음성인식기의 변별력있는 학습 알고리즘들

나 경민°, 전 범 기, 이 주 헌, 안 수 길
서울대학교 대학원 전자공학과

Discriminative Training Algorithms for Speech Recognizers

KyungMin NA, BumKi JEON, JooHun LEE and SouGuil ANN
Dept. of Electronics Eng., Seoul National University

초 록

기존의 음성인식기들은 일반적으로 간단하면서도 성능이 우수한 계층별 학습에의해서 설계된다. 계층별 학습은 통계적 패턴인식에서의 ML (Maximum Likelihood) 추정기법처럼 모델간의 독립성이 보장되고 무한한 양의 학습데이타가 주어진다는 가정에 기초하고 있다. 그러나, 대상어휘집합에 음운학적으로 유사한 어휘가 많이 포함되어 있는 인식문제에 있어서는 모델간의 독립성이 보장되지 못하고, 실제 주어지는 학습데이타의 양도 제한되므로 기존의 학습알고리즘에는 한계가 있다. 따라서 본 논문에서는 그러한 가정상의 문제점으로 생기는 인식기의 성능저하를 개선할 수 있는 변별력있는 학습알고리즘들을 검토하고 그의 일반적인 접근방법들에 대해서 논의한다.

I. 서 론

현재까지 제안된 음성인식모델들은 크게 DTW (Dynamic Time Warping) [1], HMM (Hidden Markov Models) [2], 신경회로망모델 [3-5] 로 나눌 수 있다. 각 모델들은 그 특성에 맞는 학습알고리즘을 갖고 있으며, 최근에는 고립단어인식에서 연결단어나 연속음성인식분야로 그 적용이 확대되고 있다. 그러한 인식기들은 기본적으로 모델파라메터, 학습데이타, 결정규칙으로 구성되고, 학습알고리즘이란 주어진 학습데이타를 이용해서 최적의 모델파라메터를 추정하는 방법으로 볼 수 있다. DTW에서 참조패턴을 얻는데 사용되는 여러가지 클러스터링 기법들, HMM의 ML 추정기법인 Baum-Welch 알고리즘 등은 대표적인 학습알고리즘이다. 그러한 기존의 학습알고리즘들은 대부분 각 계층의 모델들이 독립이고 무한한 양의 학습데이타가 주어진다는 가정에 기초하고 있다 [6][7]. 그러나, 실제의 경우에 그러한 가정이 성립되지 못하며 그로인한 인식기의 성능저하는 불가피하다. 따라서, 그와같은 추정오차에 의한 인식율의 저하를 최대한 개선하기위해서 변별력있는 학습알고리즘이 요구된다 [8-15].

최근에 신경회로망이 음성인식에 적용되면서 변별력있는 학습의 가능성을 제시했다. 다층신경회로망을 학습시키는 오차역전파 학습알고리즘이나 LVQ (Learning Vector Quantization) 등이 그 대표적인 예이다. 그러나, 신경회로망 본래의 정적인 구조때문에 동적인 신호의 모델링에 한계를 보였고 그에 따라서 최근에는 여러가지 동적인 기법과의 결합이나 회귀하는 신경회로망(recurrent neural networks)에 관한 연구가 활발하다.

또한, 신경회로망의 학습에서와 같은 변별력있는 학습기법을 다른 인식모델로 확장하여 적용하고자 하는 연구도 최근에 관심을 모으고 있다. GPD (Generalized Probabilistic Descent) 방법은 인식오차를 소무딩함수의 형태로 근사하여 확률적인 의미에서 최소화시키는 기법으로 이미 DTW, HMM, NPM (Neural Prediction Model) 등에 적용되었다 [1][9-13]. 또한 최근에 통계적 패턴인식분야의 최소오차율분류를 근사하는 최소오차율 학습알고리즘이 제안되었다 [14][15]. GPD 방법과 최소오차율 학습알고리즘은 모두 인식오차를 최소화시키기 위한 알고리즘으로 근사적으로 Bayes 분류기에 가까운 성능을 얻기 위해서 개발되었다. 두 알고리즘간의 유사한 성질을 분석하고 그 수식화에 있어서 좀더 일반적인 방법론을 제시한다. 본 논문에서는 전통적인 통계적 패턴분류기법에 기초해서 두 알고리즘을 비교하고 그로부터 변별력있는 학습알고리즘에 대한 일반적인 방법론을 제시한다.

II. 전통적인 패턴분류기법

2.1 Bayes decision theory

대부분의 인식기에 있어서, M개의 계층 C_m , $m=1,2,\cdots,M$ 에 대해서 어떤 k차원의 학습데이타 x_n의 계층을 알고 있고, 주어진 학습데이타의 집합을 $\Omega=\{x_1,x_2,\cdots,x_N\}$라 하면, 인식기의 최적학습은 주어진 학습데이타집합 Ω에 기초해서 최적의 모델파라메터집합 Λ와 그에 합당한 결정규칙을 찾는 것이 된다.

패턴분류에 있어서 잘 알려진 통계적 방법인 Bayes decision theory에 의하면 주어진 파라메터집합 Λ에 대해서 사후확률(a posteriori probability) $P_\Lambda(C_m|x)$를 완전히 알고 있을때, 최소오차확률(minimum probability of error)을 얻을 수 있는 Bayes 결정규칙은 식 (1)과 같이 주어진다.

$$C(x)=C_i \ , \ if \ P_\Lambda(C_i|x) = \max_j P_\Lambda(C_j|x) \qquad (1)$$

식 (1)에서 $C(\cdot)$은 분류연산(classification operation)을 나타낸다. 그러나, 일반적으로 사후확률의 정확한 형태를 알기 어려우므로 Bayes 규칙을 이용해서 식 (1)의 $P_\Lambda(C_i|x)$ 대신에 사전확률(a priori probability) $P(C_i)$과 계층조건부확률(class-conditional probability) $p_\Lambda(x|C_i)$의 곱을 결정규칙으로 사용한다. 그러면, 사전확률들이 모두 같다는 가정하에서 결정문제는 결국 $p_\Lambda(x|C_i)$에 의해서만 기술되고, 그에 따라서 사후확률 $P_\Lambda(C_i|x)$ 대신 $p_\Lambda(x|C_i)$를 추정하여 분류기를 설계하는 것을 일반적으로 ML (Maximum Likelihood) 학습방법이라 한다. 즉, 계층 i의 모델파라메터들을 λ_i라 하고 각 λ_i가 서로 독립이라고 가정하면 전체파라메터집합은 $\Lambda=\{\lambda_1\lambda_2\cdots,\lambda_M\}$이 되고 그에 따라서 각 계층의 설계표본들로부터 해당되는 모델파라메터들을 독립적으로 추정할 수 있다. 따라서, ML 개념의 학습시에는 각 계층의 모델파라메터집합들이 서로 독립이고 무한한 양의 학습데이타가 주어진다는 가정이 필요하다. 그리고, Bayes decision theory 자체도 분류문제가 확률척도로 기술가능하고, 관련된 확률척도가 파라메터집합 Λ의 함수로서 그 형태가 알려져있으며, 주어지는 설계표본집합에 대해서 그 모델파라메터들을 추정할 수 있는 효과적인 방법이 존재한다는 여러가지 가정들에 기초하고 있다.

2.2 최소오차율분류(minimum-error-rate classification)

어떤 입력 x의 실제 계층이 C_j인 경우에 그 입력을 C_i로 분류하는 어떤 행위를 α_i라고 정의하고, α_i라는 행위를 함으로써 발생하는 손실(loss)을 $\lambda(\alpha_i|C_j)$라 정의하자. 즉, 어떤 x를 관찰하고 α_i라는 행위를 했는데 실제로는 x가 C_j에 포함되면, $\lambda(\alpha_i|C_j)$의 손실이 일어났다고 할 수 있다. 그러면, conditional risk라고 알려진 손실의 기대치 $R(\alpha_i|x)$는 다음과 같이 정의된다.

$$R(\alpha_i|x)=\sum_{j=1}^{M}\lambda(\alpha_i|C_j)P_\Lambda(C_j|x) \qquad (2)$$

최소오차율분류를 위해서는 특별히 대칭손실함수 혹은 one-zero 손실함수라고 부르는 다음과 같은 손실함수가 사용된다.

$$\lambda(\alpha_i|C_j)=\begin{cases}0 & i=j\\ 1 & i\neq j\end{cases} \quad i,j = 1,2,\cdots,M \qquad (3)$$

식 (3)의 손실함수는 정확한 결정에 대해서는 아무런 손실도 주지 않고, 틀린 결정에 대해서는 1의 손실을 준다. 이 손실함수에 대한 risk는 다음과 같이 평균오차확률(average probability of error)이 된다.

$$\begin{aligned}R(\alpha_i|x) &= \sum_{j=1}^{M}\lambda(\alpha_i|C_j)P_\Lambda(C_j|x)\\ &= 1-P_\Lambda(C_i|x)\end{aligned} \qquad (4)$$

Bayes 결정규칙은 식 (4)의 conditional risk를 최소화시키는 것이다. 따라서, 최소오차율분류를 위해서는 사후확률 $P_\Lambda(C_i|x)$를 최대화시키는 계층 C_i를 택해야 한다.

2.3 분별함수를 사용하는 분류기

확률척도에 대한 대안으로 분별함수(discriminant functions)를 사용하여 분류기를 설계하는 경우가 많다. 일반적으로 적당한 분별함수집합 $g_m(x;\Lambda)$, $m=1,2,\cdots,M$이 주어지면, 결정규칙은 식 (5)와 같이 주어진다.

$$C(x)=C_i \ , \ if \ g_\Lambda(x;\Lambda) = \max_j g_\Lambda(x;\Lambda) \qquad (5)$$

여기서 최적분류기를 설계하는 문제는 sample risk를 최소화시키는 분별함수의 최적파라메터집합을 구하는 것이다. 그러나, 설계표본을 분류하는데서 발생하는 평균비용으로

정의되는 sample risk가 일반적으로 불연속적인 함수이기 때문에 경사법에 의한 최적화에 어려움이 있기때문에 실제로는 다루기 쉬운 다른 스칼라 비용함수를 도입하여 최소화시키는 목적함수로 사용한다. 잘 알려진 스칼라 비용함수로는 perceptron criterion과 summed squared error criterion이 있다. 이러한 스칼라 비용함수를 사용할 때의 문제점은 결정규칙이 함수의 형태로 스칼라 비용함수에 포함되어 있지 않고, 그러한 스칼라 비용함수와 최소오차율을 위한 비용함수가 일치되지 못하다는데 있다. 더우기 기존의 스칼라 비용함수를 이용한 학습은 ML 개념의 학습이어서 각 계층간의 결정경계(decision boundary)를 최적화시킬 수 없다는 문제점도 남아있다.

III. GPD 방법에 기초한 변별력있는 학습

확률적강하법(probabilistic descent method)은 S. Amari가 제안한 학습법으로 혼동되기 쉬운 두 모델간에 각각 강화학습과 반강화학습을 시키는 것을 주된 내용으로 하고 있다 [7]. 이것을 B. H. Juang 등이 일반화시켜서 모든 모델들을 학습에 고려하는 확장된 알고리즘으로 최근에 제안한 것이 GPD (Generalized Probabilistic Descent) 방법이다 [8-10]. GPD 방법은 최소분류오차수식화(minimum classification error formulation)를 통해서 인식오차의 수를 스무딩된 손실함수(loss function)의 형태로 근사하여 그 손실을 확률적으로 최소화시키는 방법으로 최근에 DTW, HMM, NPM 등에 적용되었다 [1][9-13].

3.1 최소분류오차수식화

이 수식화는 3 단계의 과정을 거쳐서 인식오차의 손실함수를 근사하도록 하는 과정이다. 경사법의 적용이 쉽도록 모든 수식화가 모델파라메터에 연속인 형태로 진행된다.

(1) 음이 아닌 분별함수(discriminant function) $g_i(x;\Lambda)$

$$C(x) = C^i \quad if\ g_i(x;\Lambda) = \max_j g_j(x;\Lambda) \tag{6}$$

정의된 분별함수가 인식의 결정규칙으로 사용된다. 이와같이 결정규칙이 학습에 참여하게 된다는 점에서 기존의 방법보다 나은 결과를 기대할 수 있다.

(2) 오분류척도(misclassification measure) $d_k(x;\Lambda)$

$$d_k(x;\Lambda) = -g_k(x;\Lambda) + \left\{ \frac{1}{M-1} \sum_{j \neq k} g_j(x;\Lambda)^\xi \right\}^{\frac{1}{\xi}} \tag{7}$$

이 오분류함수의 도입이 기존의 방법과 가장 다른점이다. 오분류함수의 값이 크면 클수록 더욱 오인식될 가능성이 높고 반대로 작으면 작을수록 잘 인식이되는 것을 나타낸다. 식 (7)의 ξ를 적당히 조절함으로서 각 모델들간의 최적화에 참여하는 정도들이 조정된다. 극단적인 경우에 $\xi \to \infty$이면 수식이 다음과 같이 된다.

$$d_k(x;\Lambda) = -g_k(x;\Lambda) + g_i(x;\Lambda) \tag{8}$$

식 (8)의 경우가 바로 S. Amari가 제안한 확률적경사법과 같아지게된다. 즉, 가장 혼동될 가능성이 높은 계층 C^i의 모델만이 학습에 참가하게된다.

(3) 오인식율을 근사하는 비용함수(cost function) 혹은 손실함수(loss function) $l_k(d_k)$

$$l_k(d_k(x;\Lambda)) = \frac{1}{1+e^{-\alpha d_k}} \tag{9}$$

여러 종류의 비용함수를 사용할 수 있으나 식 (9)의 sigmoid 함수가 가장 널리 쓰인다. 이것은 식 (3)의 손실함수를 일반화시켜 근사한 것으로 볼 수 있다. 이 비용함수는 오차확률의 근사로 해석할 수 있다.

3.2 확률적강하법(probabilistic descent method)

확률적강하법을 사용하여 위에서 정의한 손실함수의 기대치를 확률적으로 최소화시키면 원하는 알고리즘을 얻을 수 있다.

$$L(\Lambda) = E l_k(x;\Lambda) \tag{10}$$

$$\Lambda_{t+1} = \Lambda_t + \delta\Lambda_t \quad where\ \delta\Lambda_t = -\eta_t U \nabla l_k \tag{11}$$

위에서 U는 positive-definite 행렬이고, 학습율 η_t는 $\sum_{t=1}^{\infty} \eta_t \to \infty$와 $\sum_{t=1}^{\infty} \eta_t^2 < \infty$를 만족하도록 시간에 따라 줄어나간다 [8].

3.3 GPD 방법에 의한 일반적인 학습알고리즘

위에서 정의된 수식과 확률적강하법에 의해서 일

반적인 학습알고리즘을 유도하면 다음과 같다.

$$(\lambda_m)_{t+1} = (\lambda_m)_t + \eta_t \alpha l_m (1.0 - l_m) \frac{\partial g_m}{\partial \lambda_m} \quad \text{for } x \in C_m, \tag{12.a}$$

$$(\lambda_l)_{t+1} = (\lambda_l)_t - \eta_t \alpha l_m (1.0 - l_m) \frac{\partial g_l}{\partial \lambda_l} \quad \text{for all } l \neq m. \tag{12.b}$$

위의 식에서 알 수 있듯이 (12.a)는 비용 $l_m(1.0-l_m)$이 가중된 경사강하법이고 (12.b)는 같은 비용이 가중된 경사상승법이다. 전자는 강화학습이고 후자는 반강화학습으로 볼 수 있다. 참여하는 비용 $l_m(1.0-l_m)$은 quadratic 형태로 l_m = 0.5에서 최대가 되고 그 주위에서는 급격히 감소되는 형태이다. l_m = 0.5는 식 (9)로부터 d_m = 0를 의미하고 그것은 현재의 입력이 가장 혼동되는 위치, 즉 결정경계면에 위치함을 나타낸다. 즉, 혼동될 가능성이 높으면 높을수록 더욱 많은 양의 학습을 하게된다. 또한, 이미 잘 분류하고 있거나 전혀 개선의 여지가 없는 outlier들에 대해서는 비용이 0에 가까워져서 학습에 주는 영향을 줄일 수 있다.

IV. 최소오차율 학습알고리즘

분별함수를 기반으로하는 분류기의 출력은 확률값이 아니므로 Bayes decision theory에 근거한 최소오차율 분류가 불가능하다. 그러나, 분류기의 출력들을 정규화시켜서 식 (1)의 사후확률의 추정치로 사용하면 근사적으로 최소오차율분류를 위한 학습알고리즘을 얻을 수 있다. 패턴분류기를 사후확률추정기로 보고 어떤 입력패턴에 대해 그에 대응하는 계층의 사후확률을 최대로하도록 학습시키는 것이다 [14][15].

그러므로, 먼저 분류기의 출력을 효과적으로 정규화시키는 과정이 필요하다. 이 과정은 여러가지 방법으로 수식화가 가능한데 예를 들어서 출력이 항상 영보다 크거나 같다면 다음과 같은 수식화로 가능하다. scaling factor를 α라 하면,

$$f_m(x;\Lambda) = \frac{(g_m(x;\Lambda))^\alpha}{\sum_{i=1}^{M}(g_i(x;\Lambda))^\alpha}. \tag{13}$$

분류기의 출력의 부호에 상관 없이 사용하려면 다음과 같은 정규화된 지수형태인 "softmax" 함수가 적절하다 [2].

$$f_m(x;\Lambda) = \frac{\exp(\alpha g_m(x;\Lambda))}{\sum_{i=1}^{M}\exp(\alpha g_i(x;\Lambda))} \tag{14}$$

각각의 설계표본들이 서로 독립이라고 가정하면, 정규화 과정을 거쳐서 다음과 같은 전체 비용함수를 정의할 수 있다.

$$L(\Lambda) = \prod_{i=1}^{M}\prod_{x\in C_i} f_i(x;\Lambda) \tag{15}$$

식 (15)의 비용함수를 최대화시키는 것이 설계의 목표이다. 그러나, 비용함수가 식 (14)의 곱으로 표현되어 있으므로 일반적인 경사법을 적용하기 어렵다. 따라서, 확률적강하법을 적용하여 분류기를 설계한다. 이 방법에 의하면 개별적인 사후확률추정치인 $f_m(x;\Lambda)$들에 대해서만 강하법을 적용한다.

$$\Lambda_{t+1} = \Lambda_t - \eta_t U \nabla(-f_i(x;\Lambda)) \tag{16}$$

또한, 강하법은 비용함수를 최소화시키는 방법이므로 $f_i(x;\Lambda)$ 앞에 (-) 부호를 붙여서 수식화했는데 그것은 경사상승법과 같은 효과를 내서 식 (15)을 최대화시키는 목적에 합당하다.

또한, 기존의 강하법을 사용하기 위해서 식 (15)의 자연로그를 취하고 음수부호를 붙여서 새로운 비용함수를 정의하면 다음과 같다.

$$L'(\Lambda) = -\ln L(\Lambda) = -\sum_{i=1}^{M}\sum_{x\in C_i} \ln f_i(x;\Lambda) \tag{17}$$

식 (10)에 경사강하법을 적용하면 다음과 같은 학습식을 얻을 수 있다.

$$\nabla L'(\Lambda) = \sum_{i=1}^{M}\sum_{x\in C_i} \nabla(-\ln f_i(x;\Lambda)) \tag{18}$$

$$\Lambda_{t+1} = \Lambda_t - \eta\nabla(-\ln f_i(x;\Lambda)) \tag{19}$$

제안하는 알고리즘의 특성을 더 자세히 파악하기 위해서 식 (14)을 사용하여 구체적인 학습알고리즘을 유도하겠다. 다른 정규화 방법에 대해서도 쉽게 유도할 수 있으므로 분류기의 특성에 따라서 설계자가 사용할 식을 결정하면 된다. 먼저 식 (16)에 의한 학습알고리즘은 다음과 같다.

$$(\lambda_m)_{t+1} = (\lambda_m)_t + \eta_t \alpha f_m (1.0 - f_m) \frac{\partial g_m}{\partial \lambda_m} \quad \text{for } x \in C_m, \tag{20.a}$$

$$(\lambda_l)_{t+1} = (\lambda_l)_t - \eta_t \alpha f_m f_l \frac{\partial g_l}{\partial \lambda_l} \quad \text{for all } l \neq m. \tag{20.b}$$

또한, 식 (17)에 의한 학습알고리즘도 다음과 같이 쉽게 유도할 수 있다.

$$(\lambda_m)_{t+1}=(\lambda_m)_t+\eta\alpha(1.0-f_m)\frac{\partial g_m}{\partial \lambda_m} \quad \text{for } x\in C_m, \tag{21.a}$$

$$(\lambda_l)_{t+1}=(\lambda_l)_t-\eta\alpha f_l\frac{\partial g_l}{\partial \lambda_l} \quad \text{for all } l\neq m. \tag{21.b}$$

편의상 $f_l=f_l(x;\Lambda)$이라고 표현했다. 식 (20)과 (21)의 기본적인 차이는 학습에 참여하는 비용에 있다. 식 (20)에서는 학습에 참여하는 비용이 $f_m(1.0-f_m)$과 f_mf_l인데 비하여 식 (21)에서는 각각 $(1.0-f_m)$과 f_l이다. 학습에 의해서 f_m이 1에 가까워지면 갈수록 f_l은 0에 가까워진다. 학습 초기에는 같은 f_m과 f_l에 대해서 식 (20)의 비용보다 식 (21)의 비용이 크므로 비용함수가 빨리 줄어들고, 학습이 어느정도 진행된 후에는 $(1.0-f_m)$과 f_l이 모두 0에 가까워짐으로 적어도 국부적으로 최적인 해를 얻을 수 있다. 그러나, 식 (21)의 경우에는 outlier들과 이미 잘 분류하고 있는 경우의 학습데이타들이 미치는 영향을 제거하기 위해서 f_m = 0.5를 기준으로 어느 정도의 창(window)을 취해서 그 안의 경우만 학습시킴으로서 계산량도 줄이고 안정적인 결과를 얻을 수 있다. 따라서, 그러한 창을 도입하면 식 (21)의 학습알고리즘이 식 (20)보다 빨리 안정적으로 수렴함을 알 수 있다.

V. GPD 방법에 대한 재해석과 비교

5.1 GPD 방법에 대한 최소오차율분류 적용

이제 GPD 방법을 최소오차율분류의 범위에서 재해석하여 보자. GPD 방법에서 정의되는 손실함수는 오차의 확률을 스무딩시켜서 모델링한다. 그러나, Bayes decision theory에 따르면 최소오차율분류는 손실함수의 risk (4)를 최소화하는 것이다. 그러한 의미에서 식 (9)에 정의된 함수는 식 (4)를 근사한다고 볼 수 있다. 따라서, 다음과 같은 관계를 얻을 수 있다.

$$P_\Lambda(C^t) \approx 1-R(a_t|x) \approx 1-l_t(x;\Lambda) \tag{22}$$

이러한 관점에서 식 (15)를 이용하면 위에서 얻은 GPD에 의한 학습식 (12)를 얻을 수 있다. 또한, 식 (17)과 같은 맥락에서 다음과 같은 새로운 알고리즘의 유도도 가능하다.

$$(\lambda_m)_{t+1}=(\lambda_m)_t+\eta\alpha l_m\frac{\partial g_m}{\partial \lambda_m} \quad \text{for } x\in C_m, \tag{23.a}$$

$$(\lambda_l)_{t+1}=(\lambda_l)_t-\eta\alpha l_m\frac{\partial g_l}{\partial \lambda_l} \quad \text{for all } l\neq m. \tag{23.b}$$

위의 식에도 적절한 창을 도입하면 식 (12)보다 빨리 안정적으로 수렴한다는 것을 알 수 있다.

5.2 알고리즘 비교

이상에서 최종적으로 식 (12), (20), (21), (23)의 네가지 알고리즘을 얻었다. 식 (12)와 (20)으로 부터 얻어진 결과로부터 식 (21)과 (23)에 적절한 창을 도입하는 것이 계산량이나 수렴도에서 유리함을 알 수 있다. 또한, 모든 알고리즘이 강화학습과 반강화학습으로 구성되어 변별력을 향상시키는 목적에 부합됨을 알 수 있다. 또한, 각 학습에 참여하는 비용의 개념도 학습의 목적에 맞는 형태임을 알 수 있다. 식 (12)와 (23)은 모델간의 출력에 일종의 거리개념을 도입하였고, 식 (20)과 (21)에는 정규화기법이 사용되었다고 볼 수 있다. 따라서, 이러한 종류의 알고리즘의 개발에 있어서 출력들간의 유사도를 재는 척도가 중요함을 알 수 있다. 또한, 이러한 알고리즘들이 근사적으로 오차의 확률을 모델링함으로 그에 포함되는 스케일링 요소가 실제의 적용에서 매우 까다로운 점으로 남게된다.

VI. 실험및 결론

위에서 검토한 알고리즘들을 HCNN (Hidden Control Neural Network)을 이용한 한국어 단음절 "가,나,다,라,마,바,사,아,자"를 인식하는데 적용한 결과 평균적으로 기존 알고리즘에서 발생된 오인식 갯수의 약 20 % 가량이 감소되었다. 실험에 사용된 화자는 총 20인으로 각 2회씩 발성하여 각각 학습과 실험에 사용했다. 수렴속도면에서도 예측한대로 식 (12)보다는 (23)이, 식 (20)보다는 (21)이 빨리 수렴하고 결과도 약간 좋았다.

위에서 유도된 알고리즘들은 다른 음성인식기에 대해서도 적용이 가능하며 실제로는 다른 영역의 패턴인식기에도 적용이 가능한 일반적인 알고리즘이다. 앞으로 여러가지 스케일링 요소에 대한 특성과 다른 모델에의 적용 등의 과제가 남아 있으며, 모델간의 유사도에 대한 새로운 척도에 대한 연구도 계속될 것이다.

표 1. 인식율 비교

Table 1. Comparison of recognition rates

데이타	기존방법	식 (12)	식 (23)	식 (20)	식 (21)
학습데이타	48.3 %	53.9 %	54.5 %	53.9 %	54.5 %
시험데이타	87.2 %	94.4 %	95.0 %	94.4 %	95.0 %

참 고 문 헌

[1] P. C. Chang and B. H. Juang, "Discriminative training of dynamic programming based speech recognizer," *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 2, pp. 135-143, 1993.

[2] H. Bourlard and C. J. Wellekens, "Links between Markov models and multilayer perceptrons," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 12, pp. 1167-1178, 1990.

[3] K. Iso and T. Watanabe, "Large vocabulary speech recognition using neural prediction model," *Proc. ICASSP-91*, pp. 57-60, 1991.

[4] J. Tebelskis, A. Waibel, B. Petek and O. Schmidbauer, "Continuous speech recognition using linked predictive neural network," *Proc. ICASSP-91*, pp. 61-64, 1991.

[5] E. Levin, "Hidden control neural architecture modeling of nonlinear time varying systems and its applications," *IEEE Trans. Neural Networks*, vol. 4, no. 1, pp. 109-116, 1993.

[6] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[7] S. Amari, "A theory of adaptive pattern classifiers," *IEEE Trans. Electronic Computers*, vol. EC-16, no. 3, pp. 299-307, 1967.

[8] B. H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Processing*, pp. 3043-3054, 1992.

[9] W. Chou, B. H. Juang and C. H. Lee, "Segmental GPD training of HMM based speech recognizer," *Proc. ICASSP-92*, pp. 473-476, 1992.

[10] P. C. Chang, S. H. Chen and B. H. Juang, "Discriminative analysis of distortion sequences in speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 3, pp. 326-333, 1993.

[11] K. M. Na, J. Y. Rheem and S. G. Ann, "A discriminative training algorithm for predictive neural network models," *Proc. ISCAS-94*, pp. 431-434, 1994.

[12] K. M. Na, J. Y. Rheem and S. G. Ann, "A GPD-based discriminative training algorithm for predictive neural network models," *Proc. WESTPRAC-94*, pp. 997-1002, 1994.

[13] K. M. Na, J. Y. Rheem and S. G. Ann, "Discriminative training of predictive neural network models," 한국음향학회지 13권 1E호, pp. 64-70, 1994.

[14] K. M. Na, J. Y. Rheem and S. G. Ann, "Minimum-error-rate training of predictive neural network models," *Proc. ICSLP-94*, pp. S26-1.1 - S26-1.4, 1994.

[15] 나 경민, 임 재열, 안 수길, "패턴분류기를 위한 최소오차율 학습알고리즘과 예측신경회로망모델에의 적용," 대한전자공학회지 11월호에 게재 예정.

본 연구는 통신개발연구원의 통신학술연구과제의 지원으로 이루어진 것임.