 이용하여 간단히 구축될 수 있다. 그러나, HMM은 학습시에 많은 학습데이터를 필요로 하고, 비슷한 음성 신호에 대해서는 인식률이 낮은 단점을 가지고 있다. HMM을 이용하여 고립단어에 대한 전화음성 인식을 하는 경우 사용된 학습데이터의 양이 부족하면 음성인식률이 낮다.

실세계 음성인식의 활용을 위하여는 유사어의 분별

* 正會員, 韓國電子通信研究院
(Electronics and Telecommunications Research Institute)
** 正會員, 大林大學 電子情報通信科
(Dept. of Electronics, Information, and Communication, Daelim College)
** 正會員, 韓國科學技術院
(KAIST)

※ This research was supported by Brain Science and engineering Research Program sponsored by Korean Ministry of Science and Technology.
接受日字:2000年5月9日, 수정완료일:2000年8月17日

이나 잡음 처리가 매우 중요하다. 비슷한 발음의 음성 신호에 대한 처리, 잡음처리 등에 널리 쓰이는 방법이 discriminant 학습이다.[1][2] 또한 discriminant 학습은 적은 학습데이터가 필연적인 화자인식, 화자확인 분야에도 널리 적용되고 있다.[5]

본 논문에서는 연속 숫자음의 발음을 2음절 또는 3음절 등에서의 발음현상을 모두 고려하지 않고 음절단위로 나누어 각각 고립 숫자음으로 인식한다. 그리고 유사한 전화 숫자음 처리에 있어서는 discriminant 학습을 적용하여 HMM 방법을 사용하는 것보다 높은 인식률을 갖는다. 따라서 적은 비용과 계산량으로 효과적인 한국어 숫자 음성인식 시스템을 구축할 수 있음을 보여준다.

Ⅱ. HMM을 활용한 음성인식

HMM을 기반으로 한 음성인식 시스템은 서론에서 언급한 몇 개의 단점이 있으나 현재까지는 널리 쓰이고 있는 방법이다. 일반적으로 HMM은 N개의 상태 수를 가진 left-to-right 모델로 구성되며, 상태별 초기확률을 나타내는 $\Pi=\{\pi_i : 1\le i\le N\}$, 상태 전이 확률 행렬 $A=\{a_{ij} : 1\le i,j\le N\}$, 상태별 관찰 확률을 나타내는 $B=\{b_i : 1\le i\le N\}$을 구성 원소로 하는 벡터 $\Lambda=(\Pi, A, B)$로 모델을 표현한다.

본 논문에서 HMM 모델은 그림 1에서와 같이 개별 숫자음 데이터에 대해 3개 또는 4개의 상태 수를 가지고, 상태별 전이는 2개의 상태만 가능한 left-to-right 모델로 하였다.

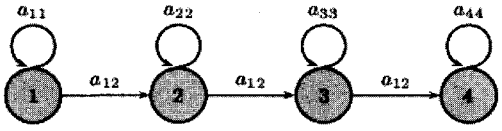


그림 1. 상태 전이도
Fig. 1. State transition diagram.

주어진 음성 벡터 $X=(x_1, x_2, \cdots, x_{T_f})$에 대해 Viterbi 알고리즘에 의한 최적 상태변화가

$$q=(q_0, q_1, \cdots, q_{T_f})$$

라면 최대 유사정도(maximum likelihood)는

$$L(X)=\sum_q \pi_{q_0} \prod_{t=1}^{T_f} a_{q_{t-1}q_t} b_{q_t}(x_t)$$

로 계산된다. 여기에서 T_f는 프레임의 개수이고, $b_{q_t}(x_t)$는 상태 j 에서 x_t를 관찰하게 될 확률을 나타내며, 상태별로 모인 벡터들을 군집한 후, 군집별로 평균과 분산을 이용한 정규분포로 계산된다. HMM은 보통 최대유사정도 값을 계산함으로써 음성 패턴을 찾아가는 통계적인 모델로 특징 지울 수 있다. 그러나 HMM은 많은 학습자료를 필요로 한다는 단점을 지닌다.

Ⅲ. HMM을 기반한 Discriminant 학습

Discriminant 학습은 비슷한 발음 등으로 인식오류의 가능성이 있는 다른 음성모델을 함께 고려하는 학습시스템으로서 MCE (Minimum Classification Error)를 기반으로 한다. MCE를 최소로 하는 방향으로 HMM 학습 결과를 개선시키는 discriminant 학습은 화자확인, 화자인식 등에 광범위하게 적용되고 있으며, 불충분한 학습자료를 사용한 HMM학습결과로 초래되는 인식률 저하를 개선할 수 있다. 특히 실 세계에서의 음성인식은 충분한 데이터를 수집하기 어려운 경우가 많아 통계적인 방법인 HMM의 단점을 보완할 수 있다는 점에서 실용적으로 널리 활용가능하다.

K개의 HMM학습 자료를 $\Lambda=\{\lambda_k, 1\le k\le K\}$로 표시하고, 각각의 음성 패턴 X_i, $(1\le i\le X_T)$는 K개의 음성 클래스 C_k, $(1\le k\le K)$ 중 하나에 포함된다. 여기에서 X_T는 입력 데이터의 수를 나타낸다. HMM에 기반한 discriminant 학습은 Viterbi 알고리즘에 의한 로그 유사정도 값을 discriminant 함수로 사용한다. HMM 학습의 개선은 GPD(Generalized Probabilistic Descent)방법을 사용하여 반복적으로 MCE를 최소화시키는 학습결과를 만들어낸다.

HMM에 기반한 GPD는 다음과 같은 세 가지 함수를 구성 요소로 한다.

(a) Discriminant 함수 $g_k(X_i ; \Lambda)$: 음성 패턴 X_i가 클래스 C_k에 대한 HMM 유사정도 값

$$g_k(X_i : \Lambda)=\log[L_k(X_i)]$$

을 discriminant 함수로 사용한다. 따라서 음성 패턴 X_i는 Viterbi 알고리즘에 의한 유사정도 값이 최대인 클

래스 C_k에 속하는 것으로 인식된다.

$$X_i \in C_j \quad \text{if} \quad g_j(X_i\ ;\ \Lambda) = \max_k g_k(X_i\ ;\ \Lambda)$$

(b) Misclassification 측정 $d_k(X_i\ ;\ \Lambda)$: 인식오류1에 대한 척도는

$$d_k(X_i\ ;\ \Lambda) = -g_k(X_i\ ;\ \Lambda) + G_k(X_i\ ;\ \Lambda)$$

로 정의한다. 여기에서 $G_k(X_i\ ;\ \Lambda)$는 클래스 C_k에 들어가야 할 입력데이터 X_i가 다른 클래스에 들어가는 모든 경우를 의미하며, 다음과 같이 다른 클래스들의 유사정도 값들의 기하평균의 로그 값으로 계산한다.

$$G_k(X_i\ ;\ \Lambda) = \log\left(\frac{1}{K-1}\sum_{i,j\neq k}\exp[\eta g_j(X_i\ ;\ \Lambda)]\right)^{1/\eta}$$

일반적으로 $\eta=1$로 놓는다. $d_k(X_i\ ;\ \Lambda)$의 정의로부터 입력 데이터 X_i가 올바르게 인식되었다면 d_k의 값이 음의 값을, 잘못 인식된 경우는 양의 값을 가짐을 알 수 있다.

(c) Loss 함수 $l_k(X_i\ ;\ \Lambda)$: GPD 방법은 미분을 이용해 파라미터 Λ를 반복적으로 보정해가는 방법이다. 따라서 loss 함수는 다음과 같이 0과 1 사이의 값을 갖는 sigmoid 함수로 정의된다. 이 sigmoid 함수는 단조증가이며 미분 가능한 함수이다.

$$l_k(X_i\ ;\ \Lambda) = \ell(d_k) = \frac{1}{1+\exp[-a(d_k+b)]}$$

여기에서 a(>0)는 상수이고, d_k+b 부근에서 sigmoid 함수의 기울기를 나타낸다. 또한 적당한 b값의 설정은 sigmoid 함수의 미분값이 0을 갖게 되는 것을 방지해 파라미터 Λ의 변화를 가능하게 한다. 일반적으로, GPD에 의한 파라미터 Λ의 변화는 misclassification 측정 d_k의 값이 -b의 근처에 있을 때, 결정적으로 일어나게 된다. 따라서 효율적인 GPD 알고리즘을 구현하기 위해서는

$$b = d_{\min} = -\max_{1\le i\le X_T,\,1\le k\le K}[-d_k(X_i;\Lambda)]\,\chi(X_i\in C_l)$$

로 설정하는 것이 좋다. 이 때, χ는 다음과 같이 정의된다.

$$\chi(W) = \begin{cases} 1, & W\text{가 참일 때} \\ 0, & W\text{가 거짓일 때} \end{cases}$$

이와 같은 세 가지 구성요소를 가진 GPD방법에 의해

$$\Lambda_{t+1} = \Lambda_t - \varepsilon_t \nabla\,\ell(X_t\ ;\ \Lambda)|_{\Lambda=\Lambda_t}$$

을 따라 수렴하는 파라미터 Λ를 구하는 것으로서 결국, GPD에 의한 discriminant 학습은 다음과 같은 평균 오차율을 최소로 하는 파라미터 Λ를 찾아가는 방법이다.

$$\Lambda = \frac{1}{N}\sum_{i=1}^{X_T}\sum_{k=1}^{K} l_k(X_i\ ;\ \Lambda)\ \chi(X_i\in C_k).$$

여기에서 ε_t는 매우 작은 수로서 이 논문의 실험에서는 $\varepsilon_t = 1 - t/75$를 사용하였다.

클래스 C_j, $(1\le j\le K)$의 HMM 파라미터에 대한 미분 $\nabla l_k(X\ ;\ \Lambda)$는

$$\nabla_{\Lambda_j}[l_k(X\ ;\ \Lambda)] \frac{\partial \ell_k}{\partial d_k}\frac{\partial d_k}{\partial g_j}\nabla_{\Lambda_j}[g_j(X\ ;\ \Lambda)]$$

로 유도되며, 각각의 편미분은

$$\frac{\partial \ell_k}{\partial d_k} = a\cdot \ell_k(1-\ell_k), \tag{1}$$

$$\frac{\partial d_k}{\partial g_j} = \begin{cases} -1, & j=k\text{ 일 때}, \\ \dfrac{\exp[\eta g_j(X\ ;\ \Lambda)]}{\sum_{n,n\neq k}\exp[\eta g_n(X\ ;\ \Lambda)}, & j\neq k\text{ 일 때} \end{cases}$$

이다. HMM 파라미터에 대한 $\nabla_{\Lambda_j}[g_j(X\ ;\ \Lambda)]$는 클래스 j에서 전이확률 행렬의 원소 $a_{q_t,\,q_{t+1}}$, $1\le t\le N$에 대한 편미분과 상태별 관찰 확률 $b_q(X)$의 계산에 필요한 평균(μ), 분산(σ), cluster 계수 c 대한 편미분으로 나타난다.[2][5]

식 (1)에서 d_k가 b에서 조금만 떨어져 있어도 $\frac{\partial \ell_k}{\partial d_k}$의 값은 지수 함수 exp의 특성으로 인해 0에 가까운 값을 갖게된다. 따라서 b값을 설정하게 한 입력 데이터 X와 극히 제한된 수의 입력 데이터를 제외하고는 GPD에 의한 파라미터 보정에 영향을 주지 못한다.

Ⅳ. 실험결과

한국어 고립 숫자음 음성인식 실험을 위해 30대 남자 5인의 화자가 '일', '이', …, '구'의 숫자음을 각 15회씩, 전화기를 통해 발음한 것을 모아 개별 단어에 대해

모두 75개의 음성데이터를 HMM 학습 데이터로 사용하였다. 인식 실험을 위한 데이터는 같은 숫자음을 20인이 5회씩 전화기를 통해 발음한 것을 사용하였다.

또한 연속 숫자음 인식 실험을 위해서는 5인의 화자가

'삼일'(31), '오일'(51), '육일'(61), '칠일'(71),
'삼이'(32), '오이'(52), '육이'(62), '칠이'(72),
'삼칠'(37), '오칠'(57), '육칠'(67), '칠칠'(77),

과 같은 네 가지의 발음이 나온다. '일'에 대한 HMM 학습은 '일', '밀', '길', '밀'을 발음한 것을 잘라 각 20개씩, 모두 80개의 데이터을 4회씩 발음한 전화음성 데이터를 사용하였다. 연속 숫자음에서는 앞에서 기술한대로 두 숫자음의 뒤에 나오는 발음이 '삼일'(31)의 경우, '일', '밀', '길', '밀'를 한꺼번에 사용하였다. 마찬가지로 '이'에 대해서도 32(사미), 52(오이), 62(유기), 72(치리)에서 나오는 '미', '이', '기', '리'를 잘라 20개씩, 모두 80개의 데이터를 한꺼번에 사용하였다. '칠'에 대한 발음도 같은 방법으로 사용하였다.

음성 인식에 필요한 특징벡터는 0.015초의 구간으로 잘라 12차원 cepstrum, 12차원 delta-cepstrum의 모두 24차원 벡터로 추출하였다.

각 숫자음에 대해서 '일', '삼'과 같이 받침이 있는 숫자음은 4개의 상태를 가진 모델로, '이', '오'와 같이 받침이 없는 숫자음은 3개의 상태를 가진 모델로 설정하였다. k-means 알고리즘[3]을 이용하여 상태별로 군집수는 모두 3개로 하였으며, HMM 모델은 연속형 left-to-right 모델로 설정하였다. 공분산 행렬(covariance matrix)은 대각형을 채용하였다.

실험결과 고립 숫자음의 경우에는 HMM 알고리즘의 효율성이 충분히 드러나 일반적으로 전화를 통한 1음절에서 4음절 사이의 단어에 대해서 70개에서 100개 정도의 음성 학습 데이터를 가지고, 100여개의 테스트 데이터에서 95%이상의 높은 인식률을 보여 주었다.

연속 숫자음의 경우에서 '영'과 '공'의 경우는 20개의 테스트 데이터에서 100% 인식되어 문제점이 없으며, '공'과 '구'의 경우에도 실험결과 뚜렷이 구별되었다. 그러나 '일'과 '이'와 발음이 비슷한 '칠'의 경우는 인식률이 매우 저조하였다. 발음이 비슷한 '일', '이', '칠' 세 숫자음에 국한한 인식률은 표 1과 같다

표 1. HMM을 활용한 인식률
Table 1. Recognition rate using HMM.

숫자음	'일'	'이'	'칠'
실험데이터의 수	56	56	56
인식률	100%	100%	78.57%

표 1에서와 같이 '칠'에 대한 인식률이 저조한 것은 '칠'의 발음이 '일'로 인식되고 있기 때문이다. 이는 '일'에 대한 HMM 학습이 '일', '밀', '길', '밀'을 포함하도록 매우 폭넓게 이루어졌기 때문이라 생각된다.

'칠'에 대한 낮은 인식률을 높이기 위해 discriminant 학습을 이용하였다. Discriminant 학습에 사용한 데이터는 '일'로 인식되는 '칠'의 발음 가운데 d_k의 값이 작고(이는 두 발음이 비슷하여 최대유사정도 값의 차이가 작게 나타남을 의미한다), 모든 '칠'에 대한 음성이 갖는 최대유사정도 값의 평균에 가까운 것을 2개 정도로 하여도 충분하였다.

Discriminant 학습을 이용하여 HMM 파라미터의 보정을 한 후, 다시 인식률을 실험한 결과는 표 2와 같다. 실험 결과 '칠'에 대한 인식률은 약 20% 향상된 것으로 나타났다.

표 2. Discriminant 학습을 활용한 인식률
Table 2. Recognition rate using discriminant learning.

숫자음	'일'	'이'	'칠'
실험데이터의 수	56	56	56
인식률	100%	100%	98.21%

V. 결 론

HMM 방법은 충분한 학습 데이타가 주어지면 높은 인식률을 가져올 수 있다. 그러나 실 세계에서 충분한 데이터를 수집하고 가공해야 하는 어려운 경우가 많으며 계산상의 복잡도가 높아지는 단점이 있다. 특히 연속 숫자음의 경우에는 유사한 숫자음에 대한 분별력이 떨어져 인식율의 저하를 가져온다.

유사한 연속 숫자음 인식에서 발음이 뚜렷이 구별되는 다른 숫자음에 대해서는 문제가 없었으나 '일'과 '이'와 발음이 비슷한 '칠'의 경우에 HMM 방법의 인식률이 떨어졌다. 그것은 '일'에 대해 너무 포괄적인 학습이

이루어져 '칠'에 대한 인식률 저하를 초래하였다. 즉, '칠'에 대한 실험데이터들이 '일'로 인식되는 결과를 낳았다. 이러한 결과에 대한 개선책으로 discriminant 학습을 적용하였다. '칠'의 경우 discriminant 학습 전과 학습 후의 인식률을 비교한 결과 20%의 인식률 향상을 가져와 1음절 인식으로서도 다른 숫자음과 충분히 분별력을 가질 수 있었다. 또한 적은 계산량과 적은 데이터에서도 인식률을 향상시켰다. 따라서 비슷한 발음으로 인해 인식률의 저하되거나, 포괄적인 학습으로 인하여 인식률이 낮은 경우에 discriminant 학습을 사용해서 높은 인식률을 얻을 수 있다.

따라서 다양하게 발음되는 숫자음에서도 HMM과 discriminant 학습을 이용해 하나의 학습자료로 만들면 일반적인 빔 탐색에 의해서도 실 세계에서 원하는 속도와 인식률을 가진 연속 전화 숫자음 인식 시스템을 구축할 수 있다.

참 고 문 헌

[1] Jiqing Han, Munsung Han, Gyu-Bong Park, Jeongue Park and Wen Gao, "Discriminative earning of additive noise and channel distortions for robust speech recognition", in Proc. ICASSP-98 pp.81-84.

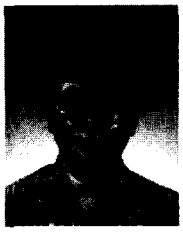
[2] Biing-Hwang Juang, Wu Chou and Chin-Hui Lee, "Minimum classification error rate methods for speech recognition", IEEE Trans. Speech and Audio Process., 5 (3), pp.257-265, 1997, 5

[3] Biing-Hwang Juang and L. Rabiner, "The segmental K-means algorithm for estimating parameters of hidden Markov models", IEEE Trans. Audio Speech Signal Process., 38, pp. 1639-1641, 1990.

[4] Chin-Hui Lee and Lawrence R. Rabiner, "A frame-synchronous Network Search Algorithm for connected word recognition", IEEE Trans. Acoust. Speech Signal Processing, 37, pp. 1649-1658, 1989, 11.

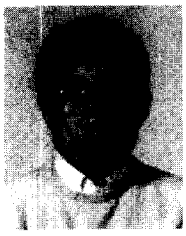
[5] Chi-Shi Liu, Chin-Hui Lee, Wu Chou, Biing-Hwang Juang and Aaron E. Rosenberg, "A study on minimum error discriminative training for speaker recognition", J. Acoust. Soc Am., 97 (1), pp.637-648, 1995, 1.

저 자 소 개



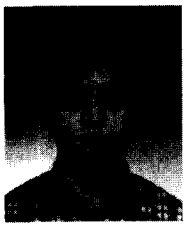
韓 汶 星(正會員)

1977 서울대학교 자연과학대학 수학과 졸업(학사). 1977~1979 전국경제인연합회 경제기술조사부. 1980~1981 한국 IBM. 1981~1988 Indiana 대학교 Computer Science 박사과정. 1989~현재 한국전자통신연구원 책임연구원. 주관심분야 : 음성인식, 신경망, AI



崔 完 洙(正會員)

1977 고려대학교 전자공학과 졸업(학사). 1983~1988 미네소타주립대학교 Computer Science 졸업(석사, 박사수료). 1993 고려대학교 전자공학과 졸업(박사). 1993~현재 대림대학 전자정보통신과 조교수. 주관심분야 : 음성인식, CAD



權 賢 稷(正會員)

1992 서울대 자연과학대학 수학과 졸업(학사). 1992~2000 한국과학기술원 수학과 졸업(석사, 박사). 2000~현재 한국과학기술원 수학과 Post Doc. 주관심분야 : 음성인식, 수치해석