# NON-UNIFORM ERROR CRITERIA FOR AUTOMATIC PATTERN AND SPEECH RECOGNITION

*Qiang Fu, Dwi Sianto Mansjur, Biing-Hwang Juang*

Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332
{qfu, dwi, juang}@ece.gatech.edu

## ABSTRACT

The classical Bayes decision theory [1] is the foundation of statistical pattern recognition. Conventional applications of the Bayes decision theory result in ubiquitous use of the maximum *a posteriori* probability (MAP) decision policy and the paradigm of distribution estimation as practice in the design of a statistical pattern recognition system. In this paper, we address the issue of non-uniform error criteria in statistical pattern recognition, and generalize the Bayes decision theory for pattern recognition tasks where errors over different classes have different degrees of significance. We further propose extensions of the method of minimum classification error (MCE) [2] for a practical design of a statistical pattern recognition system to achieve empirical optimality when non-uniform error criteria are prescribed. In addition, we apply our method upon speech recognition tasks. In the context of automatic speech recognition (ASR), we present a variety of training scenarios and weighting strategies under our framework. The experimental demonstrations for both general pattern recognition and continuous speech recognition are provided to support the effectiveness of our new approach.

***Index Terms***— Non-Uniform error cost, Weighted MCE training

## 1. INTRODUCTION

The classical Bayes decision theory [1] is the foundation of statistical approach to the problem of pattern recognition. Bayes' analysis of the pattern recognition problem is built upon the notion of an *expected* system performance, as opposed to the evaluation of any particular instances of recognition decisions. Consider a pattern recognition task involving $M$ classes of events or patterns (e.g., the task of recognizing a handwritten digit with $M = 10$). An unknown pattern, say $X$, is observed and recognized as belonging to one of the $M$ classes. Thus, a recognizer is a function $C$ that maps $X$ to a class identity denoted by $C_i$, where $i \in I_M = \{i, i = 1, 2, \ldots, M\}$. We denote this function as a decision function $C(X)$. Obviously, some decisions are likely to be correct while others wrong, and correct decisions are preferred over wrong decisions. In other words, every decision is associated with a cost which can be expressed as an entry $\epsilon_{ij}$ in an $M \times M$ matrix where $i, j \in I_M$, signifying the cost in identifying a pattern from the $j^{th}$ class as one of the $i^{th}$ class. Suppose at our disposal we have the knowledge of the *a posteriori* probabilities $P(C_i|X), \forall i \in I_M$. Then, following the teaching of Bayes, given $X$, the conditional cost of making a decision of $C(X) = C_i$ can be defined as [1]

$$R(C_i|X) = \sum_{j=1}^{M} \epsilon_{ij} P(C_j|X) \tag{1}$$

and the system performance in terms of the *expected* loss is

$$\mathcal{L} = E\{R(C(X)|X)\} = \int R(C(X)|X)p(X)dX \tag{2}$$

Traditionally. a simple error count is used as the cost of recognition decision with

$$\epsilon_{ij} = \begin{cases} 1, & i \neq j \\ 0, & i = j \end{cases} \tag{3}$$

which is a typical error cost function. This cost function is the most intuitive and prevalent performance measure in pattern recognition as it is related to the probability of error in simple terms. We can institute the decision policy with the cost function of (3) as

$$C(X) = \arg\min_i R(C_i|X) = \arg\max_i P(C_i|X) \tag{4}$$

The expected loss of (2) will be minimized due to the fact that $p(X)$ is non-negative. This is the ubiquitous maximum *a posteriori* (MAP) decision rule that guarantees minimum system cost, or Bayes risk [1].

Note that the above Bayes decision theory requires that the *a posteriori* distribution needs to be available to the system. The result of (4) has led to the conventional paradigm of distribution estimation as a fundamental step towards the design of a pattern recognition system. However, due to lack of knowledge of the analytical functional form of the *a posteriori* distribution, the optimality of the distribution estimation method is usually hindered. Further more, when the cost function is not uniform, the best decision policy is not necessarily the one that achieves maximum *a posteriori* probability. Instead, we may want an "optimal" decision policy which can accomplish the *minimum error cost/risk.* The non-uniform or asymmetric error cost function is quite common in real world applications. For example, a keyword spotting system may consider misrecognizing "key" words unacceptable, while errors of functional words such as "a" or "the" may not be considered consequential. It is necessary to revisit the Bayes decision theory and discuss the validity of the conventional MAP policy when non-uniform error criteria are employed. The purpose of this paper is therefore to reformulate a framework for such applications in pattern recognition. In addition, we attempt to provide what may be considered a reasonable system design methodology to follow the circumstances that the real data distribution is unknown.

1-4244-1484-9/08/$25.00 ©2008 IEEE

In this paper, in addition to Fgeneral pattern recognition experiments, we apply our framework to ASR tasks as well. Speech recognition is a typical task that the form of data distribution function (or probability density function) is usually unavailable, leaving no assurance of any optimality if using distribution estimation methods even with infinite amount of data. Further more, the validity of non-uniform error cost can be justified in many speech recognition tasks. We present corresponding non-uniform error training criteria according to different training scenarios in the context of ASR applications.

This paper is organized as follows. In Section 2, we discuss Bayes' original analysis of optimal decision as applied to general cases involving non-uniform error cost. Also, we point out that the method of MCE, which departs from the conventional paradigm of distribution estimation for pattern recognition, provides a fitting framework for incorporation of non-uniform error cost functions. In Section 3, we apply our framework to the ASR applications and discuss some training scenarios. We demonstrate the results of our new methodology through several general pattern recognition and speech recognition tasks in Section 4. Conclusions and future work are provided in Section 5.

## 2. DECISION POLICY WITH NON-UNIFORM ERROR COST AND WEIGHTED MCE METHOD

### 2.1. Minimum Risk (MR) rule

The conditional risk of (1) and the expected loss of (2) are general expressions of the system performance without imposing any particular conditions on the error cost function $\epsilon_{ij}$. Again, since $p(X)$ is non-negative,

$$\min \mathcal{L} = \min_C E\{R(C(X)|X)\} = \int \min_C R(C(X)|X)p(X)dX \tag{5}$$

To achieve the minimum risk, the recognizer function must implement the following policy,

$$C(X) = \arg\min_{C_i} R(C_i|X) = \arg\min_{C_i} \sum_{j=1}^{M} \epsilon_{ij} P(C_j|X) \tag{6}$$

We call this the *minimum risk* (MR) rule. For a non-uniform error cost function, we generally require that $\epsilon_{ij} = 0$ for $i = j$ and $\epsilon_{ij} \geq 0$ for $i \neq j$.

### 2.2. Weighted MCE with Non-Uniform Error Cost

The incorporation of a class-dependent non-uniform error cost function incurs two factors that require careful consideration. First, the system needs to implement the MR rule defined in (6). We need to embed this *decision rule* (or decision operation) in a functional form so that optimization can be performed to obtain the values of the system parameter set. Second, as the overall system performance is defined over a non-uniform error cost function, the particular decision for each of the training token becomes an integral part of the performance measure and has to be included in the objective function for optimization. The second factor is unique because once a decision is rendered by the recognizer, what matters is not only if the decision is right or wrong but how much error cost the decision actually incurs. We shall see how these factors are taken into account in the proposed schemes for non-uniform error cost minimization.

Note that the execution of (6) obviously requires the knowledge of the *a posteriori* probability $P(C_j|X), \forall j \in I_M$. In real applications, the *a posteriori* distribution needs to be learned with class identity labels, as part of the conventional design paradigm for a recognition system. The estimated posterior distribution (for all classes and over the entire space of $X$) may have to be substantially more accurate in the non-uniform cost case than in the uniform case, because one may argue that the rank order of posterior probabilities (as required in uniform cost situations) is likely to be less sensitive to small deviations than the quantities themselves (which is required in the non-uniform cost). Therefore, to implement the MR decision rule in practice, we introduce the idea of the *discriminant function* as follows.

Let $g_i(X; \Lambda) \geq 0$ be a discriminant function for the $i^{th}$ class, $i = 1, 2, \ldots, M$ where $\Lambda$ is the parameter set that defines the function. The recognition decision is reached according to

$$C(X) = \arg\max_i g_i(X; \Lambda) \tag{7}$$

That is, the recognizer chooses the class that leads to the largest value among all discriminants evaluated on $X$. Obviously, If the true *a posteriori* probability is available, a monotonically decreasing function of the conditional risk of (1) (to switch $\min$ into $\max$ operation) would be appropriate. For example,

$$g_i(X; \Lambda) = \exp\{-R(C_i|X)\} = \exp\left\{-\sum_{j \in I_M} \epsilon_{ij} P(C_j|X)\right\} \tag{8}$$

However, when the above approximation of the *a posteriori* distribution can not be ensured, one may opt for other discriminant functions based on some reasonable convention. One example is the use of hidden Markov models (HMM) as the discriminant functions in ASR applications.

To accumulate the error cost of each training token into the objective function, the expected system loss of (2) needs to be expressed in terms of the empirical loss (yet to be defined) with the decision rule embedded in it. For clarity, let $i_X = C(X)$ be the identity index as decided by the recognizer and $j_X$ be the true identity index of X. Also, $\Omega = \{X^{(n)}\}_{n=1}^N$ is the set of training tokens. A *single token* realized cost is defined as

$$l_{i_X}(X; \Lambda) = \epsilon_{i_X j_X} \tag{9}$$

Therefore if the empirical system loss is defined over the realized token-based costs (rather than the expected cost on realized tokens), an alternative non-uniform cost will result:

$$L = \frac{1}{N} \sum_{X \in \Omega} \epsilon_{i_X j_X} \rightarrow \int \epsilon_{i_X j_X} p(X) dX \tag{10}$$

Suppose that each class is prescribed a discriminant function $g_j(X; \Lambda), \forall j$. Define the recognizer function as

$$C(X) = i_X = \arg\max_i g_i(X; \Lambda) \tag{11}$$

The empirical system loss of (10) based on $\Omega$ is then

$$L = \frac{1}{N} \sum_{X \in \Omega} \sum_{i \in I_M} \sum_{j \in I_M} \epsilon_{ij} \mathbf{1}[j_X = j] \mathbf{1}\{i = \arg\max_k g_k(X; \Lambda)\} \tag{12}$$

Note that in the above the indicator function $\mathbf{1}[j_X = j] = \mathbf{1}[X \in C_j]$.

The remaining challenge in designing a discriminative training algorithm with non-uniform error cost is to turn the objective function, $L$ in (12), into an appropriate smooth function of the parameter so as to allow numeric optimization. Consider

$$L = \sum_{j \in I_M} L_j \tag{13}$$

and

$$L_j = \frac{1}{N} \sum_{X \in \Omega} \left( \sum_{i \in I_M} \epsilon_{ij} \mathbf{1} \left\{ i = \arg\max_k g_k(X; \Lambda) \right\} \right) \mathbf{1}[X \in C_j] \tag{14}$$

That is, $L_j$ is the empirical error cost collected over all training tokens in $\Omega$ with $j_X = j$. The approximation then needs to be made to the summands. This can be accomplished by

$$\sum_{i \in I_M} \epsilon_{ij} \mathbf{1}\{i = \arg\max_k g_k(X; \Lambda)\} \approx \sum_{i \in I_M} \epsilon_{ij} \frac{g_i(X; \Lambda)}{G(X; \Lambda)} \tag{15}$$

where

$$G(X; \Lambda) = \left[ \sum_{i \in I_M} g_i^{\eta}(X; \Lambda) \right]^{1/\eta} \tag{16}$$

Note that as $\eta \to \infty$,

$$\frac{g_i(X; \Lambda)}{G(X; \Lambda)} \approx \left\{ \begin{array}{ll} 1, & G(X; \Lambda) = \max_k g_k(X; \Lambda) \\ 0, & \text{otherwise} \end{array} \right. \tag{17}$$

Finally, the smoothed empirical system cost is

$$L \approx \frac{1}{N} \sum_{X \in \Omega} \sum_{j \in I_M} \left( \sum_{i \in I_M} \epsilon_{ij} \frac{g_i(X; \Lambda)}{G(X; \Lambda)} \right) \mathbf{1}[X \in C_j] \tag{18}$$

which is a continuous function of the parameter set $\Lambda$. We name Eq. (18) as the objective function of the *weighted MCE* method. The parameter set $\Lambda$ can be updated using gradient descent methods (e.g., generalized probabilistic descent method (GPD)) [2][3]. We have derived updating equations for Gaussian mixture models (GMM) in [4].

## 3. NON-UNIFORM ERROR CRITERIA FOR SPEECH RECOGNITION

Speech recognition is an important category of pattern recognition applications. In brief, there are two training scenarios in ASR. In the first case, the training and recognition decisions are on the same linguistic level of the performance measure. For example, the acoustic model is trained on the *phone* level and the evaluation metric is the weighted *phone* error rate (PER). In this case, the loss of wrong recognition decisions represents the recognizer's performance directly. We call this scenario **intra-level training**. The second and the most common circumstance in practice is **inter-level** training in which the training and recognition decisions are not on the same linguistic level as the performance measure. For example, the training and the recognition are on the *phone* level but the system evaluation measure is the weighted *word* error rate (WER). In this case, the system performance is not evaluated by the recognition error loss. Hence, minimizing the cost of wrong recognition decisions does not directly optimize the recognizer's performance in terms of the evaluation measure. To alleviate this inconsistency, the error weighting strategy should be built in a cross-level fashion.

### 3.1. Error Weighting for Intra-Level Training

Assume that the training is on the *phone* level and the evaluation measure is the weighted *phone* error rate (PER). The phone sequence $PH = (ph_1, ph_2, \ldots, ph_{L_k})$ is the label of the $k$th training token in a training set with totally $K$ tokens. $X_k = \{X_{k,l_k}\}_{l_k}^{L_k}$ is the $k$th token that is segmented into $L_k$ segments corresponding to the phone sequence. The objective function for the weighted MCE in this case could be written as

$$\mathcal{L}'_{W-MCE} = \sum_k \sum_{l_k=1}^{L_k} \epsilon_{ij} \frac{g_i(X; \Lambda)}{G(X; \Lambda)} \mathbf{1}(X_{k,l_k} \in C_j) \tag{19}$$

### 3.2. Error Weighting for Inter-Level Training

Assume that in this case, the training is on the phone model and the performance metric is the weighted word error rate. The first weighting mechanism we are discussing is the user-defined weighting. Let the word sequence $W = (w_1, w_2, \ldots, w_{L_k})$ be the label of the $k$th training token in a training set with totally $K$ tokens. Each word $w_{l_k}$ contains a phone sequence as $ph_{l_k}^1, ph_{l_k}^2, \ldots, ph_{l_k}^{N_k}$. $X_k = \{X_{k,l_k,n_k}\}_{l_k}^{L_k}$ is the $k$th token that is segmented into $L_k$ segments corresponding to the word sequence. Hence, the weighted MCE for the inter-level training can be written as:

$$\mathcal{L}''_{W-MCE} = \sum_k \sum_{l_k=1}^{L_k} \sum_{n_k=1}^{N_k} \epsilon_{ij} \frac{g_i(X; \Lambda)}{G(X; \Lambda)} \mathbf{1}(X_{k,l_k,n_k} \in C_j) \mathcal{E}(w_{l_k}) \tag{20}$$

where $\mathcal{E}(w_{l_k})$ is the word-level (or higher level) cost imposed on word $w_{l_k}$. This objective function utilizes the class-dependent error cost from the higher level (e.g., word level) to control the optimization of the parameters on the lower level (e.g., phone level).

## 4. EXPERIMENTS

We present two groups of experiments to show the effectiveness of our framework. We first conduct computer simulations to demonstrate the effectiveness of our method in the ideal situation, in which we know the analytical form of the data distribution so that we can model the scoring function $g_i(X; \Lambda)$ as a function of the non-uniform error cost as of Eq. (8). We then present speech recognition experiments, in which real data distribution is unknown and the scoring function $g_i(X; \Lambda)$ is assumed to be an HMM model. Because of the limited space, we only display the inter-level training experiments.

### 4.1. Experiments on General Pattern Recognition

In our experiments, there are 3 classes and the data of each class is 2-dimensional and generated by a GMM with 4 mixture components. We generate 9 data sets for each of the 3 classes using 4-mixture GMMs, which correspond to the data set size from 128 (which is $2^7$) to 32768 (which is $2^{15}$). Fig.1 shows the GMM model contours with 1024 data samples. The entire data set into the training and test set by a ratio of $80/20$ and the class prior probabilities are assumed to be 0.2, 0.3, and 0.5 respectively. The baseline models are initialized using the EM algorithm [1]. The scoring function $g_i(X; \Lambda)$ is modeled as of Eq.(8) and the detailed parameter updating equations are derived in [4].

Fig.2 compares the performance in terms of the empirical error cost of (13) computed using a non-uniform cost function between the baseline and the MCE-trained models. The performance comparison

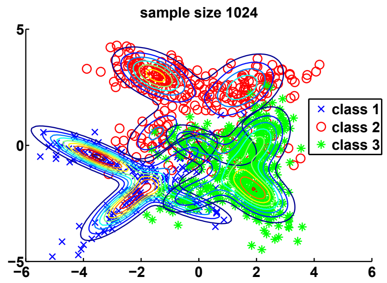
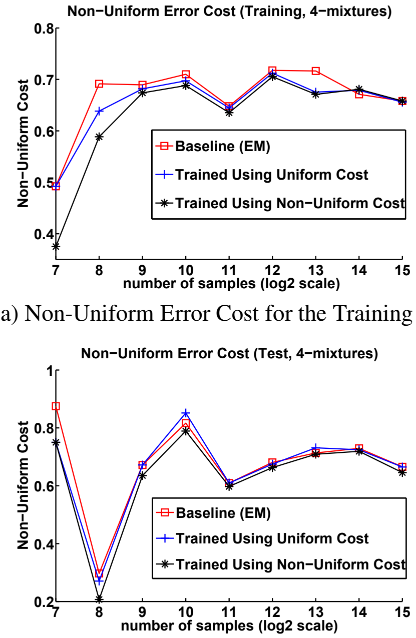


**Fig. 1**. GMM models contours with 1024 data samples.

assessed on the training set is plotted in the upper panel and that on the test set is in the bottom panel. In each panel, the horizontal axis represents the number of training data samples. The square sign, the "+" sign and the star sign denote the baseline, the model trained by the conventional MCE method with a *uniform* cost function, and the model trained by the weighted MCE method with a *non-uniform* cost function, respectively. The non-uniform error cost function is assumed to be

$$[\epsilon_{ij}] = \begin{bmatrix} 0 & 7 & 3 \\ 2 & 0 & 8 \\ 4 & 6 & 0 \end{bmatrix}$$

We can observe that since the performance measure is the non-uniform error cost, the model trained by the MCE method with the non-uniform cost matrix (i.e., a matched-objective condition) shows the best performance in most operating points due to the consistency between the training objective and the performance measure.



(a) Non-Uniform Error Cost for the Training Set

(b) Non-Uniform Error Cost for the Test Set

**Fig. 2**. Non-uniform error cost for training and test set when modeling $g(X;\Lambda)$ as a function of the non-uniform error cost

### 4.2. Experiments on Speech Recognition

The weighted MCE method in the inter-level training scenario is investigated on the TIDIGITS database [5]. The main challenge in this scenario is to adjust the cost of phone errors so that the word error rate is optimized. The baseline is built based on 22 HMM models for all phonemes in digit words 0-9 plus "oh" using EM algorithm. Each model has 3 states and each state contains 32 Gaussian mixtures. The training features are 12MFCCs + $\Delta$ + $\Delta^2$ + energy. We use a non-uniform error cost function that is created through a transformation of the phoneme confusion matrix.

In Table 1, we compare the performance of the MCE method using uniform and non-uniform error cost matrices for inter-level training. We can see that for both the word error rate and sentence error rate, the weighted MCE method achieves better performance.

**Table 1**. Performance comparison between the conventional MCE training and the weighted MCE method with the non-uniform error cost matrix

| | Word Error Rate | Sentence Error Rate |
|---|---|---|
| Baseline | 1.53 | 4.64 |
| MCE | 1.40 | 4.36 |
| Weighted MCE | 1.32 | 3.97 |

## 5. CONCLUSIONS AND FUTURE WORK

In this paper, we start from the Bayes decision theory and derive a framework to solve the problem of the optimal classifier design with non-uniform error cost assessments. We apply our method on both general pattern recognition problems and speech recognition tasks. Experiments show that our method is effective in minimizing non-uniform error cost according to system requirements.

In the future, more details with regard to the non-uniform error criteria will be discussed upon more complex tasks. Techniques of building efficient non-uniform error cost functions also need further exploration.

## 6. REFERENCES

[1] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, John Wiley, New York, NY, 2001, 2nd edition.

[2] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 257–265, May 1997.

[3] E. McDermott and T. J. Hazen, "Minimum classification error training of landmark models for real-time continuous speech recognition," in *ICASSP-2004*, Montreal, Canada, May 2004, pp. 937–940.

[4] Q. Fu, D. Mansjur, and B.-H. Juang, "Pattern recognition with non-uniform error criteria," *submited to IEEE Transaction on Pattern Analysis and Machine Intelligence*.

[5] R. G. Leonard, "A database for speaker-independent digit recognition," in *ICASSP-1984*, 1984, p. 42.11.