WHMM에 적용가능한 MCE/GPD 학습알고리듬에 관한 연구

Derivation of MCE/GPD Training Algorithm Applicable to Weighted Hidden Markov Models

최 홍 섭*
(Hong-Sub Choi*)

요 약

본 논문에서는 잘 알려진 변별학습 방법인 MCE/GPD방법을 WHMM에 적용시켜 그 학습알고리듬을 유도하고, E-set에 대한 실험결과를 제시한다. 유도된 알고리듬은 여러개의 혼동 가능한 클래스의 HMM에 대해서 학습이 가능하다는 점에서 기존에 제시된 적응학습 알고리듬의 자연스러운 일반화라 할 수 있다. E-set에 대한 인식실험 결과 학습에이터에 대해서 15%, 시험데이터에 대해서 12% 정도의 인식율 개선을 얻을 수 있었다.

ABSTRACT

This paper derives a new training alorithm for WHMM using the well-known MCE/GPD method with experimental results on the E-set. The derived algorithm generalizes the conventional adaptive training algorithm for WHMM, which means that HMMs of multiple competing classes can be trained at the same time. The recognition results on the E-set have shown about 15% and 12% improvement for training and test data, respectively.

I. 서 론

현재까지 음성인식 분야에서 가장 널리 사용되고 있는 인식모델은 HMM(hidden Markov models)이다[1]. 그것은 HMM이 화자 간의 스펙트럼 변화(spectral variations)와 발성속도 변화(speaking rate variations)를 동시에 확률적으로 모델링할 수 있기 때문이다. 일반적으로 HMM은 ML(maximum likelihood) 추정의 일종인 Baum-Welch 알고리듬으로 학습되고, 인식시에는 Viterbi 복호기가 사용된다. ML추정은 그림1과 같이 어떤 클래스 C_i에 속하는 학습데이터 집합 D_i를 그 클래스를 모델링하는 HMM의 파라메터 집합 λ_i, $i=1,\cdots,M$의 추정에만 사용하여 각 집합 내의 학습데이터들을 발생시킬 확률을 최대로 하도록 HMM의 파라메터 집합을 추정하는 것이다.

이론적으로 무한히 많은 학습데이터가 사용 가능하면 이러한 ML추정에 의해서 각 HMM을 Bayes 분류기에 가까운 성능을 갖도록 학습시킬 수 있다. 그러나, 실제의 많은 경우에는 제한된 수의 학습데이터만이 사용 가능하고, 그런 경우가 아니라도 되도록이면 적은 수의 학습데이터를 사용하여 효과적으로 HMM을 학습시키는 것이 바람직하다. 그러나, 기존의 ML학습은 전체 학습데이터 집합을 클래스 별로 나누고 각 클래스의 HMM파라메터를 해당되는 학습데이터 집합만을 이용해서 추정하므로, 주어진 학습데이터에 담긴 정보를 전부 이용하지 못하게 된다. 따라서 E-set과 같이 유사한 어휘로 구성된 인식문제의 경우에는 각 클래스간의 변별력(discrimination)이 현저히 떨어지게 된다. 이러한 점에서 각 학습데이터를 해당되는 클래스의 HMM파라메터 추정에만 사용하는 것보다는 모든 클래스의 HMM파라메터 추정에 사용하는 것이 각 클래스간의 변별력을 최대화할 수 있다. 즉, 그림2와 같이 전체 학습데이터 집합을 모든 크래스의 HMM파라메터 추정에 사용하는 것이다.

최근 몇년 동안에 이와 같은 음성인식기의 변별학습(discriminative training)에 대한 연구가 활발히 진행되고 있는 추세이다[2]-[10]. 초기에는 heuristic한 방법인 정정학습(corrective training)[2]과 정보이론에 기초한 MMI (maximum mutual information)[3], MDI(minimum discrimination information) 학습[4]이 발표되었다 고전적인 적응분류기(adaptive classifier) 학습이론과 Bayes 분류이론에 기초하여 변별학습에 대한 이론적인 연구를 시작한 것은 Bell 연구소의 B. H. Juang 등이다[5]. 그들은 Amari의 방법을 일반화시킨 MCE(minimum classification error) 수식화와 GPD(generalized probailistic descent) 알고리듬

*대진대학교 전자공학과
Dept. of Electronic Eng., Daejin University
접수일자:1996년 12월 16일

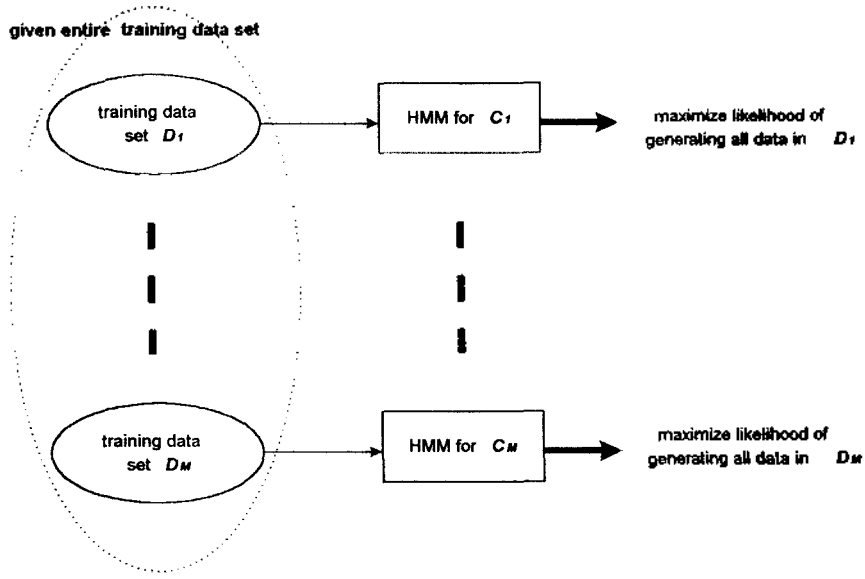


그림 1. ML추정에 의한 HMM의 학습
Fig 1. ML training of HMM

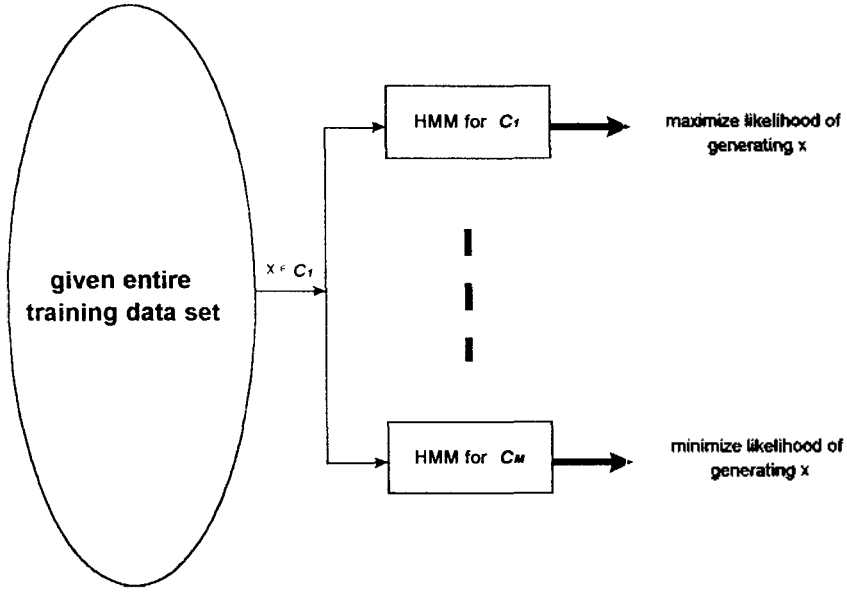


그림 2. 변별학습의 개념
Fig 2. Concept of discriminative training

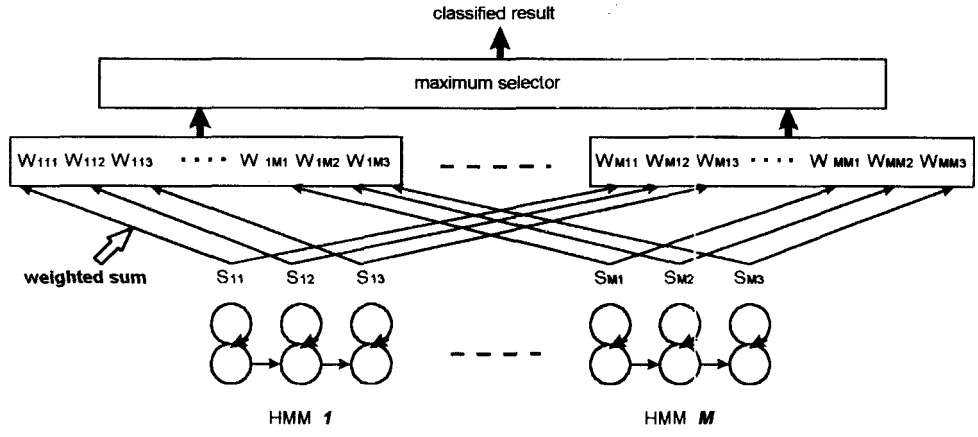
을 제안하였고, DTW(dynamic time warping), HMM, 신경회로망 등의 학습에 적용하였다. MCE 수식화는 주어진 학습데이터에 대한 모든 클래스의 HMM출력들에 대해서 평탄화된 거리개념의 오분류척도(misclassification measure)와 그 오분류척도의 값을 특정 범위내로 제한하는 손실함수(loss function)를 정의하여 오차율(error-rate)을 근사하는 수식화 과정이다. 그러한 MCE 과정을 거쳐 얻어지는 손실함수에 대해서 PD(probabilistic descent) 알고리듬을 적용하여 각 학습데이터에 대해서 pattern by pattern으로 손실함수를 최소화하도록 파라메터 값들을 조정하는 것이 MCE/GPD 알고리듬이다.

HMM의 변별학습에 관한 또 다른 연구방향으로 각 HMM의 파라메터는 고정시키고 각 HMM의 상태에서 발생되는 state-likelihood 값들에 변별에 따른 가중함수를 도입하는 방법이 있다. 그에 대한 연구는 Su와 Lee가 제안한 WHMM(weighted HMM)[7]에서 시작되었고, Wolfertetter와 Ruske는 그러한 방법을 연속음 인식에 확장하는 방법을 제안하였다[8]. 또한, 최근에는 다층 신경회로망(multilayer perceptron)을 사용하는 방법과 가중치의 합에 제한을 가해서 연속음과 고립단어 인식에 모두 적

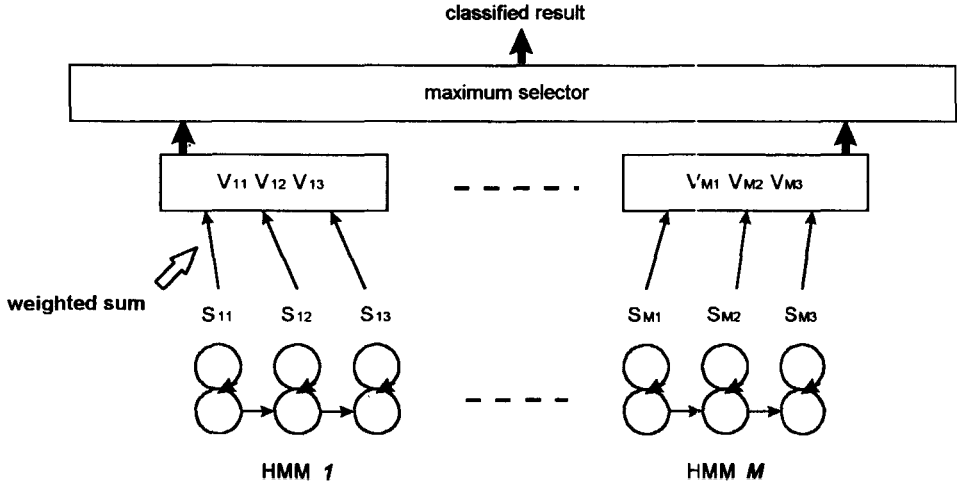
용할 수 있도록 하는 방법도 제안되었다[9][10]. 그러나, 엄밀히 말하면 Su와 Lee의 WHMM과 그후에 발표된 방법들은 약간 다르다. Su와 Lee의 WHMM은 주어진 학습 데이터에 대해서 각 HMM에서 계산되는 state-likelihood 값들을 연결하여 하나의 새로운 특징벡터로 변환하고, 이 특징벡터를 선형분류기로 분류해내는 것이다. 그에 비해서 후에 발표된 연구들은 각 HMM의 각 state에 개별적인 가중치를 도입하고 그러한 가중치들을 이용한 weighted likelihood를 이용해서 인식을 수행하는 것으로 이러한 방법들은 연속음인식으로의 확장이 용이하다. 그림3에 그 차이점을 도시하였다.

Su와 Lee는 WHMM의 학습을 위해서 Amari가 제안한 적응학습 알고리듬에 강인성(robustness)을 향상시킨 개선된 학습알고리듬을 제안하였다. 본래 Amari가 제안한 알고리듬은 오분류가 발생한 경우에만 현재의 학습데이터에 대해서 오분류된 클래스와 경쟁하도록 학습이 수행되나, Su와 Lee가 제안한 알고리듬은 강인성을 높이기 위해서 margin이라는 개념을 도입하여 오분류가 발생하지 않아도 오분류 가능성이 높은 학습데이터에 대해서는 적응학습을 수행하도록 했다. 또한, 분류에 중요한 성분을 위주로 분류를 수행하여 보다 높은 인식율을 얻을 수 있도록 divergence의 개념을 사용하는 2단계 분류기를 제안하였다.

본 논문에서는 그러한 WHMM의 학습에 현재 가장 널리 사용되고 있는 MCE/GPD 방법을 도입하여 그에 따른 학습알고리듬을 유도하고, 100명의 화자가 2회 발성



(a) WHMM 방법



(b) state-likelihoods weighting 방법

그림 3. WHMM과 state-likelihoods weighting 방법
Fig 3. WHMM and state-likelihoods weighting approaches

한 E-set에 대해서 인식 실험을 수행한다. MCE/GPD방법은 Amari방법의 일반화이므로 이러한 알고리듬의 사용은 Su와 Lee방법의 자연스러운 확장이며 일반화라 할 수 있다. Su와 Lee의 방법이 다분히 heuristics한 방법이라면, MCE/GPD방법은 이론적인 해석에 의해서 그들의 방법을 특별한 경우로 포함하게 된다.

본 논문의 구성은 다음과 같다. 먼저 2장과 3장에서는 WHMM과 MCE/GPD 방법에 대해서 간단히 설명하고, 4장에서는 MCE/GPD방법을 사용해서 WHMM을 학습시키는 알고리듬을 유도한다. 5장에서는 E-set에 대한 실험결과와 분석을 제시하고, 6장에서 결론을 기술한다.

Ⅱ. WHMM(weighted hidden Markov models)

기존의 HMM은 어떤 입력에 대해서 Viterbi 복호화에 의해서 얻어지는 확률(likelihood)을 기초로 인식을 수행한다. 이때 계산되는 확률은 그림4와 같이 Viterbi 복호화에 의한 최적분할경로(optimal segmentation path)에 따라 결정되는 각 상태(state)에서의 상태확률(state-likelihood)들의 곱이라고 할 수 있다. 그러나, E-set({b, c, d, e, g, p, t, v, z})과 같이 끝이 모두 'c'와 같은 발음으로 끝나는 단어들의 경우에는 대부분의 상태들이 유사하게 되고, 따라서 앞부분의 짧은 자음부분에 의해서만 변별될 수 있는 문제에 대해서는 기존의 방법으로는 충분히 인식해낼 수 없게 된다. 이 경우에는 공통적인 모음부분의 영향을 배제시키는 것이 변별에 더 좋은 결과를 얻을 것으로 기대된다. 따라서, 그러한 공통된 부분에 대해서는 상태공유(state-sharing)와 같은 방법을 사용하는 것이 바람직하나 어떤 상태들을 어떻게 공유시켜야 하는 지를 알기 어려운 경우가 많다. 따라서, 각 상태확률들을 연결하여 새로운 특징벡터로 사용하여 변별력을 높이려는 알고리듬으로 제안된 것이 WHMM이다[7].

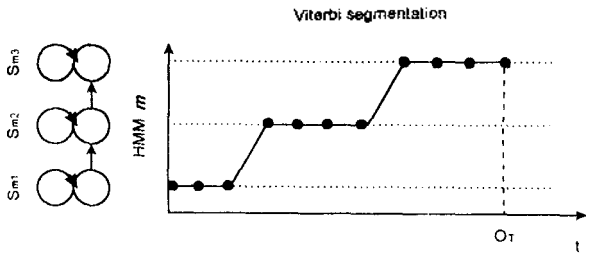


그림 4. Viterbi 분할과정

WHMM의 경우에 모든 HMM에서의 상태확률들을 연결하여 새로운 특징벡터를 만들어 낸다. 따라서, E-set에 대해서 5-state HMM를 사용한 경우에는 새로운 특징벡터의 차원은 45차가 된다. 이 방법은 그림3(a)와 같이 HMM을 새로운 특징벡터를 얻기 위한 변환(transform)으로 사용하고, 그로부터 얻어지는 결과를 선형분류기에 입력시켜서 분류를 수행하는 것이라 할 수 있다. Su와 Lee는 학습알고리듬으로 Amari의 적응학습 알고리듬에 강인성을 첨가한 변형된 알고리듬을 사용하였고, 인식에 중요한 성분을 사용하기 위해서 subspace projection을 사용하는 2단계 인식알고리듬을 제안하였다. 그러나, 이 방법은 어떤 학습데이터에 대해서 가장 혼동되기 쉬운 모델만을 학습에 이용하므로, E-set과 같이 여러개의 혼동되는 클래스가 존재하는 문제에서는 한계가 있다. 따라서 본 논문에서는 그에 대한 일반적인 확장방법인 MCE/GPD방법을 WHMM의 학습에 적용한다.

Ⅲ. MCE/GPD 알고리듬

MCE(minimum classification error)/GPD(generalizaed probabilistic descent) 알고리듬은 최근에 DTW, HMM, 신경회로망 등 많은 인식모델의 학습에 도입되고 있는 변별학습 알고리듬이다. Bayes 오차율을 근사하는 최적 분류기의 학습을 위해서 다음과 같은 세 단계의 MCE 수식화를 통해서 손실함수를 정의하고, 그 함수에 확률적 강하법을 적용하여 최종적인 알습알고리듬을 유도한다.

단계 1. 변별함수(discriminant function)의 정의

클래스 i의 음이 아닌 변별함수 $g_i(x;\lambda)$를 정의한다. 이 변별함수는 인식규칙과 밀접한 관계가 있다. HMM의 경우에는 일반적으로 log-likelihood가 출력되므로, 그 값에 -1을 곱해서 양의 값으로 만들고 그 값들 중 가장 작은 값을 택하여 인식결과로 출력하게 된다.

$$C(x)=C_i \text{ for } g_i(x;\lambda)=\max_j g_j(x;\lambda) \tag{1}$$

$C(\cdot)$는 분류연산, x는 어떤 입력, λ는 파라메터 집합이다.

단계 2. 오분류척도(misclassification meausre)의 정의

오분류척도 $d_i(x;\lambda)$는 분류시의 혼동 가능성을 출력간의 거리함수로 표현한 것이다. 이 척도의 도입이 기존의 방법들과의 가장 큰 차이이고, 이 과정에 의해서 다른 모든 클래스의 파라메터들이 하나의 기준에 통합되게 되어 변별학습이 가능해진다. 여러가지 형태의 오분류척도들을 정의하는 것이 가능하지만 일반적으로 다음의 형태가 많이 사용된다.

$$d_i(x;\lambda)=g_i(x;\lambda)-\left\{\frac{1}{M-1}\sum_{j,j\neq i}g_j(x;\lambda)^{-\eta}\right\}^{\frac{1}{\eta}} \tag{2}$$

위의 수식에서 우측의 두번째 항은 min( )연산을 평탄화시킨 일반적인 연산으로 볼 수 있고, η 값에 따라서 참여하는 혼동되는 클래스의 범위가 조절됨을 알 수가 있다. 특히, $\eta\rightarrow\infty$이면 다음과 같이 되어 2-클래스 Bayes 분류기와 같이 된다.

$$d_i(x;\lambda) = g_i(x,\lambda) - g_{most\ confusible\ with\ class\ i}(x;\lambda) \quad (3)$$

단계 3. 손실함수(loss function)의 정의

손실함수는 입력데이터를 분류하는데 감수하는 위험(risk)과 관련된 함수이다. 그 역할은 (2)의 값을 일정한 범위로 제안하여 오차확률을 근사하도록 하는 것이다. 일반적으로 0-1 손실함수의 확장인 시그모이드함수(sigmoid function)가 주로 사용된다.

$$l_i(d_i(x;\lambda)) = \frac{1}{1+e^{-d_i(x;\lambda)}} \quad (4)$$

식(4)의 손실함수는 하나의 학습데이터에 대해서 정의된다. 식(4)에 대한 기대치가 위험으로 정의되고, 위험을 최소화하는 분류기가 Bayes분류기이다. 실제로 그 기대치를 얻을 수 없으므로, 모든 학습데이터에 대한 손실함수의 합 $L(\lambda)$을 확률적 강하법으로 최소화 시켜야 한다. 학습을 μ_t가 다음과 같이 일정한 조건을 만족시키면 확률적 강하법으로 어떤 기준을 확률적으로 최소화시킬 수 있다는 것이 증명되어 있다.

$$L(\lambda) = E\{l_i(x;\lambda)\} \quad (5)$$

$$\lambda_{t+1} = \lambda_t + \delta\lambda_t, \text{ where } \delta\lambda_t = -\mu_t U \nabla l_i(x;\lambda) \quad (6)$$

$$\sum_{t=1}^{\infty} \mu_t \to \infty, \sum_{t=1}^{\infty} \mu_t^2 \to \infty \quad (7)$$

Ⅳ. WHMM의 MCE/GPD 학습 알고리듬

본 논문에서는 2장의 WHMM을 3장의 MCE/GPD방법으로 학습시키는 알고리듬을 유도한다. 알고리듬의 유도는 MCE수식화의 첫 단계인 변별함수의 정의로 시작되고, 나머지 두 단계의 정의는 식(2)와 식(4)를 그대로 사용한다. m번째 HMM의 상태 p에서의 상태 로그확률을 S_p^m라 하면 변별함수는 다음과 같이 정의된다.

$$g_i(x;\lambda) = \sum_{r=1}^{M} \sum_{p=1}^{P} W_p^{i,r} S_p^r \quad (8)$$

위의 정의로부터 손실함수를 정의하고 확률적 경사법을 적용하면 다음과 같은 학습알고리듬을 얻게 된다.

$$(W_p^{i,r})_{t+1} = (W_p^{i,r})_t + \mu_t l_i(1-l_i) S_p^r \quad (9)$$

$$(W_p^{j,r})_{t+1} = (W_p^{j,r})_t - \mu_t l_i(1-l_i) \nu_j S_p^r \quad (10)$$

$$\nu_j = \left\{ \frac{1}{M-1} \left[\frac{g_j^{-\eta}}{\sum_{k,k\neq i} g_k^{-\eta}} \right]^{-\eta-1} \right\}^{\frac{-1}{\eta}} \quad (11)$$

Ⅴ. E-set에 대한 인식 실험 결과

본 논문에서는 150명의 미국인이 2회 발음한 E-set 데이터베이스를 사용하여 multi-speaker 모드의 실험을 수행하였다. 150명의 1회 발음데이터로 학습시키고, 나머지 데이터는 시험에 사용하였다. 음성은 3.3 kHz로 저역필터링하고 8 kHz로 표본화한 후 전형적인 LPC-front 과정을 거쳐서 12차의 LPC 켑스트럼계수와 12차의 델타 켑스트럼계수로 구성된 24차의 벡터를 한 프레임의 특징벡터로 사용하였다. 학습율은 0.01에서 선형적으로 감소시켰고, 총 학습회수는 300회이다. 실험에 사용된 HMM은 128개의 코드북 크기를 갖는 5-state left-to-right 모델이다. 실험결과는 표1에 요약하였다.

표 1. 인식실험 결과

Table 1. Speech recognition results.

데이터	HMM	$\eta=\infty$	$\eta=5$	$\eta=30$
학습데이터	76.4%	91.9%	92.4%	92.2%
시험데이터	53.8%	65.3%	66.1%	66.1%

실험결과 제안하는 방법에 의해서 시험데이터의 인식율이 약 12% 내외가 향상됨을 알 수 있고, 학습데이터에 대해서도 15% 이상이 개선됨을 알 수 있다.

Ⅵ. 결 론

본 논문에서는 WHMM의 학습을 위한 MCE/GPD 알고리듬을 유도하고, E-set에 대한 인식실험을 통해서 약 14-15% 정도의 인식율 개선을 보였다. 유도된 알고리듬은 여러 개의 혼동되는 클래스가 존재하는 경우에도 효과적으로 적용이 가능한 일반화된 알고리듬이다. 그러나, 학습에 요구되는 계산량은 기존의 방법에 비해서 클래스의 수에 비례해서 늘어난다는 문제점이 있다. 그러므로, 이에 대해서 적절한 기준 등을 도입해서 불필요한 계산을 줄일 수 있는 방법에 대한 연구가 필요하다.

또한 선형분류가 대신에 다른 여러가지 신경회로망과의 혼용도 가능할 것이다. 앞으로 그에 대한 연구를 계속해 나갈 것이다.

참 고 문 헌

1. L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257-286, Feb. 1989.
2. L. R. Bahl, P. F. Brown, P. V. de Souza and R. L. Mercer, "A new algorithm for the estimation of hidden Markov model parameters," in *Proc. ICASSP'88*, pp. 493-496, 1989.
3. L. R. Bahl, P. F. Brown, P. V. de Souza and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. ICASSP'86*, pp. 49-52, 1986.
4. Y. Ephraim, A. Dembo and L. R. Rabiner, "A minimum discrimination information approach for hidden Markov modeling," in *Proc. ICASSP'88*, pp. 25-28, 1988.

5. B. H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," IEEE Trans. Signal Processing, vol. 40, no. 12, pp. 3043-3054, 1992.
6. W. Chou, B. H. Juang and C. H. Lee, "Segmental GPD training of HMM based speech recognizer," in *Proc. ICASSP'92*, pp. 473-476, 1992.
7. K. Y. Su and C. H. Lee, "Speech recognition using weighted HMM and subspace projection approaches," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 1, part I, pp. 69-79, 1994.
8. F. Wolfertetter and G. Ruske, "Discriminative state-weighting in hidden Markov models," in *Proc. ICSLP'94*, pp. 219-222, 1994.
9. Y. J. Chung and C. K. Un, "Multilayer perceptrons for stae-dependent weightings of HMM likelihoods," *Speech Communication*, vol. 18, pp. 79-89, 1996.
10. O. W. Kwon and C. K. Un, "Discriminative weighting of HMM state-likelihoods using the GPD method," *IEEE Signal Processing Letters*, vol. 3, no. 9, pp. 257-259, 1996.

▲최 홍 섭(Choi Hong-Sub) 1957년 10월 3일생

1985년 2월 : 서울대학교 전자공학과 (공학사)
1987년 2월 : 서울대학교 전자공학과 (공학석사)
1988년 3월~1993년 2월 : 서울대학교 공학연구소 조교
1994년 8월 : 서울대학교 전자공학과 (공학박사)
1995년 3월~현재 : 대진대학교 전자공학과 전임강사