CHAPTER 3

Discriminative Learning: A Unified Objective Function

In this chapter, a unified account is provided for three classes of objective functions developed in discriminative training of hidden Markov models (HMMs). These are: maximum mutual information (MMI), minimum classification error (MCE), and minimum phone error/minimum word error (MPE/MWE). We also compare our unified form of these objective functions with another popular unified form in the literature.

3.1 INTRODUCTION

Popular discriminative parameter learning techniques are (1) MMI [6, 14, 17, 34, 49, 52]; (2) MCE [8, 20, 24, 25, 31–33, 42, 44, 46], and (3) MPE and closely related MWE [12, 38–41]. In addition to a general overview on the above classes of techniques, this book has a special focus on three key areas in discriminative learning: objective function, optimization method, and algorithmic properties. This chapter is devoted to the first area, where we provide a unified view of the three discriminative learning objective functions, MMI, MCE, and MPE/MWE, for classifier parameter optimization, from which structural insight and the relationships among them are derived. In this chapter, we concentrate on a unified objective function that gives rise to various special cases associated with different levels of performance optimization for pattern recognition tasks — including the performance optimization levels of superstring unit, string unit, and substring unit.

After giving an introduction to the discriminative learning criteria of MMI, MCE, and MPE/MWE, we show that under certain assumptions, the objective functions of MMI, MCE, and MPE/MWE criteria (with multiple training tokens) can be formulated and unified into a rational-function form. From that, relations among MMI, MCE, and MPE/MWE criteria are studied. In discussing these topics, some familiarities of HMMs are assumed, such as those described in standard textbooks (e.g., [43, 47]).

3.2 A UNIFIED DISCRIMINATIVE TRAINING CRITERION

The main purpose of this chapter is to provide a general and concise introduction to three types of optimization criteria, MMI, MCE, and MPE/MWE, for discriminative parameter learning, and then to formulate a unified criterion that subsumes the three criteria as special cases. The process of this formulation offers insight into the fundamental relationship among MMI, MCE, and MPE. Another insight gained is on how these special cases correspond to distinct levels of pattern recognition performance optimization. MMI gives performance optimization for superstring sequences. MCE gives performance optimization for string sequences. And MPE/MWE gives performance optimization for substring sequences.

3.2.1 Notations

As the notations throughout this book, we denote by Λ the parameter set of the generative model (e.g., HMM or a Gaussian distribution) expressed in terms of a joint statistical distribution:

$$p(X, S|\Lambda) = p(X|S, \Lambda)P(S) \tag{3.1}$$

on the observation training data sequence X and on the corresponding label sequence S, where we assume the parameters in the "language model" $P(S)$ are not subject to optimization. We denote by R the number of training tokens and use $r = 1, \ldots, R$ as the index for "token" or "string" (e.g., a single sentence or utterance) in the training data, and each token consists of a "string" of an observation data sequence: $X_r = x_{r,1}, \ldots, x_{r,Tr}$ of length T_r with the corresponding label (e.g., word) sequence: $S_r = w_{r,1}, \ldots, w_{r,Nr}$ of length N_r. That is, S_r denotes the correct label sequence for token r; in effect, $w_{r,i}$ is the ith word in the word sequence of S_r. Furthermore, we use s_r to denote all possible label sequences for the rth token, including the correct label sequence S_r and all other incorrect label sequences. For the iterative learning approach discussed in this book, we denote by Λ' the model parameters computed from the immediately previous iteration.

3.2.2 The Central Result

The central result presented in this chapter is that all three discriminative learning criteria, MMI, MCE, and MPE/MWE, can be formulated as the following unified form of a rational function as the objective function (which can be readily subject to a special way of optimization discussed later):

$$O(\Lambda) = \frac{\sum_{s_1, \ldots, s_R} p(X_1, \ldots, X_R, s_1, \ldots, s_R|\Lambda) \cdot C_{\text{DT}}(s_1, \ldots, s_R)}{\sum_{s_1, \ldots, s_R} p(X_1, \ldots, X_R, s_1, \ldots, s_R|\Lambda)} \tag{3.2}$$

where the summation over $s = s_1,\ldots, s_R$ in (3.2) denotes the coverage of all possible label sequences (both correct and incorrect ones) in all R training tokens. (This huge number of terms will be drastically simplified during the optimization step, which we shall discuss in detail later.) In (3.2), $X = X_1,\ldots, X_R$ denotes the collection of all observation data sequences (strings) in all R training tokens, which we also call "superstring." $p_\Lambda(X_1,\ldots X_R, s_1\ldots, s_R)$ is the joint distribution for the superstring of data $X_1,\ldots, X_R$ and an arbitrary possible super label sequence assignments $s_1,\ldots, s_R$. The discriminative training (DT) function $C_{DT}(s_1\ldots, s_R)$ in (3.2) differentiates MMI, MCE, and MPE/MWE, each with a specific form of $C_{MMI}(s_1\ldots, s_R)$, $C_{MCE}(s_1\ldots, s_R)$, and $C_{MPE}(s_1\ldots, s_R)$, respectively. We will derive these specific forms in the subsequent sections of this chapter. Note that $C_{DT}(s_1\ldots, s_R)$ in (3.2) is a quantity that depends only on the label sequence $s_1, \ldots, s_R$, and are independent of the parameter set Λ to be optimized.

We now introduce the criteria of MMI, MCE, and MPE/MWE separately as in the standard literature, and then provide detailed derivations to reformulate each of them into the unified rational function form of (3.2). This then will enable the use of powerful and unified optimization techniques based on GT applied specifically to rational functions.

3.3 MMI AND ITS UNIFIED FORM

3.3.1 Introduction to MMI Criterion

In information theory, mutual information *I(X,S)* between data X and their corresponding labels/symbols S measures the amount of information obtained, or the amount of reduction in uncertainty, through a noisy information-transfer channel after observing output labels/symbols. In designing the noisy channel, it is obvious that one desires to increase the information attainment by maximizing *I(X,S)*. Quantitatively, mutual information is defined as

$$I(X,S) = \sum_{X,S} p(X,S) \log \frac{p(X,S)}{p(X)p(S)} = \sum_{X,S} p(X,S) \log \frac{p(S|X)}{p(S)} = H(S) - H(S|X) \qquad (3.3)$$

where $H(S) = -\sum_S p(S) \log p(S)$ is the entropy of S, and $H(S|X)$ is the conditional entropy:

$$H(S|X) = -\sum_{X,S} p(X,S) \log P(S|X) \qquad (3.4)$$

Assume that $P(S)$ ("language model") and hence $H(S)$ is given (i.e., with no parameters to optimize). Then maximizing mutual information of (3.3) becomes equivalent to minimizing conditional entropy of (3.4) with respect to its parameters. Because $P(S|X)$ in (3.4) is unknown, $H(S|X)$ can only be estimated using the sample average:

$$H(S|X) \cong \hat{H}_\Lambda(S|X) = -\frac{1}{R}\sum_{r=1}^{R} \log p(S_r|X_r) = -\frac{1}{R}\sum_{r=1}^{R} \log \frac{p(X_r|S_r,\Lambda)P(S_r)}{p(X_r)}$$

Hence, maximizing mutual information (MMI) is equivalent to maximizing

$$O_{\mathrm{MMI}}(\Lambda) = \sum_{r=1}^{R} \log \frac{p(X_r|S_r,\Lambda)P(S_r)}{P(X_r)} = \sum_{r=1}^{R} \log \frac{p(X_r|S_r,\Lambda)P(S_r)}{\sum_{s_r} p(X_r|s_r,\Lambda)P(s_r)} \tag{3.5}$$

where $P(s_r)$ is the "language model" probability for an arbitrary sentence token s_r. The MMI criterion equals the logarithm of the posterior probability of the correct sentence S_r, or "good model," given their observation sequences. This posterior probability takes into account all models, good (S_r) or bad (s_r excluding S_r), as shown in the denominator of (3.5). (In practice, a scale κ has been applied empirically to all the probability terms in (3.5) for generalization purposes in implementing MMI [40]. This issue will not be addressed in this paper).

3.3.2 Reformulation of the MMI Criterion into Its Unified Form

It is straightforward to reformulate the problem of optimizing (3.5) into that of optimizing the unified form of (3.2), because (3.5) is essentially a rational function due to the logarithm. To continue the reformulation, we construct the monotonically increasing function of exponentiation for (3.5). This gives

$$\tilde{O}_{\mathrm{MMI}}(\Lambda) = \exp[O_{\mathrm{MMI}}(\Lambda)] = \prod_{r=1}^{R} \frac{p(X_r,S_r|\Lambda)}{\sum_{s_r} p(X_r,s_r|\Lambda)} = \frac{p(X_1,\ldots,X_R,S_1,\ldots,S_R|\Lambda)}{\sum_{s_1,\ldots,s_R} p(X_1,\ldots,X_R,s_1,\ldots,s_R|\Lambda)} \tag{3.6}$$

The latter step uses the assumption that different training tokens are independent of each other. It is noteworthy that each multiplier in (3.6) can be viewed as a model-based expected gain, that is,

$$\frac{p(X_r,S_r|\Lambda)}{\sum_{s_r} p(X_r,s_r|\Lambda)} = 1 - \sum_{s_r \neq S_r} P(s_r|X_r,\Lambda) = 1 - \overbrace{\sum_{s_r} \underbrace{(1-\delta(s_r,S_r))}_{0-1 \text{ loss}} P(s_r|X_r,\Lambda)}^{\Lambda \text{ based expected loss}}$$

We now rewrite (3.6) in the form of a rational function

$$\tilde{O}_{\mathrm{MMI}}(\Lambda) = \frac{\sum_{s_1,\ldots,s_R} p(X_1,\ldots,X_R,s_1,\ldots,s_R|\Lambda) C_{\mathrm{MMI}}(s_1,\ldots,s_R)}{\sum_{s_1,\ldots,s_R} p(X_1,\ldots,X_R,s_1,\ldots,s_R|\Lambda)} \tag{3.7}$$

where

$$C_{\mathrm{MMI}}(s_1,\ldots,s_R) = \prod_{r=1}^{R} \delta(s_r,S_r) \tag{3.8}$$

is a quantity that depends only on the sentence sequence $s_1,\ldots,s_R$. In (3.8), $\delta(s_r, S_r)$ is the Kronecker delta function, that is, $\delta(s_r, S_r) = \begin{cases} 1 \text{ if } s_r = S_r \\ 0 \text{ otherwise} \end{cases}$.

We note that MMI is a discriminative performance measure at the "superstring" level in that it aims to improve the conditional likelihood on the entire training data set instead of on each individual string (token). This is reflected by the product form of the function in (3.8). $C_{\text{MMI}}(s_1, \ldots, s_R)$ in (3.8) can be interpreted as the binary function (as "accuracy count") of the "superstring" $s_1, \ldots, s_R$, which takes a value of 1 if the superstring $s_1, \ldots, s_R$ is correct and zero otherwise. Correspondingly, $O_{\text{MMI}}(\Lambda)$ can be interpreted as the average superstring accuracy count of the full training data set, which takes a continuous value between 0 and 1.

3.4 MCE AND ITS UNIFIED FORM

The key result of this section is to reformulate another popular discriminative criterion, that is, MCE, into the same form of the rational function as in (3.7), except that the accuracy-count function $C(\cdot)$ takes a summation form instead of a product form. This correspondingly gives the string-level discriminative performance measure for MCE, contrasting with the superstring level for the MMI criterion just described. We now first provide a concise introduction to the basic concept and conventional formulation of MCE.

3.4.1 Introduction to the MCE Criterion

MCE learning was originally introduced for multiple-category classification problems where the smoothed error rate is minimized for isolated "tokens" [2, 24]. It was later generalized to minimize the smoothed "sentence token" or string-level error rate [8, 25], which is known as "embedded MCE." The MCE objective function is defined first based on a set of discriminant functions and a special type of loss function. Then model parameters are estimated to minimize the expected loss that is closely related to the recognition error rate of the classifier.

In embedded MCE training, for the rth training token, a set of discriminant functions is defined as the log likelihood of data based on the correct as well as competing strings:

$$g_{s_r}(X_r;\Lambda) = \log p(X_r, s_r|\Lambda) \tag{3.9}$$

Then the decision rule of the classifier/recognizer can be expressed as

$$C(X_r) = \overset{*}{s}_r \text{ iff } \overset{*}{s}_r = \arg\max_{s_r} g_{s_r}(X_r;\Lambda) \tag{3.10}$$

For sequential pattern recognition tasks such as continuous speech recognition, usually only the N most confusable competing strings, $s_{r,1}, \ldots, s_{r,N}$, are considered in MCE. Note these N

confusable competing strings change dynamically after each training iteration. In practice, they are regenerated after each iteration through N-best decoding based on the parameters Λ' obtained at the immediately previous iteration. The N-best strings can be defined inductively by

$$\begin{aligned} s_{r,1} &= \arg\max_{s_r : s_r \neq S_r} \log p(X_r, s_r|\Lambda) \approx \arg\max_{s_r : s_r \neq S_r} \log p(X_r, s_r|\Lambda') \\ s_{r,i} &= \arg\max_{s_r : s_r \neq S_r, s_r \neq s_{r,1}, \ldots, s_{r,i-1}} \log p(X_r, s_r|\Lambda) \approx \arg\max_{s_r : s_r \neq S_r, s_r \neq s_{r,1}, \ldots, s_{r,i-1}} \log p(X_r, s_r|\Lambda') \quad i = 2, \ldots, N. \end{aligned} \tag{3.11}$$

Next, a misclassification measure $d_r(X_r, \Lambda)$ is defined to emulate the decision rule for utterance r, that is, $d_r(X_r, \Lambda) \geq 0$ implies misclassification and $d_r(X_r, \Lambda) < 0$ implies correct classification,

$$d_r(X_r, \Lambda) = -g_{S_r}(X_r; \Lambda) + G_{S_r}(X_r; \Lambda) \tag{3.12}$$

where $G_{S_r}(X_r; \Lambda)$ is a function that represents the score of incorrect strings competing with the correct string S_r.

For 1-best MCE training ($N = 1$), only the best-incorrect-string $s_{r,1}$ is considered as the competitor. In this special case, $G_{S_r}(X_r; \Lambda)$ clearly becomes

$$G_{S_r}(X_r; \Lambda) = g_{s_{r,1}}(X_r; \Lambda) \tag{3.13}$$

However, for the general case where $N > 1$, different definitions of $G_{S_r}(X_r; \Lambda)$ can be used. One popular definition takes the following form [25]:

$$G_{S_r}(X_r; \Lambda) = \log \left\{ \frac{1}{N} \sum_{i=1}^{N} p^{\eta}(X_r, s_{r,i}|\Lambda) \right\}^{\frac{1}{\eta}} \tag{3.14}$$

Another popular form of $g_{S_r}(X_r; \Lambda)$ and $G_{S_r}(X_r; \Lambda)$ (the latter has similar effects to (3.14) and was used in [46]) is

$$\begin{cases} g_{S_r}(X_r; \Lambda) = \log p^{\eta}(X_r, S_r|\Lambda) \\ G_{S_r}(X_r; \Lambda) = \log \sum_{i=1}^{N} p^{\eta}(X_r, s_{r,i}|\Lambda) \end{cases} \tag{3.15}$$

where η is a scaling factor for joint probability $p(X_r, s_r | \Lambda)$. In this paper, we adopt $G_{S_r}(X_r; \Lambda)$ with the form of (3.15) and set $\eta = 1$ for simplicity and mathematic tractability. (We will discuss the $\eta \neq 1$ case in Chapter 6.)

Now we define the MCE loss function, which, for a single utterance r, is typically a sigmoid function as originally proposed in [24, 25]:

$$l_r(d_r(X_r,\Lambda)) = \frac{1}{1+\mathrm{e}^{-\alpha d_r(X_r,\Lambda)}} \tag{3.16}$$

where α is the slope of the sigmoid function, often determined empirically. As presented in [21] (p. 156), we also use $\alpha = 1$ for simplicity in exposition. In practice, however, α is usually set to be a value less than 1; we will discuss this more general case in Chapter 6. Note that the loss function of (3.16) emulates the desirable zero–one classification error count.

Using the misclassification measure in the form of (3.12) and (3.15) (with $\eta = 1$), we can rewrite the loss function for one training string token as

$$l_r(d_r(X_r,\Lambda)) = \frac{\sum\limits_{s_r, s_r \neq S_r} p(X_r, s_r|\Lambda)}{\sum\limits_{s_r, s_r \neq S_r} p(X_r, s_r|\Lambda) + p(X_r, S_r|\Lambda)} = \frac{\sum\limits_{s_r, s_r \neq S_r} p(X_r, s_r|\Lambda)}{\sum\limits_{s_r} p(X_r, s_r|\Lambda)} \tag{3.17}$$

which can be viewed as an model-based expected loss of classifying X_r to S_r, after putting it in the following form:

$$l_r(d_r(X_r,\Lambda)) = \sum_{s_r \neq S_r} P(s_r|X_r,\Lambda) = \sum_{s_r} \underbrace{(1-\delta(s_r,S_r))}_{0-1 \text{ loss}} P(s_r|X_r,\Lambda)$$

Then, because the error count sums over training tokens, the loss function for all R training tokens is naturally defined to be:

$$L_{\mathrm{MCE}}(\Lambda) = \sum_{r=1}^{R} l_r(d_r(X_r,\Lambda)) \tag{3.18}$$

Here, we emphasize the summation in (3.18) for combining all string tokens for MCE. Because each loss function approximates the string error count, the total empirical error count rightfully becomes the sum of all independent individual string error counts. This forms a sharp contrast to the MMI case as in (3.6), where a product of probabilities is constructed in pooling all string tokens.

Now, minimizing the overall loss function of $L_{\mathrm{MCE}}(\Lambda)$ in (3.18) is equivalent to maximizing the following MCE objective function:

$$O_{\mathrm{MCE}}(\Lambda) = R - L_{\mathrm{MCE}}(\Lambda) = \sum_{r=1}^{R} \left[1 - \frac{\sum\limits_{s_r, s_r \neq S_r} p(X_r, s_r|\Lambda)}{\sum\limits_{s_r} p(X_r, s_r|\Lambda)} \right] = \sum_{r=1}^{R} \frac{p(X_r, S_r|\Lambda)}{\sum\limits_{s_r} p(X_r, s_r|\Lambda)} \tag{3.19}$$

3.4.2 Reformulation of the MCE Criterion Into its Unified Form

Unlike the MMI case, the MCE objective function as expressed in (3.19) is a sum of rational functions rather than a rational function in itself, and hence it would not be amenable to the special form of GT-based optimization. The state-of-the-art techniques for optimizing the MCE objective function have been based on gradient descent, which is called generalized probabilistic descent (GPD) [8, 24, 25]. As one original contribution of this paper, we reformulate the MCE objective function as a true rational function in a nontrivial fashion. This not only unifies the earlier disparate types of objective functions and offers insights into their differences and similarities, but, more importantly, it enables the use of GT as an alternative technique to GPD for faster and more effective optimization.

The reformulation proceeds as follows:

$$O_{\mathrm{MCE}}(\Lambda) = \sum_{r=1}^{R} \frac{\sum_{s_r} p(X_r, s_r|\Lambda)\delta(s_r, \mathcal{S}_r)}{\sum_{s} p(X_r, s_r|\Lambda)} \tag{3.20}$$

$$= \underbrace{\frac{\sum_{s_1} p(X_1, s_1|\Lambda)\delta(s_1, \mathcal{S}_1)}{\sum_{s_1} p(X_1, s_1|\Lambda)}}_{:=O_1} + \underbrace{\frac{\sum_{s_2} p(X_2, s_2|\Lambda)\delta(s_2, \mathcal{S}_2)}{\sum_{s_2} p(X_2, s_2|\Lambda)}}_{:=O_2}$$

$$+ \underbrace{\frac{\sum_{s_3} p(X_3, s_3|\Lambda)\delta(s_3, \mathcal{S}_3)}{\sum_{s_3} p(X_3, s_3|\Lambda)}}_{:=O_3} + \cdots + \underbrace{\frac{\sum_{s_R} p(X_R, s_R|\Lambda)\delta(s_R, \mathcal{S}_R)}{\sum_{s_R} p(X_R, s_R|\Lambda)}}_{:=O_R}$$

$$= \frac{\sum_{s_1}\sum_{s_2} p(X_1, s_1|\Lambda)p(X_2, s_2|\Lambda)[\delta(s_1, \mathcal{S}_1) + \delta(s_2, \mathcal{S}_2)]}{\sum_{s_1}\sum_{s_2} p(X_1, s_1|\Lambda)p(X_2, s_2|\Lambda)} + O_3 + \cdots + O_R$$

$$= \frac{\sum_{s_1 s_2} p(X_1, X_2, s_1, s_2|\Lambda)[C_{\mathrm{MCE}}(s_1 s_2)]}{\sum_{s_1 s_2} p(X_1, X_2, s_1, s_2|\Lambda)} + O_3 + \cdots + O_R$$

$$= \frac{\sum_{s_1 s_2 s_3} p(X_1, X_2, X_3, s_1, s_2, s_3|\Lambda)\,[C_{\mathrm{MCE}}(s_1 s_2 s_3)]}{\sum_{s_1 s_2 s_3} p(X_1, X_2, X_3, s_1, s_2, s_3|\Lambda)} + \cdots + O_R \tag{3.21}$$

$$= \frac{\sum_{s_1,\ldots,s_R} p(X_1, \ldots, X_R, s_1, \ldots, s_R|\Lambda)\, C_{\text{MCE}}(s_1, \ldots, s_R)}{\sum_{s_1,\ldots,s_R} p(X_1, \ldots, X_R, s_1, \ldots, s_R|\Lambda)} \tag{3.21}$$

where $C_{\text{MCE}}(s_1, \ldots, s_R) = \sum_{r=1}^{R} \delta(s_1, S_R)$. The final result of (3.21) gives the rational function fitting exactly to the unified form of (3.2). The correctness of (3.21) can also be proved directly by induction, which we leave to readers as an exercise. $C_{\text{MCE}}(s_1, \ldots, s_R)$ in (3.21) can be interpreted as the string accuracy count for $s_1, \ldots, s_R$, which takes an integer value between 0 and R as the number of correct strings in $s_1, \ldots, s_R$. Correspondingly, $O_{\text{MCE}}(\Lambda)$ can be interpreted as the average string accuracy count of the full training data set.

3.5 MINIMUM PHONE/WORD ERROR AND ITS UNIFIED FORM

3.5.1 Introduction to the MPE/MWE Criterion

In this section, we introduce yet another commonly used discriminative training objective function, MPE or MWE, in speech recognition, developed originally in [38, 40]. In contrast to MMI and MCE described earlier, which are aimed at the superstring level and at the string level of recognition performance optimization, respectively, MPE/MWE is aimed for performance optimization at the substring level. In speech recognition, a string corresponds to a sentence, and a substring as a constituent of the sentence can be words or phones. Because performance measures of speech recognition are often the word or phone error rates rather than the sentence error rate, it has been argued that MPE/MWE is a more appropriate criterion to optimize than the MMI and MCE criteria [40].

The MPE objective function that needs to be maximized is defined as

$$O_{\text{MPE}}(\Lambda) = \sum_{r=1}^{R} \frac{\sum_{s_r} p(X_r, s_r|\Lambda) \sum_{s_r} A(s_r, S_r)}{\sum_{s_r} p(X_r, s_r|\Lambda)} \tag{3.22}$$

where $A(s_r, S_r)$ is the raw phone (substring) accuracy count in the sentence string s_r (proposed originally in [38, 40]). Specifically, $A(s_r, S_r)$ is the total phone (substring) count in the reference string S_r minus the sum of insertion, deletion, and substitution errors of s_r computed based on S_r.

The MPE criterion (3.22) equals the model-based expectation of the raw phone accuracy count over the entire training set. This becomes clear by rewriting (3.22) as

$$O_{\text{MPE}}(\Lambda) = \sum_{r=1}^{R} \sum_{s_r} P(s_r|X_r, \Lambda) A(s_r, S_r)$$

where

$$p(s_r|X_r,\Lambda) = \frac{p(X_r,s_r|\Lambda)}{p(X_r|\Lambda)} = \frac{p(X_r,s_r|\Lambda)}{\sum_{s_r} p(X_r,s_r|\Lambda)}$$

is the model-based posterior probability over which the expectation is taken in defining the MPE objective function of (3.22). It can be shown that MPE criterion provides an upper bound of true Bayes risk on the substring (e.g., phone) level.

When $A(s_r, S_r)$ in (3.22) is changed from the raw phone accuracy count to another raw substring accuracy for words $A_l(s_r, S_r)$, we have the virtually equivalent definition of the MWE criterion:

$$O_{\text{MWE}}(\Lambda) = \sum_{r=1}^{R} \frac{\sum_{s_r} p(X_r|s_r,\Lambda)\ P(s_r) A_l(s_r, S_r)}{\sum_{s_r} p(X_r|s_r,\Lambda)\ P(s_r)} \tag{3.23}$$

and hence, throughout this book, we merge these two into the same MPE/MWE category.

3.5.2 Reformulation of the MPE/MWE Criterion Into Its Unified Form

The MPE/MWE objective function is also a sum of multiple rational functions instead of a single rational function, and hence it is difficult to derive GT formulas, as pointed out in [40] (Section 7.2, p. 92). The state-of-the-art techniques for optimizing the MPE/MWE objective functions have been based on a weak-sense auxiliary function (WSAF) proposed in [40], where the difficulty of formulating a rational function and the desire of moving away from traditional gradient descent have been eloquently discussed. In this paper, we propose to reformulate the MPE/MWE objective functions as a unified rational function in the form of (3.2). This enables an alternative technique to WSAF for optimization but with guaranteed convergence in the algorithm's iteration. The reformulation of the MPE/MWE criteria (3.22)–(3.23) in the unified form of rational function follows the same steps as in the preceding MCE case. Note that (3.22)–(3.23) are in the same form as (3.20), except for the replacement of $\delta(s_r, S_r)$ by $A(s_r, S_r)$ or $A_1(s_r, S_r)$. Then, the same steps starting from (3.20) to (3.21) lead to the reformulated results for MPE/MWE:

$$O_{\text{MPE}}(\Lambda) = \frac{\sum_{s_1,\ldots,s_R} p(X_1,\ldots,X_R,s_1,\ldots,s_R|\Lambda) C_{\text{MPE}}(s_1,\ldots,s_R)}{\sum_{s_1,\ldots,s_R} p(X_1,\ldots,X_R,s_1,\ldots,s_R|\Lambda)} \tag{3.24}$$

where

$$C_{\text{MPE}}(s_1,\ldots,s_R) = \sum_{r=1}^{R} A(s_r, S_r)$$

and

$$O_{\text{MWE}}(\Lambda) = \frac{\sum_{s_1,\ldots,s_R} p(X_1,\ldots,X_R,s_1,\ldots,s_R|\Lambda) C_{\text{MWE}}(s_1,\ldots,s_R)}{\sum_{s_1,\ldots,s_R} p(X_1,\ldots,X_R,s_1,\ldots,s_R|\Lambda)} \tag{3.25}$$

where

$$C_{\text{MWE}}(s_1,\ldots,s_R) = \sum_{r=1}^{R} A_l(s_r,S_r)$$

Note that $C_{\text{MPE}}(s_1, \ldots, s_R)$ in (3.24) or $C_{\text{MWE}}(s_1, \ldots, s_R)$ in (3.25) can be interpreted as the raw phone or word (substring unit) accuracy count within the "superstring" $s_1, \ldots, s_R$. Its upper-limit value is the total number of phones or words in the full training data (i.e., the correct superstring $S_1, \ldots, S_R$). The actual value may be negative if many insertion errors occur. Correspondingly, $O_{\text{MPE}}(\Lambda)$ and $O_{\text{MWE}}(\Lambda)$ can be interpreted as the average raw phone or word accuracy count of the full training data set.

3.6 DISCUSSIONS AND COMPARISONS

3.6.1 Discussion and Elaboration on the Unified Form

We first provide a summary of the previous sections in this chapter, where a rational-function form of the discriminative training (DT) objective function or criterion is established as in (3.2) that unifies MMI, MCE, and MPE/MWE. In this unified form, the choice of the set of label sequences and the form of the generic function $C_{\text{DT}}(s_1, \ldots, s_R)$ determine the particular DT criterion, as summarized in Table 3.1. As an example shown in Table 3.1, for MMI, we have the specific function $C_{\text{DT}}(s_1, \ldots, s_R) = \prod_{r=1}^{R} \delta(s_r, S_R)$. For MPE, the function becomes $C_{\text{DT}}(s_1, \ldots, s_R) = \sum_{r=1}^{R} A(s_r, S_r)$. For MCE with general N-best competitors where $N > 1$, $C_{\text{DT}}(s_1, \ldots, s_R) = \sum_{r=1}^{R} \delta(s_r, S_r)$. For 1-best MCE ($N = 1$), s_r belongs to only the subset $\{S_r, s_{r,1}\}$. Equation (3.2) allows direct comparisons of the MMI, MCE, and MPE/MWE criteria. The most important insight offered by the unified framework of (3.2) is that the difference of these three types of criteria is embedded only in the weighting of alternative strings, where the weights (i.e., $C_{\text{DT}}(s_1, \ldots, s_R)$) are independent of the model parameters Λ to be learned.

TABLE 3.1: A unified rational-function form of the DT objective function (3.2), where differences in CDT(s_1, …, s_R) distinguish MMI, MCE, and MPE/MWE and the number of "competing token candidates" distinguishes N-best and 1-best versions of the MCE

OBJECTIVE FUNCTIONS	$C_{\mathrm{DT}}(S_R)$	$C_{\mathrm{DT}}(S_1, \ldots, S_R)$	LABEL SEQUENCE SET USED IN DT
MCE (*N*-best)	$\delta(s_r, S_r)$	$\sum_{r=1}^{R} C_{\mathrm{DT}}(s_r)$	$\{S_r, s_{r,1}, \ldots, s_{r,N}\}$
MCE (1-best)	$\delta(s_r, S_r)$	$\sum_{r=1}^{R} C_{\mathrm{DT}}(s_r)$	$\{S_r, s_{r,1}\}$
MPE	$A(s_r, S_r)$	$\sum_{r=1}^{R} C_{\mathrm{DT}}(s_r)$	all possible label sequences
MWE	$A_l(s_r, S_r)$	$\sum_{r=1}^{R} C_{\mathrm{DT}}(s_r)$	all possible label sequences
MMI	$\delta(s_r, S_r)$	$\prod_{r=1}^{R} C_{\mathrm{DT}}(s_r)$	all possible label sequences

Note that the overall $C_{\mathrm{DT}}(s_1, \ldots, s_R)$ is constructed from its constituents $C_{\mathrm{DT}}(s_r)$'s in individual string tokens by either summation (for MCE, MPE/MWE) or product (for MMI).

As pointed out in [40], MPE/MWE has an important difference from MCE and MMI in that the weighting given by the MPE/MWE criteria to an incorrect string (sentence token) depends on the number of wrong substrings (wrong phones or words) within the string. MCE and MMI make a binary distinction based on whether the entire sentence string is correct or not, which is not desirable when reduction of substring errors (e.g., word errors in speech recognition) is the main goal of the sequential pattern recognition tasks. This distinction is most clearly seen by the sum of the binary function $C_{\mathrm{DT}}(s_1, \ldots, s_R) = \sum_{r=1}^{R} \delta(s_r, S_r)$ for MCE and the sum of nonbinary functions $C_{\mathrm{DT}}(s_1, \ldots, s_R) = \sum_{r=1}^{R} A(s_r, S_r)$for MPE/MWE, both within the same unified framework. This key difference gives rise to the distinction of the substring level versus the string level of recognition performance optimization associated with MPE/MWE and MCE, respectively. As it performs the sentence or string-level optimization, MCE tends to push and pack errors into a few sentence

tokens so as to create as many error-free token "strings" as possible. It sacrifices word/phone (substring) errors in order to reduce string errors, which may not be desirable when high word or phone accuracy is the goal of continuous speech or phonetic recognition.

Furthermore, the product instead of the summation form of the binary function associated with MMI, that is, $C_{\mathrm{DT}}(s_1, \ldots, s_R) = \prod_{r=1}^{R} \delta(s_r, S_R)$, makes it clear that MMI achieves performance optimization at the superstring level. That is, as long as any one single sentence token has an error, the product of the Kronecker delta functions becomes zero. Therefore, all the summation terms in the numerator of (3.2) are zero except for the one corresponding to the correct label/transcription sequence. This "superstring" level performance measure is apparently less desirable than MCE or MPE/MWE, as has been shown extensively in experiments [31, 38–40].

Another insight gleaned from the unified form of the objective function (3.2) is that in the case of having only one sentence token (i.e., $R = 1$) in the training data and when the sentence contains only one phone, then all three MMI, MCE, and MPE/MWE criteria become identical. This is because in this case $C_{\mathrm{DT}}(s_1, \ldots, s_R)$ becomes identical for all these criteria. The difference surfaces only when the training set consists of multiple sentence tokens. In this realistic case, the difference lies only in $C_{\mathrm{DT}}(s_1, \ldots, s_R)$ as the Λ-independent weighing factor (as well as in the set of competitor strings), whereas the general rational-function form for the three criteria remains unchanged.

The major benefit of unifying the discriminative training criteria into a single rational function as in (3.2) is that we can then extend the same, well-established framework of GT to optimize all these major discriminative training criteria. Moreover, it also provides a new possibility of applying other, more advanced rational function optimization methods to the various discriminative training criteria. As an example, Jebara [22, 23] proposed a novel optimization method for rational functions as an alternative to the traditional GT method. In [22], the reverse Jensen's inequality method was developed and described, based on which an elegant solution for rational function optimization for HMMs with exponential-family densities was constructed. We will review this method in Appendix V, showing that our unified framework of (3.2) is directly subject to the application of this method and that this is not the case for another framework that we review below.

3.6.2 Comparisons With Another Unifying Framework

In recent papers [31, 46], an approach was proposed to unify a number of discriminative learning methods including MMI, MPE, and MPE/MWE (the earlier paper [46] did not include MPE/MWE). Functional similarities and differences among MMI, MCE, and MPE/MWE criteria were noted and discussed in [31, 46]. The framework proposed in this paper takes an additional step of unifying these criteria in a common rational-function form, and GT-based discriminative

TABLE 3.2: Choices of the smoothing function f (z), alternative word sequences M_r, and exponent weight h in (3.26) for various types of DT criteria

CRITERIA	SMOOTHING FUNCTION $f(z)$	ALTERNATIVE WORD SEQUENCES M_r	H
MCE (N-best)	$-1/[1+\exp(2qz)]$	$\{s_r\}$ excluding S_r	≥1
MCE (1-best)	$-1/[1+\exp(2qz)]$	$\{s_{r,1}\}$	N/A
MPE/MWE	$\exp(z)$	all possible label sequence $\{s_r\}$	1
MMI	z	all possible label sequence $\{s_r\}$	1

This is modified from the original table in [46].

learning is applied to this generic rational-function, which includes MMI, MCE, and MPE/MWE criteria as special cases. This is significant from two perspectives. First, it provides a more precise and direct insight into the fundamental relations among MMI, MCE, and MPE/MWE criteria at the objective function level based on the common rational-function form. Second, it enables a unified GT-based parameter optimization framework that applies to MMI, MCE, MPE/MWE, and other discriminative criteria.

In [31], it was based on the objective function of the following form (rewritten using the mathematical notations of this paper for easy comparison):

$$O(\Lambda) = \frac{1}{R}\sum_{r=1}^{R} f\left(\frac{1}{\eta}\log\frac{\sum_{s_r} p^{\eta}(X_r, s_r|\Lambda)\, C_{\mathrm{DT}}(s_r, S_r)}{\sum_{s_r \in M_r} p^{\eta}(X_r, s_r|\Lambda)}\right) \tag{3.26}$$

where $C_{\mathrm{DT}}(s_r)$ takes the same value as in our Table 3.1. The choices of the smoothing function $f(z)$, the competing word sequences M_r, and the weight value η in (3.26) are provided in Table 3.2 for the different types of DT criteria. In Table 3.2, q is the slope of a sigmoid smoothing function.

Equation (3.26) is a generic description of the objective functions from MMI, MCE, and MPE/MWE. However, it is not at the definitive level of a unified form of a rational function. It indicates that different discriminative criteria can have a similar form of kernel and differ by the criterion dependent smoothing function $f(z)$ that modulates the kernel. The objection function of (3.26) is a sum of the smoothing functions. In the approach presented in this chapter, we show that

objective functions from MMI, MCE, and MPE/MWE criteria can have a definitive common rational-function form (3.2), and for each discriminative criterion, the objective function differs only by a model-independent quantity $C_{\mathrm{DT}}(s_1, \ldots, s_R)$.

On the other hand, as shown in Table 3.2, $f(z)$ is a nonlinear function for MPE/MWE and MCE criteria. Therefore, the original GT solution [14], while directly applicable to MMI with $f(z)$ being an identity function and z being a logarithm of rational function, is not directly applicable to the objective functions of MPE/MWE and MCE criteria. To circumvent this difficulty, the limiting procedures of [26, 27] are needed, in which the original objective function is approximated by a sequence of polynomials through Taylor series expansion (in a neighbor of the current model parameters). Based on that, the GT-based parameter optimization of [14] can be applied to each of the partial sum, a polynomial with finite number of terms. But for a nonpolynomial analytic function, the Taylor series expansion consists of infinite number of terms. It needs to justify the limiting process that the GT for polynomials with finite number of terms can be extended to the limit case as the number of terms goes to infinity, for example, the existence of a uniform bounded constant D for all partial sums of the Taylor series expansion in GT-based parameter optimization. The unified rational-function approach described in this paper departs from the work of Macherey et al. [31] and Schlüter et al. [46], because it is free from the Taylor series expansion and it maps the objective functions from MMI, MCE, and MPE/MWE criteria into a definitive rational-functional form (3.2). Therefore, the GT-based parameter optimization framework of [14] can be directly applied. Moreover, this approach allows new rational function optimization methods (e.g., the method based on reverse Jensen's inequality [22]) to be applied, upon which algorithmic properties of the parameter optimization procedure can be constructively established and justified.