DISCOVERING RELATIONS AMONG DISCRIMINATIVE TRAINING OBJECTIVES

Qi Li

Li Creative Technologies (LcT), Inc.
New Providence, NJ 07974, USA
qili@ieee.org; www.licreativetech.com

ABSTRACT

In this paper, the relations among several discriminative training objectives for speech and speaker recognition, language processing, and dynamic pattern recognition are derived and discovered through theoretical analysis. Those objectives are the *minimum classification error* (MCE), *maximum mutual information* (MMI), *minimum error rate* (MER), and a recently proposed *generalized minimum error rate* (GMER) objectives. The results show that all the objectives are related to the *a posteriori* probability and error rates, and the MCE and GMER objectives are more general and flexible than the MMI and MER objectives. These results can help in understanding the discriminative objectives, in improving recognition performances, and in discovering new training algorithms jointly with objectives.

1. INTRODUCTION

It has been reported that the discriminative training techniques provide significant improvements in recognition performance compared to the traditional maximum likelihood (ML) objective in speech and speaker recognition as well as language processing. Those discriminative objectives include the *minimum classification error* (MCE) [1], *maximum mutual information* (MMI) [2], *minimum error rate* (MER) [8], and a recently proposed *generalized minimum error rate* (GMER) [3, 4] objectives, as well as others.

Among those objectives, the MCE and MMI have been used for years and have both shown good performances over the ML objective. Consequentially, some research have been conducted to compare the performances through experiments or some degree of theoretical analysis (e.g. [5, 6, 7]); however, the experimental comparisons are limited to particular tasks and the results are not general enough to help us understand the detailed mechanisms; on the other hand, the previous theoretical analyses are not conclusive or adequate enough to show the relations among those objectives. In this paper, we intend to derive and discover the relations among the four objectives theoretically and conclusively without any bias to any particular tasks. The theoretical results are further validated by experimental results.

In the following, the *a posteriori* probability will be used as the tool to facilitate the comparisons. We will first review the relation between error rates and the *a posteriori* probability, and then derive the relations between each of the discriminative objectives and the *a posteriori* probability; furthermore, we will establish the relations among the different objectives.

Error Rates vs. the a Posteriori Probability: In an M-class classification problem, we are asked to make a decision to identify a sequence of observations, $\mathbf{x}$, as a member of a class, say, C_i. The true identity of $\mathbf{x}$, say C_j, is unknown, except in the design or training phase in which observations of known identity are used as reference for parameter optimization. We denote event α_i as the action of identifying an observation as class C_i. The decision is correct if $i = j$; otherwise, it is incorrect. It is natural to seek a decision rule that minimizes the probability of error, or empirically, the error rate, which entails a zero-one loss function:

$$\mathcal{L}(\alpha_i|C_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j. \end{cases} \qquad i, j = 1, ..., M \tag{1}$$

It assigns no loss to a correct decision and assigns a unit loss to an error. The probabilistic risk of α_i corresponding to this loss function is

$$\begin{aligned} R(\alpha_i|\mathbf{x}) &= \sum_{j=1}^{M} \mathcal{L}(\alpha_i|C_j) P(C_j|\mathbf{x}) \\ &= \sum_{j \neq i} P(C_j|\mathbf{x}) = 1 - P(C_i|\mathbf{x}) \end{aligned} \tag{2}$$

where $P(C_i|\mathbf{x})$ is the *a posteriori* probability that $\mathbf{x}$ belongs to C_i. Thus, the zero-one loss function links the error rates to the *a posteriori* probability. To minimize the probability of error, one should therefore maximize the *a posteriori* probability $P(C_i|\mathbf{x})$. This is the basis of Bayes' maximum *a posteriori* (MAP) decision theory and is also referred to as *minimum error rate* (MER) [8] in an ideal setup.

Qi (Peter) Li was with Bell Labs, Lucent Technologies, Murray Hill, NJ 07974.

We note that the *a posteriori* probability $P(C_i|\mathbf{x})$ is often modeled as $P_{\lambda_i}(C_i|\mathbf{x})$, a function defined by a set of parameters λ_i. Since the parameter set λ_i has a one-to-one correspondence with C_i, we write $P_{\lambda_i}(C_i|\mathbf{x}) = P(\lambda_i|\mathbf{x})$ and other similar expressions without ambiguity.

If we consider all M classes and all data samples, an objective for MER can be defined as:

$$\max J(\Lambda) = \frac{1}{N}\sum_{k=1}^{M}\sum_{i=1}^{N_k} P(\lambda_k|\mathbf{x}_{k,i}) \tag{3}$$

where N_k is the total number of training data of class k, $N = \sum_{k=1}^{M} N_k$, and $\mathbf{x}_{k,i}$ is the ith feature vector of class k. Λ is a set of model parameters, $\Lambda = \{\lambda_k\}_{k=1}^{M}$.

Neural Networks vs. the a Posteriori Probability: We note that it has shown that multilayer neural networks trained by backpropagation on a sum-squared error objective can approximate the true *a posteriori* probability in a least-square sense [8]. In this paper, we focus on the objectives that have been applied to speech and speaker recognition or other dynamic pattern recognition problems.

2. MINIMUM CLASSIFICATION ERROR VS. THE A POSTERIORI PROBABILITY

The minimum classification error (MCE) objective was derived through a systematic analysis on classification errors. It introduced a misclassification measure to embed the decision process in the overall minimum classification error formulation. During the derivation, it was also considered that the misclassification measure is continuous with respect to the classifier parameters. The empirical average cost as the typical objective in the MCE algorithm was defined as [1]:

$$\min L(\Lambda) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{M} \ell_k(d_k(\mathbf{x}_i);\Lambda)1(\mathbf{x}_i \in C_k). \tag{4}$$

where M and N are the total numbers of classes and training data, and $\Lambda = \{\lambda_k\}_{k=1}^{M}$. It can be rewritten as:

$$\min L(\Lambda) = \frac{1}{N}\sum_{k=1}^{M}\sum_{i=1}^{N_k} \ell_k(d_k(\mathbf{x}_{k,i});\Lambda) \tag{5}$$

where N_k is the total number of training data of class k, $N = \sum_{k=1}^{M} N_k$, and $\mathbf{x}_{k,i}$ is the ith feature vector of class k. ℓ_k is a loss function and a sigmoid function is often used for it:

$$\ell_k(d_k) = \frac{1}{1+e^{-\zeta d_k+\alpha}}, \qquad \zeta > 0 \tag{6}$$

where d_k is a class misclassification measure defined as [9]:

$$d_k(\mathbf{x}) = -g_k(\mathbf{x};\Lambda) + \log\left[\frac{1}{M-1}\sum_{j\neq k}\exp[\eta g_j(\mathbf{x};\Lambda)]\right]^{1/\eta} \tag{7}$$

where $\mathbf{x} = \mathbf{x}_{k,i}$. When function $g(.)$ in (7) is a logarithm of probability as used in many applications, the class misclassification measure in (7) can be rewritten as:

$$d_k(\mathbf{x}) = -\log p(\mathbf{x}|\lambda_k) + \log\left[\frac{1}{M-1}\sum_{j\neq k} p(\mathbf{x}|\lambda_j)^{\eta}\right]^{1/\eta}. \tag{8}$$

When $\eta = 1$, we have

$$d_k(\mathbf{x}) = -\log\frac{p(\mathbf{x}|\lambda_k)}{\sum_{j\neq k}\frac{1}{M-1}p(\mathbf{x}|\lambda_j)}. \tag{9}$$

It can be further presented as:

$$d_k(\mathbf{x}) = -\log\frac{p(\mathbf{x}|\lambda_k)P_k}{\sum_{j\neq k} p(\mathbf{x}|\lambda_j)P_j} \tag{10}$$

where $P_k = 1$ and $P_j = \frac{1}{M-1}$, and they are similar to the *a priori* probability if we conduct a normalization.

To facilitate our further comparison, we convert the minimization problem to a maximization problem. Let

$$\tilde{d}_k(\mathbf{x}) = -d_k(\mathbf{x}) = \log\frac{p(\mathbf{x}|\lambda_k)P_k}{\sum_{j,j\neq k} p(\mathbf{x}|\lambda_j)P_j}, \tag{11}$$

and take it into the sigmoid function in (6). Assuming $\zeta = 1$ and $\alpha = 0$, we have

$$\begin{aligned}
\ell_k(\tilde{d}_k) &= \frac{1}{1+e^{-\tilde{d}_k}} && (12)\\
&= \frac{p(\mathbf{x}|\lambda_k)P_k}{p(\mathbf{x}|\lambda_k)P_k + \sum_{j\neq k} p(\mathbf{x}|\lambda_j)P_j} && (13)\\
&= \frac{p(\mathbf{x}|\lambda_k)P_k}{\sum_{j=1}^{M} p(\mathbf{x}|\lambda_j)P_j}. && (14)
\end{aligned}$$

Thus, the objective in (5) is simplified to:

$$\begin{aligned}
\max \tilde{L}(\Lambda) &= \frac{1}{N}\sum_{k=1}^{M}\sum_{i=1}^{N_k}\frac{p(\mathbf{x}_{k,i}|\lambda_k)P_k}{\sum_{j=1}^{M} p(\mathbf{x}_{k,i}|\lambda_j)P_j} && (15)\\
&= \frac{1}{N}\sum_{k=1}^{M}\sum_{i=1}^{N_k} P(\lambda_k|\mathbf{x}_{k,i}). && (16)
\end{aligned}$$

This demonstrates that the MCE objective can be simplified to MER as defined in (3) and linked to the *a posteriori*

probability if we make the following assumptions:

$$P_k = 1 \quad (17)$$

$$P_j = \frac{1}{M-1} \quad (18)$$

$$\eta = 1 \quad (19)$$

$$\zeta = 1 \quad (20)$$

$$\alpha = 0. \quad (21)$$

Among the parameters, $P_k = 1$ and $P_j \leq 1$ imply that the MCE objective weighs the true class higher or equal to the competing classes. The parameter η plays a role of Holder norm in (7). By changing η, the weights between the true class and competing classes can be further adjusted. The rest of the parameters, ζ and α, are related to the sigmoid function. α represents the shift of the sigmoid function. Since other parameters can play a similar role, α is usually set to zero. ζ is related to the slope of the sigmoid function. For different tasks and data distributions, different values of ζ can be selected to achieve the best performance. ζ is one of the most important parameters in the MCE objective, and it makes the MCE objective flexible and adjustable to different tasks and data distributions.

3. MAXIMUM MUTUAL INFORMATION VS. MINIMUM CLASSIFICATION ERROR

The objective of *maximum mutual information* (MMI) was defined in [2] as:

$$I(k) = \log \frac{p(\mathbf{x}_{k,i}|\lambda_k)P_k}{\sum_{j=1}^{M} p(\mathbf{x}_{k,i}|\lambda_j)P_j}. \quad (22)$$

If we consider all M models and all data as in the above discussions, the complete objective for MMI training is:

$$\max I(\Lambda) = \sum_{k=1}^{M}\sum_{i=1}^{N_k} \log \frac{p(\mathbf{x}_{k,i}|\lambda_k)P_k}{\sum_{j=1}^{M} p(\mathbf{x}_{k,i}|\lambda_j)P_j} \quad (23)$$

$$= \sum_{k=1}^{M}\sum_{i=1}^{N_k} \log P(\lambda_k|\mathbf{x}_{\mathbf{k},\mathbf{i}}). \quad (24)$$

By comparing (16) and (24), we observe that the difference between the simplified version of the MCE objective and the MMI objective is primarily in the logarithm. Since the logarithm is a monotonically increasing function and N is a constant, a procedure to optimize (24) is equivalent to optimize (16); therefore, the MMI objective in (24) is equivalent to:

$$\max \tilde{I}(\Lambda) = \frac{1}{N}\sum_{k=1}^{M}\sum_{i=1}^{N_k} P(\lambda_k|\mathbf{x}_{k,i}) \quad (25)$$

which equals the simplified version of MCE objective in (16) or the MER objective in (3).

4. GENERALIZED MINIMUM ERROR RATE VS. OTHER OBJECTIVES

In order to derive a set of close-form formulas for fast parameter estimation, we defined the generalized minimum error rate (GMER) objective as [3, 4]:

$$\max \tilde{J}(\Lambda) = \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N_m} \ell(d_{m,n}) \quad (26)$$

where $\ell(d_{m,n}) = \frac{1}{1+e^{-\zeta d_{m,n}}}$ is a sigmoid function, and

$$d_{m,n} = \log p(\mathbf{x}_{m,n}|\lambda_m)P_m - L_m \log \sum_{j\neq m} p(\mathbf{x}_{m,n}|\lambda_j)P_j, \quad (27)$$

where $0 < L_m \leq 1$ is a weighting scalar. Intuitively, L_m represents the weighting between true class m and competing classes $j \neq m$. When $L_m < 1$, it means that true class m is more important than the competing classes. When $L_m = 1$, it means the true class and competing classes are equally important. The exact value of L_m can be determined during estimation based on the constraint that estimated covariance matrixes must be positive definite.

The sigmoid function plays a role similar to its role in the MCE objective. It actually provides different weights to different training data. For that data that is hardly ambiguous in its classification, the weight is close to 0 (i.e., decisively wrong) or 1 (i.e., decisively correct); for the data near the classification boundary, the weighting is in-between. The slope of the sigmoid function is controlled by the parameter $\zeta > 0$. Its value can be adjusted based upon the data distributions in specific tasks.

When $L_m = 1$ and $\zeta = 1$, we have that the $\tilde{J}$ in (26) equals to J in (3), and the GMER objective in (26) is simplified to the MMI or MER objectives. The GMER objective is equal to the MCE objective if we have: $L_m = 1$, $P_k = 1$, and $P_j = \frac{1}{M-1}$ on the GMER objective, and have $\eta = 1$ and $\alpha = 0$ on the MCE objective.

The new GMER objective is more general and flexible than both MMI and MER objectives. It also remtains the most important parameters ζ from the MCE objective. The weighting parameter η in MCE was replaced by L_m in the GMER objective. The most important factor is that the GMER objective is simpler than MCE; thus, we can derive a new set of closed-form formulas for fast parameter estimation for discriminative training [3, 4].

5. EXPERIMENTAL COMPARISONS

The results from the above theoretical analysis are consistent with the experimental results reported from different research sites.

In speaker verification, ML (maximum likelihood), MMI, and MCE objectives were compared using the NIST 1996

Table 1. COMPARISONS ON TRAINING ALGORITHMS

Objectives	Algorithms	Learning Parameters	Relation to Post. Prob.
ML	Closed form/EM	None	Not related
MMI	Closed form	D	Same
MCE	Gradient descent/GPD	Learning rates	Extended
GMER	Closed form	None	Extended

evaluation dataset by Ma, *et al.* [7]. There are 21 male target speakers and 204 male impostors. The reported relative equal-error-rate reductions compared to the ML objective are 3.2% and 7.0% for MMI and MCE, respectively.

In speech recognition, MLE, MMI, and MCE objectives were compared using a common database by Reichl and Ruske [5]. It was found that both MMI and MCE objectives can have speech recognition performance improvements over the ML objective. The absolute error-rate reduction in the MMI objective is 2.5% versus 5.3% in the MCE objective.

In speaker identification, we compared the ML and GMER objectives using an 11-speaker group from the NIST 2000 dataset. For the testing durations of 1, 5, and 10 seconds, the ML objective had error rates of 31.4%, 6.59%, and 1.39% while the GMER objective has error rates of 26.81%, 2.21%, and 0.00%. The relative error rate reductions are 14.7%, 66.5%, and 100%, respectively. For the best results, the weighting scalars L_m were determined by the algorithm ($L_m \neq 1.0$), and the slope of sigmoid function is $\zeta = 0.8$. Based on our above analysis, this implies that the GMER objective also outperforms the MMI and MER objectives.

The above experimental results showed that by adjusting the additional parameters, the MCE and GMER can provide better performances than others.

6. DISCUSSIONS AND CONCLUSIONS

For pattern recognition or classification, the objectives and optimization methods for parameter estimation are related to each other, and they both play important roles in solving real-world problems, in terms of recognition accuracy and training speed. We summarize the comparisons in Table 1.

Regarding optimization methods, in general, closed-form formulas are more efficient than a gradient-descent kind of approach. However, not every objective has the closed-form formulas. When an objective is complicated, such as the MCE objective, it has less of a chance to derive closed-form formulas. For the MMI objective, a closed-form parameter estimation algorithm was derived; however, there is a constant D in the algorithm and the value of the constant needs to be pre-determined for parameter estimation. Like the learning rate in gradient-descent methods, it is difficult to determine the value of D as reported in literature. The GMER algorithm is developed under our belief that for the best performances, in terms of recognition accuracy and training speed, the objective and optimization method should be developed jointly. The GMERs recognition accuracy is similar to MCE while the training speech is close to the EM algorithm used in the ML estimation.

It has been argued that it is not intuitive how the MMI objective relates to error rates. From the above discussions, the answer is straightforward because we have linked the MMI objective to the *a posteriori* probability and error rates. If we want to further investigate the differences between the MCE objective in (5) and MMI objectives in (24), the differences are mainly in the parameter set listed from (17) to (21). In theory, those parameters provide the flexibility to adjust the MCE object for different recognition tasks and data distributions; therefore, MCE object is more general compared to the MMI and MER objectives.

In conclusion, the theoretical analysis in this paper indicates that the discriminative objectives used in speech and speaker recognition are all related to the *a posteriori* probability and error rates. While the MMI is directly from the *a posteriori* probability, the MCE and GMER objectives can be equivalent to the *posteriori* probability under some assumptions. The results from this paper show that the MCE and GMER objectives extend the *a posteriori* probability-based objectives, and they are more general and flexible than the MMI objective. As validated in experiments, the extension and flexibility can benefit real applications for different recognition tasks or data distributions.

The author would like to thank Dr. Biing-Hwang Juang for useful discussions.

7. REFERENCES

[1] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. on Signal Processing*, vol. 40, pp. 3043–3054, December 1992.

[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," in *Proc. IEEE ICASSP*, pp. 49–52, 1986.

[3] Q. Li and B.-H. Juang, "A new algorithm for fast discriminative training," in *Proc. IEEE ICASSP*, May 2002.

[4] Q. Li and B.-H. Juang, "Fast discriminative training for sequential observations with application to speaker identification," in *Proc. IEEE ICASSP*, April 2003.

[5] W. Reichl and G. Ruske, "Discriminant training for continuous speech recognition," in *Proc. of Eurospeech*, 1995.

[6] R. Schluter and W. Macherey, "Comparison of discriminative training criteria," in *Proc. IEEE ICASSP*, pp. 493–497, 1998.

[7] C. Ma and E. Chang, "Comparison of discriminative training methods for speaker verification," in *Proc. IEEE ICASSP*, 2003.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification, Second Edition.* New York: John & Wiley, 2001.

[9] W. Chou, "Discriminant-function-based minimum recognition error rate pattern-recognition approach to speech recognition," *Proceedings of the IEEE*, vol. 88, pp. 1201–1222, August 2000.