MAP 수식화에 의한 HMM의 변별력있는 학습 알고리듬

전 범 기°, 나 경 민*, 안 수 길*
서울대학교 전자공학과*

A Discriminative Training Algorithm for HMM Based on MAP Formulation

BumKi JEON°, KyungMin NA*, SouGuil ANN*
Dept. of Electronics Eng., Seoul Nat'l Univ.*

요 약

기존의 HMM(Hidden Markov Models)을 이용한 음성인식기는 대부분 ML(Maximum Likelihood) 추정에 기초한 Baum-Welch 알고리듬으로 학습되었다. ML학습은 기본적으로 무한한 양의 학습 데이타가 주어지고, 각 모델들이 서로 독립이라는 가정에 기초한다. 하지만 실제적인 학습의 경우에 각 모델들이 서로 독립이라고 보기 어렵고, 학습 데이타의 양도 상당히 제한되어 있어서 인식기의 변별력을 저하시키는 주된 원인이 되고 있다. 본 논문에서는 전통적인 패턴분류기법인 Bayes 결정이론에 따라 최소오차율분류(Minimum Error Rate Classification)를 위한 MAP(Maximum A Posteriori) 수식화를 유도하고, 그에 기초한 HMM의 변별력있는 학습 알고리듬을 제안한다. 최소오차율분류를 근사화한 사후확률(a posteriori probabilty)로 표현된 비용함수를 정의하고, 그 비용함수에 조건부 경사강하법을 적용한다. 제안된 알고리듬을 분류하기 어려운 한국어 단음절 (가,나,다,라,마,바,사,아,자) 인식에 적용한 결과, 기존의 ML 알고리듬으로 학습한 경우 발생한 오인식 갯수의 약 10 % 가량이 개선되었다.

I. 서 론

HMM을 음성인식에 처음으로 적용하게된 계기는 효율적인 학습 알고리듬인 Baum-Welch 알고리듬이 존재하기 때문이었다 [1]. 하지만 B-W 알고리듬의 근거인 ML 학습은 기본적으로 무한한 양의 학습 데이타가 주어지고, 각 모델들이 서로 독립인 경우 최적의 학습이 보장되는 것으로 알려져 있다. 하지만 실제적인 학습의 경우에 각 모델들이 서로 독립이라고 보기 어렵고, 학습 데이타의 양도 상당히 제한되어 있어서 인식기의 변별력을 저하시키는 주된 원인이 된다. 따라서 최근에 모델들간의 변별력을 높이기 위한 학습 알고리듬으로 실제 음성신호와 모델간의 상호정보(mutual information)를 최대가 되도록 하는 MMI (Maximum Mutual Information) 수식화에 의한 학습 알고리듬 [2], 기존의 ML학습을 모델들이 서로 독립이 아닌 경우로 확장시킨 MDI(Minimum Discrimination Information) 수식화에 의한 학습 알고리듬 [3], 그리고 최소오차분류수식화(MCEF; Minimum Classification Error Formulation)에 GPD(Generalized Probabilistic Descent) 방법을 적용한 Segmental GPD 알고리듬등이 개발되어 왔다 [4][5]. 전자의 두 학습 알고리듬은 비용함수(cost function)가 인식오차(recognition error)로부터 유도되지 않으므로 오차율을 간접적으로 감소시키는 방법인 반면 후자의 Segmental GPD 알고리듬은 비용함수가 인식오차로부터 직접 유도되므로 오차율을 직접적으로 감소시키는 방법이다.

본 논문에서는 전통적인 분류기법인 Bayes 결정이론에 따라 최소오차율분류(Minimum Error Rate Classification)를 위한 Bayes risk를 사후확률(a posteriori probability)로 표현한 후, 이를 근사화한 비용함수를 최소화하는 학습 알고리듬을 제안한다. 가능한 모든 데이타들에 대한 Bayes risk를 주어진 학습 데이타에 대해서 근사화한 비용함수는 사후확률을 최대화시키는 MAP (Maximum A Posteriori) 수식이 된다. 이와 같이 정의된 비용함수에 조건부경사강하법(constrained gradient descent) 방법을 적용하여 MAP 수식화에 기초한 HMM의 변별력있는 학습 알고리듬(MAP 알고리듬)을 유도한다.

본 논문의 구성은 다음과 같다. II 장에서는 ML 학습 알고리듬의 문제점과 기존의 학습 알고리듬에 대해서 간략히 검사한다. III 장에서는 Bayes 결정이론으로부터 Bayes risk를 사후확률을 이용하여 MAP 수식화로 근사화하고 그에 기초한 HMM의 변별력있는 학습 알고리듬을 유도한 후, IV 장에서는 제안된 학습 알고리듬의 타당성을 보이기 위해서 ML 알고리듬과 기존의 변별력있는 알고리듬과의 실험결과를 비교하고 마지막으로 V 장에서 결론을 제시한다.

II. ML 알고리듬과 기존의 변별력 있는 알고리듬

HMM의 학습 알고리듬인 Baum-Welch 재추정 알고리듬은 ML 추정에 기초하며, ML 학습은 아래와 같은 조건하에서 최적의 학습을 보장됨이 Nadas에 의해 증명되어 있다 [6].

1) 최적의 모델로부터 발생된 관찰열이 학습 데이타에 포함되고,

2) 학습 데이타가 상당히 크고,

3) 인식 모델들이 서로 독립이다.

하지만 실제적인 학습의 경우에는 위의 세 가지 조건들이 만족되지 못하므로 ML 학습이 최적의 모델을 구현한다고 보기가 어렵다. 따라서 최근에 모델들간의 변별력을 높이기 위한 MMI (Maximum Mutual Information), MDI (Minimum Discrimination Information), Segmental GPD 알고리듬등이 개발되어 왔다.

MMI알고리듬은 관찰열과 그에 대응하는 모델사이의 상호정보(mutual information)에 대한 음의 평균치를 식 (1)과 같은 비용함수로 정의하여 이를 최소화시키는 파라메터를 구하는 알고리듬이다 [2].

$$L(\Lambda) = -\sum_{m=1}^{M} I(x_n^m; C_m)$$

$$= \sum_{m=1}^{M}\left[\log\sum_{k=1}^{M} P(x_n^m|\lambda_k)P(C_k) - \log P(x_n^m|\lambda_m)\right] \quad (1)$$

MDI 알고리듬은 음성신호가 가우시안 AR(Gaussian autoregressive) 모델이라는 가정에 기초하며, 모델의 확률분포(probability distribution)에 대한 실제 음성의 추정확률분포의 변별정보(discrimination information)를 식 (2)와 같이 최소화시키는 학습 알고리듬이다 [3].

$$v(R, p_\Lambda) = \inf_{Q\in\Omega(R)} D(Q\,||p_\Lambda) \quad (2)$$

단, R는 모델의 covariance 행렬

$\Omega(R)$은 R로 생성 가능한 모든 PDF

$D(\cdot)$는 relative entropy

Segmental GPD 알고리듬은 식 (3)과 같이 분류오차를 적당한 함수의 형태로 변환시킨 최소분류오차수식화 (MCEF: Minimum Classification Error Formulation)에 GPD를 적용시킨 알고리듬이다 [5].

$$L(\Lambda) = \frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M} l_m(x_n;\Lambda)\,1(x_n\in C_m) \quad (3)$$

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

$$d_k(x;\Lambda) = -g_k(x;\Lambda) + \log\left[\frac{1}{M-1}\sum_{j\neq k}\mathrm{EXP}(\eta\, g_j(x;\Lambda))\right]^{\frac{1}{\eta}}$$

$$g_i(x;\Lambda) = \log p(x|\lambda_i)$$

MCEF로부터 유도된 비용함수를 GPD 방법에 의해 최소화시키는 파라메터 Λ에 대한 학습식을 구한다. Segmental GPD 알고리듬은 전자의 두가지 방법과는 달리 분류오차로부터 변별함수를 정의하므로, ML 학습에 의한 인식기의 오차율과 비교해서 Segmental GPD 알고리듬으로 학습된 인식기의 오차율이 감소됨을 쉽게 보일 수 있으며 이는 실험적으로 증명되어 있다.

III. Bayes 결정이론과 MAP 알고리듬

일반적으로 인식기의 평균오차확률(average probability of error)의 최소값을 얻을 수 있는 Bayes 결정이론은 식 (4)와 같다 [7].(아래의 수식들에서 $P(x)$와 $p_\Lambda(x)$는 각각 확률과 확률분포함수를 나타낸다.)

$$C(x) = C_m, \quad \text{if } p_\Lambda(C_m|x) = \max_k p_\Lambda(C_k|x) \quad (4)$$

단, $C(\cdot)$는 인식연산(recognition operation)을 나타낸다.

Bayes 결정이론에 따라 인식하는 최적의 인식기는 인식연산에 따른 손실(loss)을 나타내는 risk를 정의한 후, 이를 최소화시키는 파라메터를 갖는 인식기가 된다.

먼저, 인식해야 할 객층 $C = \{C_1, C_2, \cdots, C_M\}$에 대해서 실제 객층이 C_k인 경우 α_m을 C_m으로 인식하는 어떤 행위라고 정의하고, α_k라는 행위를 함으로써 발생하는 손실(loss)을 $\lambda(\alpha_m|C_k)$라 하자. 어떤 x를 관찰하고 α_k라는 행위를 했는데 실제로는 x가 C_k에 포함된다고 가정하면, $\lambda(\alpha_m|C_k)$의 손실이 일어났다고할 수 있다. 이로부터 conditional risk라고 알려진 손실의 기대치 $R(\alpha_m|x)$와 최종적인 overall risk는 식 (5)와 같이 정의된다.

$$R(\alpha_m|x) = \sum_{k=1}^{M}\lambda(\alpha_m|C_k)p_\Lambda(C_k|x) \quad (5\text{-}a)$$

$$R = \int\left[\sum_{m=1}^{M} R(\alpha_m|x)\right]P(x)\,dx$$

$$= \int\left[\sum_{m=1}^{M}\left\{\sum_{k=1}^{M}\lambda(\alpha_m|C_k)p_\Lambda(C_k|x)\right\}\right]P(x)\,dx \quad (5\text{-}b)$$

위의 risk를 최소화시키는 인식기가 최적의 인식기가 되며, 그 최소 risk를 Bayes risk라고 한다. 최소오차율 분류를 위해서 risk에 사용된 손실함수로써 대칭손실함수 또는

zero-one 손실함수라고 불리우는 식 (6)과 같은 함수를 사용한다.

$$\lambda(\alpha_m|C_k) = \begin{cases} 0, & m=k \\ 1, & m\neq k \end{cases} \quad 1 \le m,k \le M \qquad (6)$$

윗 식의 손실함수는 올바른 인식에 대해서는 손실을 주지 않고 틀린 인식에 대해서만 1의 손실을 준다. 이런 손실함수에 대한 risk는 식 (7)과 같은 평균오차확률이 된다.

$$R(\alpha_m|x) = \sum_{k=1}^{M} \lambda(\alpha_m|C_k)p_\Lambda(C_k|x)$$
$$= 1 - p_\Lambda(C_m|x) \qquad (7)$$

또한, Bayes 규칙으로부터 사후확률(a posteriori probability)을 사전확률(a prior probability)과 계층조건부확률(class-conditional probability)로 표현할 수 있다. 이를 이용하여 overall risk를 표현하면 식 (8)과 같다 [8].

$$R = \int \left[\sum_{m=1}^{M} R(\alpha_m|x)\right]P(x)\,dx$$
$$= \int \left[\sum_{m=1}^{M} \frac{\sum_{k=1,k\neq m}^{M} p_\Lambda(x|C_k)P(C_k)}{\sum_{k=1}^{M} p_\Lambda(x|C_k)P(C_k)}\right] P(x)\,dx \qquad (8)$$

식 (8)에서 사후확률의 증가는 overall risk의 감소가 되므로 식 (8)는 MAP 수식화된 overall risk된다.

먼저, 비용함수를 식 (8)의 overall risk를 주어진 이산 학습 데이타의 집합 $X = \{x_1, x_2, \cdots, x_N\}$에 대해서 근사화하면 식 (9)와 같이 된다.

$$L(\Lambda) = \sum_{n=1}^{N} \sum_{\substack{m=1 \\ x_n\in C_m}}^{M} \left\{ \frac{\sum_{k=1,k\neq m}^{M} p_\Lambda(x_n^m|C_k)P(C_k)}{\sum_{k=1}^{M} p_\Lambda(x_n^m|C_k)P(C_k)} \right\} P(x_n^m) \qquad (9)$$

식 (9)의 비용함수를 최소화시키기 위한 파라메터 학습식을 유도하여 이런 학습식으로 표현되는 MAP 수식화에 기초한 학습 알고리듬(MAP 알고리듬)을 제안한다.

$$\nabla_\Lambda L(\Lambda_t) = \nabla_\Lambda \left[\sum_{n=1}^{N} \sum_{\substack{m=1 \\ x_n\in C_m}}^{M} \left\{ \frac{\sum_{k=1,k\neq m}^{M} p_\Lambda(x_n^m|C_k)P(C_k)}{\sum_{k=1}^{M} p_\Lambda(x_n^m|C_k)P(C_k)} \right\} P(x_n^m) \right]$$
$$= \sum_{n=1}^{N} \sum_{\substack{m=1 \\ x_n\in C_m}}^{M} \nabla_\Lambda \left[\left\{ \frac{\sum_{k=1,k\neq m}^{M} p_\Lambda(x_n^m|C_k)P(C_k)}{\sum_{k=1}^{M} p_\Lambda(x_n^m|C_k)P(C_k)} \right\} P(x_n^m) \right] \qquad (10)$$

식 (10)에서 $P(x_n^m)$는 파라메터 Λ와 무관하므로 편미분 밖으로 나올 수 있고, 모든 학습 데이타의 발생빈도를 동일하다고 볼 수 있으며, 모델의 발생빈도도 동일하다고 볼 수 있으므로 식 (11)과 같이 변환된다.

$$\nabla_\Lambda L(\Lambda_t) = \frac{1}{N} \sum_{n=1}^{N} \sum_{\substack{m=1 \\ x_n\in C_m}}^{M} \left[\nabla_\Lambda \left(\frac{\sum_{k=1,k\neq m}^{M} p_\Lambda(x_n^m|C_k)\frac{1}{M}}{\sum_{k=1}^{M} p_\Lambda(x_n^m|C_k)\frac{1}{M}} \right) \right]$$
$$= \frac{1}{N} \sum_{n=1}^{N} \sum_{\substack{m=1 \\ x_n\in C_m}}^{M} \left[\nabla_\Lambda \left(\frac{\sum_{k=1,k\neq m}^{M} p_\Lambda(x_n^m|C_k)}{\sum_{k=1}^{M} p_\Lambda(x_n^m|C_k)} \right) \right] \qquad (11)$$

식 (11)에 조건부경사강하법인 reduced gradient descent 방법을 적용하면 식 (12)와 같이 $\nabla_\Lambda L(\Lambda_t)$가 유도된다 [9][10].

$$\frac{\partial L(\Lambda)}{\partial \lambda_l} =$$
$$\frac{1}{N} \sum_{n=1}^{N} \left[\sum_{\substack{m=1 \\ x_n\in C_m \\ m=l}}^{M} \left\{ \frac{p_\Lambda(x_n^m|C_m)}{\left(\sum_{k=1}^{M} p_\Lambda(x_n^m|C_k)\right)^2} \right\} \frac{\partial p_\Lambda(x_n^m|C_l)}{\partial \lambda_l} \right.$$
$$\left. - \sum_{\substack{m=1 \\ x_n\in C_m \\ m=l}}^{M} \left\{ \frac{\sum_{k=1,k\neq m}^{M} p_\Lambda(x_n^m|C_k)}{\left(\sum_{k=1}^{M} p_\Lambda(x_n^m|C_k)\right)^2} \right\} \frac{\partial p_\Lambda(x_n^m|C_l)}{\partial \lambda_l} \right] \qquad (12)$$

단, $1 \le l \le N$

IV. 실험결과

본 논문에서 제안한 MAP 학습 알고리듬과 기존의 학습 알고리듬을 비교하기 위해서 한국어 단음절(가,나,다,라,마,바,사,아,자)의 고립단어 인식실험을 수행하였다. 20명의 화자가 1회 발성한 180개의 음성 데이타를 사용하여 제안된 MAP 알고리듬에 의해 학습한 인식기에 대해서 학습에 사용된 음성 데이타에 대한 인식실험(실험 1)과 남은 1회 발성한 180개의 음성 데이타에 대한 인식실험(실험 2)을 수행하였다. 기존의 ML 알고리듬(B-W 알고리듬), MMI 알고리듬, 그리고 Segmental GPD 알고리듬에 의해 학습한 인식기에 대해서도 동일한 실험을 수행하여 실험 1과 실험 2의 결과와 비교하였다.

그림 1은 제안된 방법에 의한 학습시 비용함수의 감소를 나타내고 있다.

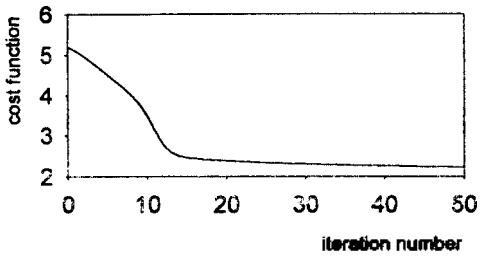

그림 1. 학습시 비용함수의 변화

표 1과 표 2에서 각각 실험 1과 실험 2에 대한 제안된 방법과 기존의 방법에 의한 오인식 갯수와 인식률을 비교하였다.

실 험 1 (180 개의 데이타)				
학습방법	ML	MMI	GPD	MAP
오인식갯수	6	4	2	2
인식률	96.7%	97.8%	98.9%	98.9%

표. 1. 실험 1에 대한 오인식 갯수와 인식률 비교

실 험 2 (180 개의 데이타)				
학습방법	ML	MMI	GPD	MAP
오인식갯수	87	83	81	81
인식률	51.7%	53.9%	55.0%	55.0%

표. 2. 실험 2에 대한 오인식 갯수와 인식률 비교

표 1과 2에서 보듯이 MAP 알고리듬에 의한 인식실험의 결과 기존의 ML 알고리듬에 의해 발생한 오차 93 개중 10 개를 줄여 발생오차의 10.8%가 감소되었다. 또한 기존의 변별력있는 알고리듬과의 비교 결과, MMI 알고리듬보다 우수한 성능을 보였으며 기존의 알고리듬중 가장 성능이 우수한 것으로 알려진 GPD 알고리듬과 비슷한 성능을 보임을 알 수가 있었다.

V. 결론

본 논문에서는 Bayes risk를 MAP 수식화하고 이를 비용함수로 하는 HMM을 위한 새로운 변별력있는 학습 알고리듬을 제안하였다. 제안된 알고리듬은 인식오차로부터 비용함수를 유도하지 않았으므로 간접 방법에 해당되지만, MMI 알고리듬과 MDI 알고리듬과는 달리 수식적으로 인식오차가 감소됨을 알 수 있다. 또한 GPD 알고리듬과 같이 학습 데이타가 속한 모델 파라메터의 경우는 강화학습을 하고 소하지 않은 모델 파라메터에 대해서는 반강화학습을 하므로 GPD 알고리듬과 비슷한 성능을 가지게 된다.

<참고문헌>

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. of IEEE*, pp. 257-286, 1989.

[2] 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," *Proc. ICASSP-86*, pp. 49-52, 1986.

[3] Y. Ephraim, A. Dembo, and L. R. Rabiner, "A minimum discrimination information approach for hidden Markov modeling," *IEEE Trans. Information Theory*, vol. 35, no. 5, pp. 1001-1013, Sep. 1989.

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

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

[6] A. Nadas, "Optimal solution of a training problem in speech recognition," *IEEE Trans ASSP*, vol. 33, no. 1, pp. 326-329, Feb. 1985.

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

[8] N. merhav and Y. Ephraim, "A Bayesian classification approach with application to speech recognition," *IEEE Trans. Signal Processing*, vol. 39, no. 10, pp. 2157-2166, Oct. 1991.

[9] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *AT&T Bell System Technical Journal*, vol. 62, no. 4, pp. 1035-1074, Apr. 1983.

[10] P. E. Gill and W. Murray, *Numerical Methods for Constrained Optimization*. New York: Academic Press 1974.