A Regularized Discriminative Training Method of Acoustic Models Derived by Minimum Relative Entropy Discrimination

Yotaro Kubo[1,3]*, Shinji Watanabe*[2]*, Atsushi Nakamura*[2]*, Tetsunori Kobayashi*[1]

[1]Department of Computer Science and Engineering, Waseda University, Tokyo, Japan
[2] NTT Communication Science Laboratory, Kyoto, Japan
[3] Lehrstuhl für Informatik 6, RWTH Aachen University, Aachen, Germany
yotaro@ieee.org, {watanabe,ats}@kecl.cslab.ntt.co.jp, koba@waseda.jp

Abstract

We present a realization method of the principle of minimum relative entropy discrimination (MRED) in order to derive a regularized discriminative training method. MRED is advantageous since it provides a Bayesian interpretations of the conventional discriminative training methods and regularization techniques. In order to realize MRED for speech recognition, we proposed an approximation method of MRED that strictly preserves the constraints used in MRED. Further, in order to practically perform MRED, an optimization method based on convex optimization and its solver based on the cutting plane algorithm are also proposed. The proposed methods were evaluated on continuous phoneme recognition tasks. We confirmed that the MRED-based training system outperformed conventional discriminative training methods in the experiments.

Index Terms: speech recognition, discriminative training, optimization

1. Introduction

Recently, several discriminative training approaches have been studied in order to directly minimize the error rate of automatic speech recognition systems. In these methods, the parameters of probabilistic models are estimated by optimizing a discriminant criterion function. Several methods for discriminative training have already been proposed along with choices of the criteria [1–3]. Although discriminative training methods significantly outperform conventional maximum likelihood training methods, the training process still includes the risk of overfitting when the amount of available training data is insufficient.

In order to avoid overfitting problems, regularization techniques, which introduce additional terms to penalize overfitting, have been investigated. The support vector machine (SVM) is one of the most successful discriminative models that utilizes a regularization technique [4]. In SVM, large-margin linear classifiers are achieved by introducing regularization terms that minimize the L2-norm of the weight vectors used in linear classifiers. Regularization techniques have also been applied to discriminative training methods for probabilistic models. The I-smoothing technique can be considered as a regularization technique that controls the estimated model parameters such that a higher likelihood as well as discriminative performance is ensured [3]. Further, as in the case of SVMs, large-margin techniques are also widely investigated in several articles [5, 6]. In the both cases, it is confirmed that the regularization techniques are efficient in order to prevent overfitting in discriminative training.

In the context of Bayesian inference, regularization is utilized by introducing a prior probability density function (pdf) of the model parameters. As compared with the Bayesian approach, conventional regularization techniques used in discriminative training lack extensibility. For example, in the Bayesian approach, various problems connected to automatic speech recognition, such as model selection problems, context clustering problems and domain adaptation problems, are formulated with unified probabilistic interpretation.

Recently, the principle of minimum relative entropy discrimination (MRED) was proposed by Jaakkola *et al.* [7] in order to provide Bayesian interpretations of discriminative model training methods. In this paper, as a core part of a full MRED speech recognition framework, we derive an improved realization method of the MRED principle for continuous-density hidden Markov models (CD-HMMs). Conventionally, this principle is applied to feature selection [8], training of Gaussian mixture models (GMMs) [9], and recognition of isolated phoneme based on kernel methods [10]. However, previous MRED methods suffer from problems when a variety of latent variables are used in the models; therefore, their application to training of CD-HMMs used for continuous speech recognition is difficult. By applying the proposed MRED method, we derived an MCE-based discriminative training method that can explicitly embed the concept of classification margin and have extensibility due to probabilistic representations of model parameters.

2. Minimum relative entropy discrimination (MRED)

In MRED, by considering all variables as random variables, the posterior pdf $P(\Theta, \xi)$ of model parameters Θ and classification margin variables ξ is estimated by minimizing the Kullback-Leibler divergence (KL divergence) between the posterior pdf and the prior pdf $P^0(\Theta, \xi)$ so that the expectation of the classification performance function $\mathcal{D}$ is greater than that of the classification margin variables ξ^i, formulated as follows:

$$\begin{aligned} &\min_{P(\Theta,\xi)} \ \mathrm{KL}[P(\Theta,\xi)||P^0(\Theta,\xi)], \\ &\text{s. t. } \left\langle \mathcal{D}(\boldsymbol{X}^i;\Theta) - \xi^i \right\rangle_{P(\Theta,\xi)} \geq 0, \quad \forall i. \end{aligned} \tag{1}$$

Here, $\boldsymbol{X}^i$ is i^{th} feature sequence in a training dataset; $\langle f(x) \rangle_{P(x)}$ denotes the expectation of $f(x)$ with respect to $P(x)$, i.e. $\langle f(x) \rangle_{P(x)} \stackrel{\text{def}}{=} \int_x P(x) f(x) dx$; and $\mathrm{KL}[f(x)||g(x)]$ denotes the KL divergence of $g(x)$ from $f(x)$, i.e. $\mathrm{KL}[f(x)||g(x)] \stackrel{\text{def}}{=} \langle \log f(x) - \log g(x) \rangle_{f(x)}$. In this formulation, inequality constraints are designed for each input sequence of the classifier. Therefore, margin variables are incorporated for each input sequence, i.e. $\xi = \{\xi^i | \forall i\}$. Because the constraint due to i^{th} feature sequence is always satisfied by reducing ξ^i, ξ^i is generally controlled by introducing a prior pdf $P^0(\xi^i)$ that favors a larger value of the margin variable ξ^i. This formulation enables us to utilize the prior knowledge of parameters and classification margin via the prior pdf $P^0(\Theta, \xi)$.

As an example, we introduce a performance function that indicates the discriminative performance of a given parameter set Θ and feature sequence $\boldsymbol{X}^i$, as follows:

$$\mathcal{D}(\boldsymbol{X}^i;\Theta) = \underbrace{\log P(\boldsymbol{X}^i, \boldsymbol{l}^i|\Theta)}_{R^i(\Theta)} - \underbrace{\log \max_{\boldsymbol{l} \neq \boldsymbol{l}^i} P(\boldsymbol{X}^i, \boldsymbol{l}|\Theta)}_{C^i(\Theta)}, \tag{2}$$

where $\boldsymbol{l}^i$ is the label sequence corresponding to the i^{th} feature sequence in the training dataset, $\boldsymbol{l}$ is a variable that denotes a label sequence, $R^i(\Theta)$ denotes a reference log-likelihood, and $C^i(\Theta)$ denotes a competitor's log-likelihood. In this paper, we used this MCE-type performance function [2]; however, an alternative choice is also applicable.

3. Application of MRED to CD-HMMs

In this section, we describe a method for performing the abovementioned optimization problem. Although MRED provides the formulation based on convex optimization, this optimization problem is not tractable due to the latent variable used in CD-HMMs. In order to make the optimization solvable, we first reformulate the original optimization problem to an auxiliary optimization problem with tighter constraints. Then, the remaining intractable parts in the constraint functions are eliminated by representing a single constraint by infinite constraints. Finally, a practical solver for the auxiliary problem is presented.

3.1. Auxiliary problem with tighter constraints

Because the inequality constraints defined in the original optimization (Eq. (1)) are non-negative constraints, the use of a lower bound function of the performance function $\mathcal{D}$ yields tighter constraints. The tighter constraints are important to ensure that the solution of the auxiliary optimization also satisfies the constraints of the original optimization.

In order to obtain a lower bound of $\mathcal{D}$, we introduce a lower bound of $R^i(\Theta)$ and an upper bound of $C^i(\Theta)$ in Eq. (2). A lower bound of $R^i(\Theta)$ can be obtained by introducing a pdf $\psi^i(\boldsymbol{q})$ of the latent variable $\boldsymbol{q}$ used in CD-HMMs (i.e. $\boldsymbol{q}$ represents sequences of mixture components and HMM states) and applying Jensen's inequality, as follows:

$$\begin{aligned} R^i(\Theta) &= \log \sum_{\boldsymbol{q}} P(\boldsymbol{X}, \boldsymbol{l}, \boldsymbol{q}|\Theta) \\ &\geq \sum_{\boldsymbol{q}} \psi^i(\boldsymbol{q}) \log P(\boldsymbol{X}, \boldsymbol{l}, \boldsymbol{q}|\Theta) + H[\psi^i] \stackrel{\text{def}}{=} \hat{R}^i[\Theta, \psi^i], \end{aligned} \tag{3}$$

where $H[\psi]$ is the entropy function, i.e. $H[\psi] \stackrel{\text{def}}{=} -\sum_{\boldsymbol{q}} \psi(\boldsymbol{q}) \log \psi(\boldsymbol{q})$. Further, an upper bound of C^i is obtained as follows:

$$\begin{aligned} C^i(\Theta) &= \log \max_{\boldsymbol{l} \neq \boldsymbol{l}^i} P(\boldsymbol{X}^i, \boldsymbol{l}|\Theta) \leq \log \sum_{\boldsymbol{l} \neq \boldsymbol{l}^i} P(\boldsymbol{X}^i, \boldsymbol{l}|\Theta) \\ &= \max_{\tilde{\psi}^i} \underbrace{\sum_{\boldsymbol{l} \neq \boldsymbol{l}^i} \sum_{\boldsymbol{q}} \tilde{\psi}^i(\boldsymbol{q}) \log P(\boldsymbol{X}, \boldsymbol{l}, \boldsymbol{q}|\Theta) + H[\tilde{\psi}^i]}_{\check{C}^i[\Theta, \tilde{\psi}^i]}. \end{aligned} \tag{4}$$

By substituting these functions (Eqs. (3) and (4)) into the original optimization defined in Eq. (1), the original optimization is reformulated to an optimization problem that has tighter constraints, defined as follows:

$$\begin{aligned} &\min_{P(\Theta,\xi)} \ \mathrm{KL}[P(\Theta,\xi)||P^0(\Theta,\xi)], \\ &\text{s.t.} \left\langle \hat{R}^i[\Theta,\psi^i] - \max_{\tilde{\psi}^i} \check{C}^i[\Theta,\tilde{\psi}^i] - \xi^i \right\rangle_{P(\Theta,\xi)} \geq 0, \quad \forall i. \end{aligned} \tag{5}$$

Here, maximization due to $\tilde{\psi}^i$ cannot be omitted since this maximization is required to maintain the strictness of the constraints.

3.2. Reformulating to a semi-infinite convex programming

In order to derive a method for performing the auxiliary optimization defined in Eq. (5), this problem is reformulated to a semi-infinite programming in this section. We focused that the maximization due to $\tilde{\psi}^i$ is equivalent to searching for the tightest constraint obtained by varying $\tilde{\psi}^i$. By considering a search method that is performed iteratively in a solver (cf. Section 3.4), we introduce an infinite sequence of search hypotheses, which includes a solution that corresponds to the tightest constraints, as $\{\tilde{\psi}^i_1, \tilde{\psi}^i_2, \cdots, \tilde{\psi}^i_k, \cdots\}$, where k is an index of the elements in this sequence which runs from 1 to infinity. Each hypothesis $\tilde{\psi}^i_k$ represents a posterior pdf of the latent variable $\boldsymbol{q}$. Because satisfying all constraints in this sequence leads to satisfaction of the tightest constraint as in Eq. (5), an equivalent optimization can be expressed as follows:

$$\begin{aligned} &\min_{P(\Theta,\xi)} \ \mathrm{KL}[P(\Theta,\xi)||P^0(\Theta,\xi)], \\ &\text{s.t.} \left\langle \hat{R}^i[\Theta,\psi^i] - \check{C}^i[\Theta,\tilde{\psi}^i_k] - \xi^i_k \right\rangle_{P(\Theta,\xi)} \geq 0, \quad \forall i, \forall k. \end{aligned} \tag{6}$$

We note that this formulation avoids the intractable maximization ($\max_{\tilde{\psi}^i}$) by introducing the infinite sequence of hypotheses. By applying this technique, the intractable optimization is reformulated as a tractable semi-infinite convex optimization.

As in the case of other convex optimization problems, a function form of the solution of the auxiliary problem (Eq. (6)) can be restricted by considering the Karush-Kühn-Tucker conditions, as follows:

$$\begin{aligned} P(\Theta,\xi) &= \frac{P^0(\Theta,\xi)}{Z(\alpha)} \exp\left\{ \sum_{i,k} \alpha^i_k \left(\hat{R}^i[\Theta,\psi^i] - \check{C}^i[\Theta,\tilde{\psi}^i_k] - \xi^i_k \right) \right\}, \\ Z(\alpha) &= \left\langle \exp\left\{ \sum_{i,k} \alpha^i_k \left(\hat{R}^i[\Theta,\psi^i] - \check{C}^i[\Theta,\tilde{\psi}^i_k] - \xi^i_k \right) \right\} \right\rangle_{P^0(\Theta,\xi).} \end{aligned} \tag{7}$$

Here, nonnegative variables $\alpha^i_k \geq 0$, called Lagrange multipliers, are introduced for each constraint in the auxiliary optimization problem. Because the number of constraints is increased to infinity, the numbers of Lagrange multipliers $\alpha \stackrel{\text{def}}{=} \{\alpha^i_k | \forall i, \forall k\}$ and the margin variables $\xi \stackrel{\text{def}}{=} \{\xi^i_k | \forall i, \forall k\}$ are also increased to infinity. Further, the problem defined in Eq. (6) has an alternative form called the dual problem, and it is expressed as follows:

$$\max_{\alpha} J(\alpha) = -\log Z(\alpha), \quad \text{s.t.}\ \alpha^i_k \geq 0, \quad \forall i, \forall k. \tag{8}$$

The optimal posterior pdf can be obtained by substituting the solution of this dual problem into Eq. (7).

Hereinafter, we introduce CD-HMM parameters as Θ and assume the independency of the parameters, i.e. $\Theta \stackrel{\text{def}}{=} \{\mu_{g,d}, \tau_{g,d}, \rho_{s,m}, | \forall g, \forall d, \forall s, \forall m\}$ and $P^0(\Theta,\xi) \stackrel{\text{def}}{=} \prod_{g,d} P^0(\mu_{g,d}, \tau_{g,d}) \prod_s P^0(\boldsymbol{\rho}_s) \prod_{i,k} P^0(\xi^i_k)$ [1], where g, d, s, and m denote the indices of all Gaussian pdfs used in CD-HMMs, dimensions, HMM states, and mixture components, respectively; $\mu_{g,d}$ and $\tau_{g,d}$ denote the mean and precision (inverse of variance) of $(g,d)^{\text{th}}$ Gaussian, respectively. The optimization function $J(\alpha)$ in the dual problem (Eq. (8)) can be decomposed into Gaussian pdf terms $J^{\texttt{GAUSS}}_{g,d}$, mixture terms $J^{\texttt{MIX}}_s$, and margin terms $J^{\texttt{MARGIN}}_{i,k}$, as follows:

$$J(\alpha) = \sum_{g,d} J^{\texttt{GAUSS}}_{g,d}(\alpha) + \sum_s J^{\texttt{MIX}}_s(\alpha) + \sum_{i,k} J^{\texttt{MARGIN}}_{i,k}(\alpha). \tag{9}$$

The following section derives a closed-form expression of each term by introducing conjugate prior pdfs.

[1] Due to space constraints, we assumed diagonal covariances in each Gaussian and omitted transition probability. However, the proposed method is naturally extensible to the full-covariance case with transition probability estimation.

3.3. Conjugate prior pdfs

In order to practically realize MRED, the use of conjugate prior pdfs is essential because it yields a closed-form objective function when combined with the abovementioned tighter optimization problem.

Conjugate pdfs for parameters of Gaussian pdfs are represented by the Gaussian-gamma distribution, defined as follows:

$$P^0(\mu_{g,d},\tau_{g,d}|\mu^0_{g,d},\gamma^0_g,\eta^0_g,\beta^0_{g,d}) \propto (\tau_{g,d})^{\eta^0_g-1/2}\exp\left\{-\beta^0_{g,d}\tau_{g,d}-\frac{\tau_{g,d}\gamma^0_g}{2}(\mu^0_{g,d}-\mu_{g,d})^2\right\}. \tag{10}$$

By using this conjugate prior pdf, the integral with respect to the Gaussian pdf parameters in $J^{\texttt{GAUSS}}(\alpha)$ (Eq. (9)) can be solved as follows:

$$J^{\texttt{GAUSS}}_{g,d}(\alpha) = \gamma_g(\alpha)\log\sqrt{2\pi}+\log\frac{\sqrt{\gamma_g(\alpha)}}{\Gamma(\eta_g(\alpha))}+\eta_g(\alpha)\log\beta_{g,d}(\alpha), \tag{11}$$

where the following are the parameters of the posterior pdf, as functions of α, derived as follows:

$$\begin{aligned}\eta_g(\alpha) &= \eta^0_g+\frac{\Delta^0_g(\alpha)}{2},\quad \gamma_g(\alpha)=\gamma^0_g+\Delta^0_g(\alpha),\\ \mu_{g,d}(\alpha) &= \frac{\gamma^0_g\mu^0_{g,d}+\Delta^1_{g,d}(\alpha)}{\gamma_g(\alpha)},\\ \beta_{g,d}(\alpha) &= \beta^0_{g,d}+\frac{1}{2}\left(\gamma^0_g\left(\mu^0_{g,d}\right)^2+\Delta^2_{g,d}(\alpha)-\gamma_g(\alpha)\mu_{g,d}(\alpha)^2\right),\end{aligned} \tag{12}$$

where

$$\begin{aligned}\Delta^0_g(\alpha) &= \sum_i\sum_k \alpha^i_k\left(\chi^0_g[\boldsymbol{X}^i,\psi^i]-\chi^0_g[\boldsymbol{X}^i,\tilde{\psi}^i_k]\right),\\ \Delta^1_{g,d}(\alpha) &= \sum_i\sum_k \alpha^i_k\left(\chi^1_{g,d}[\boldsymbol{X}^i,\psi^i]-\chi^1_{g,d}[\boldsymbol{X}^i,\tilde{\psi}^i_k]\right),\\ \Delta^2_{g,d}(\alpha) &= \sum_i\sum_k \alpha^i_k\left(\chi^2_{g,d}[\boldsymbol{X}^i,\psi^i]-\chi^2_{g,d}[\boldsymbol{X}^i,\tilde{\psi}^i_k]\right).\end{aligned} \tag{13}$$

Here, $\chi^0_g[\boldsymbol{X},\psi]$, $\chi^1_{g,d}[\boldsymbol{X},\psi]$, and $\chi^2_{g,d}[\boldsymbol{X},\psi]$ denote the occupancy, 1st-order statistics, and 2nd-order statistics, respectively, of the $(g,d)^{\text{th}}$ Gaussian pdf calculated by using the given feature sequence $\boldsymbol{X}$ and occupancy pdf ψ.

As in the case of the Gaussian parameters, the Dirichlet distribution, is introduced as the conjugate prior pdfs of the mixture weight vectors $P^0(\boldsymbol{\rho}_s)$, defined as follows:

$$P^0(\boldsymbol{\rho}_s|\boldsymbol{\phi}^0_s)\propto\prod_m(\rho_{s,m})^{(\phi^0_{s,m}-1)}. \tag{14}$$

By introducing this prior pdf, the term in the objective function (Eq. (8)) and the estimated parameters of a posterior pdf are derived as follows:

$$\begin{aligned}J^{\texttt{MIX}}_s(\alpha) &= -\log\Gamma\left(\sum_m\phi_{s,m}(\alpha)\right)+\sum_m\log\Gamma\left(\phi_{s,m}(\alpha)\right),\\ \phi_{s,m}(\alpha) &= \phi^0_{s,m}+\Delta^0_{G(s,m)}(\alpha),\end{aligned} \tag{15}$$

where $G(s,m)$ is an index that indicates the Gaussian pdf corresponding to the s^{th} HMM state and the m^{th} mixture component.

In order to control the margin variable, we introduce an exponential distribution that favors a positive value in the margin variable ξ^i_k, defined as follows:

$$P^0(\xi^i_k|c^0_{i,k},\delta^0_{i,k})\propto\exp\left\{-c^0_{i,k}\left(\delta^0_{i,k}-\xi^i_k\right)\right\},\quad(\xi^i_k\le\delta^0_{i,k}). \tag{16}$$

Here, $c^0_{i,k}$ is a hyperparameter that adjusts the importance of the corresponding constraint, and $\delta^i_{i,k}>0$ is a hyperparameter that represents a desirable amount of margin. The closed-form expression of the term $J^{\texttt{MARGIN}}_{i,k}(\alpha)$ is derived as follows:

$$J^{\texttt{MARGIN}}_{i,k}(\alpha) = -\alpha^i_k\left(s^i_k-\delta^0_{i,k}\right)+\log\left(c^0_{i,k}-\alpha^i_k\right), \tag{17}$$

where s^i_k indicates an amount of shift in the margin computed by using a score obtained by language models and the entropies of each latent variable pdf, denoted as follows:

$$s^i_k = H[\psi^i]-H[\tilde{\psi}^i_k]+\log\frac{P(\boldsymbol{l}^i)}{\sum_{\boldsymbol{l}\neq\boldsymbol{l}^i}\sum_{\boldsymbol{q}}\tilde{\psi}^i_k(\boldsymbol{q})P(\boldsymbol{l}|\boldsymbol{q})}. \tag{18}$$

3.4. A solver based on cutting plane method

Since the auxiliary optimization problem has an infinite number of variables to be optimized, as discussed in Section 3.2, the conventional convex optimization method is not suitable. In this section, in order to efficiently solve this problem, we propose an optimization method based on the cutting plane method [11]; this method gradually adds the constraints to be considered as the optimization iterates.

In the proposed algorithm, $\tilde{\psi}^i_K$, where K denote the iteration count, is estimated and added as a hypothesis of the tightest constraint at the K^{th} iteration. Then, α^i_k $(k\le K)$ is estimated to satisfy all constraints due to $\tilde{\psi}^i_k$. A newly added hypothesis $\tilde{\psi}^i_K$ can be estimated by finding a pdf that minimizes the constraint function in Eq. (6), performed as follows:

$$\begin{aligned}\tilde{\psi}^i_K &= \text{argmin}_{\tilde{\psi}}\left\langle\hat{R}^i[\Theta,\psi^i]-\check{C}^i[\Theta,\tilde{\psi}]-\xi^i\right\rangle_{P^K(\Theta,\xi)},\\ &= \text{argmax}_{\tilde{\psi}}\left\langle\check{C}^i[\Theta,\tilde{\psi}]\right\rangle_{P^K(\Theta,\xi)},\end{aligned} \tag{19}$$

where $P^K(\Theta,\xi)$ is the current estimation of $P(\Theta,\xi)$ obtained by assuming $\alpha^i_k=0$ $(\forall k\ge K)$. Because Eq. (19) implies maximum-likelihood estimation (MLE) of the latent pdf $\tilde{\psi}^i_K$, $\tilde{\psi}^i_K$ can be obtained by computing an occupation pdf by applying the forward-backward algorithm (FB) to lattices that represent a set of erroneous label sequences $(\forall\boldsymbol{l}\neq\boldsymbol{l}^i)$.

We note that the estimation of latent pdfs ψ^i corresponding to the i^{th} reference label is also important in order to obtain an accurate approximation of the constraint function. Therefore, we incorporate the update step of ψ^i in the iterative procedure in addition to the estimation of $\tilde{\psi}^i_K$. Finally, by alternately performing update of ψ^i, adding of $\tilde{\psi}^i_k$, and optimization of α^i_k, we can perform the optimization of the auxiliary problem (Eq. (6)). In particular, the following steps are iterated:

1. Update ψ^i by applying FB to the reference label,
2. Estimate $\tilde{\psi}^i_K$ by applying FB to the competitor lattices,
3. Optimize α^i_k by using ψ^i and $\tilde{\psi}^i_k$ $(\forall k\le K)$,
4. Increment K.

This algorithm enables the use of the infinite constraints by only considering finite variables to be optimized. Further, this algorithm is reasonable since the tightest constraints in each iteration dominates the other constraints in most cases.

Note that the conventional MRED method originally proposed for GMM training [9], which updates $\tilde{\psi}^i$, ψ^i and $P(\Theta,\xi)$ alternately, can also be applied to CD-HMMs. This conventional method can be regarded as a special case of the proposed method by considering a single constraint for each training datum that minimizes the constraint function. Thus, the conventional method can be implemented by using $\tilde{\psi}^i_k$ $(k=K)$ in the step 3 instead of $\tilde{\psi}^i_k$ $(\forall k\le K)$.

4. Experiments

In the experiments, we used 3,696 sentences (173,492 phonemes) from the TIMIT database for model training, and 192 sentences (7,215 phonemes; a.k.a. TIMIT core testset) for evaluation. All the training and test speech waveforms are

Table 1: Phoneme error rates of the compared methods

Method	1 mix.	2 mix.	4 mix.	8 mix.
MLE	41.8	38.1	35.4	33.0
	(40.4)	(36.5)	(33.7)	(31.2)
MMIE	38.9	35.6	33.6	32.0
	(37.5)	(34.1)	(32.1)	(30.2)
MMIE (I-smooth)	38.8	35.5	33.6	32.0
	(37.5)	(33.9)	(32.0)	(30.2)
MRED (single) [9]	39.6	36.5	34.3	32.6
	(39.2)	(34.8)	(32.4)	(30.5)
MRED (proposed)	**37.7**	**34.7**	**32.5**	**31.5**
	(**36.7**)	(**33.5**)	(**31.3**)	(**29.9**)

parametrized by Mel-frequency cepstral coefficients (12 dims MFCC) and its log-energy augmented by their derivatives and accelerations (39 dims) computed at a 10-ms frame shift with a 25-ms window size. As described in [12], we used 48 phonetic classes for training and decoding, and the phoneme error rates were calculated by using 39 broader phonetic categories. All HMMs have left-to-right 3 states for each 48 monophone model. A bi-gram (bi-phoneme) grammar model is applied during all decoding processes.

For comparison, the MLE-based system and a discriminative training method based on the maximum mutual information estimation [1] (MMIE) are implemented. Further, I-smoothing [3] technique is used to regularize the MMIE estimation. In order to evaluate the difference between realization methods, the conventional MRED training method, which was originally proposed to estimate GMMs [9], is applied to CD-HMMs (cf. Section 3.4) and compared with the proposed method (MRED (single)).

As in the case of the I-smoothing technique, the following MLE-based hyperparameters are used in the MRED systems in order to reduce the tuning uncertainty.

$$\begin{aligned} &\mu^0_{g,d} = \chi^1_{g,d}/\gamma^0, \quad \gamma^0_g = \chi^0_g, \quad \eta^0_g = \chi^0_g/2, \\ &\beta^0_{g,d} = \left(\chi^2_{g,d} - \chi^0_g(\mu^0)^2\right)/2, \quad \phi^0_{s,m} = \chi^0_{G(s,m)}, \end{aligned} \tag{20}$$

where χ^0_g, $\chi^1_{g,d}$, and $\chi^2_{g,d}$ are occupancy, 1st-order statistics, and 2nd-order statistics of the $(g,d)^{\text{th}}$ Gaussian pdf obtained by using the FB algorithm performed with a model obtained by the maximum likelihood procedure. Further, the following hyperparameter settings are used.

$$\delta^0_{i,k} = N(\boldsymbol{X}^i), \quad c^0_{i,k} = \text{constant}, \tag{21}$$

where $N(\boldsymbol{X}^i)$ is the number of frames in the i^{th} feature sequence in the training dataset. The remaining hyperparameters (language model scale factor, prior parameter $c^0_{i,k}$, learning rate parameter used in the MMIE system, and smoothing factor of I-smoothing) and the number of iterations are determined by using the development set.

Table 1 lists the phoneme error rates of the compared methods. In the table, the results evaluated by using the complete test data (62,901 phonemes) of TIMIT is presented in parentheses as supplemental information. First, we confirmed that the MRED systems successfully reduce errors as compared to MLE systems. Thus, it is confirmed that the proposed method provides posterior pdfs of the parameters that have sufficient discriminative performance.

We confirmed that the I-smoothing technique was not very helpful in this experimental setting. We consider that this is because sufficient data is provided without the overfitting effects. Even in such a configuration, it is confirmed that MRED outperformed the MMIE systems. This advantage might be attributed to the differences in their optimization methods. Because the proposed approach is convex when the statistics are fixed, it might leap local optima. Further, the proposed method outperformed the conventional MRED method that is used for training of GMMs (MRED (single)). It is considered that the proposed approach provided a more accurate approximation than the conventional one by considering wider variations of the latent variables of CD-HMMs, as discussed in Section 3.4. In fact, since relaxation of constraints that would results in oversmoothing of the posterior pdf appeared in the MRED (single) systems, the performances were lower than that of MMIE systems which correctly maximize discriminative performances. Thus, we confirmed that MRED is successfully applied to training of acoustic models used in continuous speech recognition by using the proposed method.

5. Conclusions

In this paper, we derived a discriminative training method by using the principle of minimum relative entropy discrimination (MRED) in order to provide a Bayesian interpretation of discriminative training methods. We proposed an accurate realization method of MRED to derive an efficient training method for continuous-density hidden Markov models. We confirmed that the proposed MRED method outperformed the conventional MMIE method with the I-smoothing techniques and the conventional MRED method originally proposed aiming for the estimation of Gaussian mixture models.

In the future, we intend to incorporate knowledge about the significance of errors as in the case of MPE [3] and boosted MMI/MPE [13] methods. We also intend to carry out experiments that involve large vocabulary continuous speech recognition tasks. Furthermore, comparative investigations on the prior parameters should be carried out.

Acknowledgement The authors would like to thank the valuable discussions in the Lehrstuhl für Informatik 6, RWTH Aachen University. This study was partially supported by a Grant-in-Aid for JSPS Fellows (21·04190) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

6. References

[1] L. Bahl, P. Brown, P. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. ICASSP*, 1986, pp. 49–52.

[2] E. McDermott and S. Katagiri, "String-level MCE for continuous phoneme recognition," in *Proc. EUROSPEECH*, 1997, pp. 123–126.

[3] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, 2002, pp. I–105–I–108.

[4] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th COLT*, 1992, pp. 144–152.

[5] F. Sha, "Large margin training of acoustic models for speech recognition," Ph.D. dissertation, Univ. of Pennsylvania, 2006.

[6] D. Yu, L. Deng, X. He, and A. Acero, "Large-margin minimum classification error training: A theoretical risk minimization perspective," *Computer Speech and Language*, pp. 415–429, 2008.

[7] T. S. Jaakkola, M. Meila, and T. Jebara, "Maximum entropy discrimination," *Advances in Neural Information Processing Systems*, vol. 12, pp. 470–476, 2000.

[8] F. Valente and C. Wellekens, "Maximum entropy discrimination (MED) feature subset selection for speech recognition," in *Proc. ASRU*, 2003, pp. 327–332.

[9] D. P. Lewis, "Combining kernels for classification," Ph.D. dissertation, Columbia Univ., 2008.

[10] Y. Kubo, S. Watanabe, A. Nakamura, E. McDermott, and T. Kobayashi, "A sequential pattern classifier based on hidden markov kernel machine and its application to phoneme classification," *IEEE J. Sel. Topics in Signal Process.*, 2010 (to appear).

[11] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *J. Mach. Learn. Research*, vol. 6, pp. 1453–1484, 2005.

[12] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *Acoust., Speech, Signal Process., IEEE Trans. on*, vol. 37, no. 11, pp. 1641–1648, 1989.

[13] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. ICASSP*, 2008, pp. 4057–4060.