Topics on Minimum Classification Error Rate Based Discriminant Function Approach to Speech Recognition

Wu Chou
Bell Labs., Lucent Technologies, 600 Mountain Ave., Murray Hill, NJ 07974, U.S.A., wuchou@research.bell-labs.com

ABSTRACT
In this paper, we study discriminant function based minimum recognition error rate pattern recognition approach. This approach departs from the conventional paradigm which links a classification/recognition task to the problem of distribution estimation. Instead, it takes a discriminant function based statistical pattern recognition approach and the goodness of this approach to classification error rate minimization is established through a special loss function. It is meaningful even when the model correctness assumption is known not valid. The use of discriminant function has a significant impact on classifier design, since in many realistic applications, such as speech recognition, the true distribution form of the source is rarely known precisely and without model correctness assumption, the classical optimality theory of the distribution estimation approach can not be applied directly. We discuss issues in this new classifier design paradigm and present various extensions of this approach for applications in speech processing.

1. INTRODUCTION
The advent of powerful computing devices and the success of hidden Markov modeling triggered a renewed pursuit for more powerful statistical methods to further reduce the recognition error rate and build more robust speech recognition systems across various conditions. The use of discriminant function methods in speech recognition is a promising approach. Although the statistical formulation of minimum classification error (MCE) based discriminative methods has its root in the classical Bayes decision theory, it departs from the conventional paradigm which links a classification/recognition task to the problem of distribution estimation. Instead, it takes a discriminant function based statistical pattern classification approach, and for a given family of discriminant function, optimal classifier design involves finding a set of parameters which minimize the empirical recognition error rate. One classical example of using discriminant function for classifier design in statistical literatures is the two class classification problem using linear discriminant functions. In particular, a window based method was described in [22] for the two class classification problem using linear discriminant functions that minimize the probability of classification error. In this paper, the focus is on general MCE based discriminative methods. The discriminant functions are nonlinear and often related to the structure of the statistical framework used in speech recognition such as hidden Markov models.

The reason of taking a discriminant function based approach to classifier design, as will be further elaborated, is due mainly to the fact that we lack complete knowledge of the form of the data distribution and that training data are always inadequate, particularly in dealing with speech problems. The performance of a recognizer is normally defined by its expected recognition error rate, and an optimal recognizer should be the one that achieves the least expected rate of recognition error. The difference between the distribution estimation based approach and the discriminant function based MCE approach lies in the way the recognition error is expressed and in the computational steps that would lead to the minimization of such error functions. A key to the development of the MCE method is a new error function which incorporates the recognition operation and performance in a functional form, from which the performance of the classifier can be directly evaluated and optimized.

Classifier design without assuming the knowledge of class posterior probabilities has been studied in many areas. In particular, Tsypkin [72] and Amari [1, 2] pioneered this approach for self-learning and self-organizing nets. They formulated the problem of self-learning into a classification problem which consists of optimal partitioning of the observation space into regions $X_k$ for which the expected risk $R$ is minimized. In addition, a mathematical minimization procedure, generalized probabilistic descent (GPD) algorithm or stochastic approximation, was proposed as a means for classifier design under this framework. Since then, various loss functions have been used in designing classifiers, including those popular mean-square error based loss functions. However, many tractable loss functions do not have a direct relation to the recognition error rate minimization, and therefore, albeit based on discriminant functions, they are not directly related to recognition error rate which should be the most sensible choice for classifier design.

Over the past decades, the MCE based approach has been developed to overcome the fundamental limitations of the traditional approach and to directly link the classifier design problem to classification error rate minimization. In order to alleviate the dependency on the class posterior distributions, a discriminant function based MCE approach was proposed by Juang et al. [32] as an alternative to optimal classifier design. Although this approach applies to the pattern recognition problem in general, it finds various
applications to speech recognition. It was first applied to
dynamic time warping based recognition system [12, 36].
Application to hidden Markov model based continuous
speech recognition systems was formulated as a segmental
and string model based MCE approach [14, 15], and
successful applications of this approach were reported in
[16, 21, 25, 54, 55]. This approach was further extended
to form a combined string model, in which training of
model components of the speech recognition system can
be achieved under a unified MCE framework [17, 29]. It
was applied to discriminative model combination[9, 53]
and to applications in speaker identification and verifica-
tion [48, 26, 41]. The basic idea of MCE approach
was further developed for applications in general utter-
ance verification problem [71, 51]. A general framework
of combining detection and verification in speech recog-
nition and understanding was also proposed, in which the
discriminant function based pattern recognition approach
was applied in both detection and verification processes
[35, 40].

2. DISCRIMINANT FUNCTION BASED
CLASSIFIER DESIGN

One of the important issues in classifier design is the selec-
tion of the loss function, upon which the optimal classifier
design can be characterized as the one which achieves
the minimum loss. To illustrate the concept, consider an
optimization of the loss function, upon which the optimal classifier
can be achieved under a unified MCE framework [17, 29]. It
was applied to discriminative model combination[9, 53]
and to applications in speaker identification and verifica-
tion [48, 26, 41]. The basic idea of MCE approach
was further developed for applications in general utter-
ance verification problem [71, 51]. A general framework
of combining detection and verification in speech recog-
nition and understanding was also proposed, in which the
discriminant function based pattern recognition approach
was applied in both detection and verification processes
[35, 40].

The optimal classifier from the Bayes theory implements
the “maximum a posterior” MAP decision:

\[ C(x) = C_i \ \text{iff} \ \frac{P(C_i \mid x)}{P(C_j \mid x)} = \max_i \frac{P(C_j \mid x)}{P(C_i \mid x)}. \] (5)

When true posterior probabilities are known, MAP deci-
sion implements an optimal classifier which minimize
the expected loss based on (4). The minimum error rate
achieved by MAP decision rule is called “Bayes risk”. How-
ever, if the true posterior probabilities are not known or
the decision rule is not based on the class posterior prob-
ability, then we can not use this result directly to claim the
optimality of the classifier. For most real world problems,
the posterior probabilities in Bayes decision theory are not
known priori. Our choice of the distribution used in the
classifier is often limited by the mathematical tractability
and very likely to be inconsistent with the true distribu-
tion. This means the true MAP decision can rarely be
implemented and the minimum Bayes risk generally re-
mains an unachievable lower bound.

Discriminant functions on the other hand are those func-
tions which characterize the decision rule of the classifier.
They may or may not be probability or likelihood based
functions, and they can come from different parametric
families, including those families which have no relation
to the parametric form of the class posterior distribution
as required in the classical Bayes decision theory. One
well studied family of discriminant function is the linear
discriminant function which has received considerable at-
tention and theoretical development for its design. A
minimum error rate classifier design using linear discriminant
function was studied in [22]. For M class classification
problem using discriminant functions, a set of discriminant
functions \( \{g_i(x) \mid i = 1, \ldots, M\} \) are used and the classifier
\( C(x) \) is defined such that

\[ C(x) = I \ \text{iff} \ I = \arg \max_i g_i(x). \] (6)

When the loss function \( R(C(x) \mid x) \) is specified, the prob-
lem of optimal classifier design using discriminant func-
tions becomes a minimization problem of finding a best
set of discriminant functions \( \{g_i(x) \mid i = 1, \ldots, M\} \) from a
class of discriminant functions which minimizes the ex-
pected loss \( L \) as defined in Eq. (2).

3. MCE CLASSIFIER DESIGN USING
DISCRIMINANT FUNCTIONS

As pointed in the previous section that without the knowl-
edge of the form of the true class posterior probabilities
required in the classical Bayes decision theory, classifier
design by distribution estimation often does not lead to an
optimal performance. This motivates the effort of search-

ing for other alternative criteria in classifier design. In
particular, criteria of MMI (maximum mutual informa-
tion) and MDI (minimum discriminative information) are
used in many applications[3, 57]. Although these meth-
ods demonstrate significant performance advantages over
the traditional ML approach, they are not based on a direct minimization of a loss function which links to the classification error rate.

Minimum classification error (MCE) rate based classifier design is based on a direct minimization of a loss function which relates to recognition error rate. Do-Tu et al. [22] studied MCE solution for the two class non-parametric classification problem using linear discriminant functions. A general approach for multi-class and non-linear discriminant functions are proposed by Juang et al. [32]. In MCE approach, the classifier design and parameter estimation are to correctly discriminate the observations for best recognition/classification results rather than to fit the distributions to the data. This is often achieved through a three steps process.

1) The misclassification measure in the MCE based approach is defined as

\[ d_i(X) = -g_i(X; \Lambda) + \log \left[ \frac{1}{M-1} \sum_{j, j \neq i} \exp[g_j(X; \Lambda)] \right]^{1/\eta} \]  (7)

where \( \eta \) is a positive number [14]. For an \( i^{th} \) class speech utterance \( X \), \( d_i(X) > 0 \) implies misclassification and \( d_i(X) \leq 0 \) means a correct decision. When \( \eta \) approaches \( \infty \), the term in the bracket is the \( L_\infty \) norm on the discrete integer set \( \{ j | j \neq i, j = 1...M \} \) which converges to the \( \| x \|_{\infty} \) norm and becomes \( \max_{j, j \neq i} g_j(X; \Lambda) \). By varying the value of \( \eta \) and \( M \), one can take all the competing classes into consideration, according to the individual significance, when searching for the classifier parameter \( \Lambda \).

2) The loss function is used for recognition error rate minimization. The misclassification measure of (7) is embedded in a smooth zero-one function, for which any member of the sigmoid function family is an obvious candidate. A general form of the loss function can then be defined as:

\[ \ell_i(X; \Lambda) = \ell(d_i(X)) \]  (8)

where \( \ell \) is a sigmoid function, one example of which is

\[ \ell(d) = \frac{1}{1 + \exp(-\gamma d + \theta)} \]  (9)

with \( \theta \) normally set to 0 and \( \gamma \) set to greater or equal to one. Clearly, when \( d_i(X) \) is much smaller than zero and negative, which implies correct classification, virtually no loss is incurred. When \( d_i(X) \) is positive, it leads to a penalty which becomes essentially a classification/recognition error count.

3) The classifier parameter estimation is based on the minimization of the expected loss. For any unknown object \( X \), the classifier performance is measured by

\[ \ell(X; \Lambda) = \sum_{i=1}^{M} \ell_i(X; \Lambda) \mathbb{1}(X \in C_i) \]  (10)

where \( \mathbb{1}(\cdot) \) is the indicator function. The expected loss, which is related to recognition error rate, is given by

\[ L(\Lambda) = \mathbb{E}_X[\ell(X; \Lambda)]. \]  (11)

This 3-step process in MCE approach emulates the classification operation as well as the recognition error rate based performance evaluation in a smooth functional form, suitable for classifier parameter optimization. It is important to point out that if the correct form of posterior probabilities \( P_i(x | \Lambda) \) is used as the discriminant function, the minimum loss solution in MCE approach is an approximation to the Bayes minimum risk [20].

### 3.1. Optimization Methods

Various minimization algorithms can be used to minimize the expected loss. The generalized probabilistic descent (GPD) algorithm is a powerful algorithm that can be used to accomplish this task [2]. In GPD based minimization algorithm, the target function \( L(\Lambda) \) is minimized according to an iterative procedure:

\[ \Lambda_{t+1} = \Lambda_t - \epsilon_t U_t \nabla \ell(X_t, \Lambda) \bigg|_{\Lambda=\Lambda_t} \]  (12)

where \( U_t \) is a positive definite matrix [19], \( \epsilon_t \) is a sequence of positive numbers, and \( \nabla \ell(X_t, \Lambda) \bigg|_{\Lambda=\Lambda_t} \) is the gradient function of the loss function at \( \Lambda = \Lambda_t \), and \( X_t \) is the \( t \)-th training sample used in the sequential training process. The convergence properties of GPD algorithm was studied in the literature (e.g. [19, 11]) under various names, such as stochastic approximation, etc.

MCE approach is very specific to the form and structure of the discriminant function and loss function regarding the classifier. It is relatively unrestricted to the optimization methods which can be used to minimize the expected loss. Many innovations are possible for better optimization results. In particular, methods of linear programming [58], gradient projection [28] and growth-transformation [27, 34, 57] are also used for minimization of the expected loss in MCE classifier design.

### 3.2. HMM as a Discriminant Function

There are several ways of using hidden Markov models (HMMs) as discriminant functions. In particular, the discriminant function can be based on the joint observation-state probability defined as follows:

\[ g_i(X, q; \Lambda) = P(i) X(q; \Lambda) = \pi_{q_0}^{(i)} \prod_{t=1}^{T} a_{q_{t-1} \rightarrow q_t}^{(i)} g_{q_t}^{(i)} (x_t) \]  (13)

Discriminant function defined in (13) is a natural choice for speech recognition system using Viterbi decoding, where \( a_{q_{t-1} \rightarrow q_t}^{(i)} \) is based on the optimal state sequence. The MCE training algorithm based on (13) is often called segmental GPD, and MCE estimation of HMM parameters is derived for GPD based optimization method (see [33] for details). In order to maintain the probabilistic structure of HMMs, cares must be taken and parameters are often derived through certain transformations [14, 33].
3.3. Relation of MCE and MMI

Maximum mutual information (MMI) [3] is another powerful approach used in speech recognition. The MMI approach is based on the mutual information \( I(W_c, X) \) between the acoustic observation \( X \) and its correct lexical symbol \( W_c \). For the \( N \) class classification problem, the logarithm of the mutual information has the following form

\[
I(W_c, X) = \log \frac{p(W_c, X)}{p(W_c)p(X)}
\]

\[
= \log \left( \frac{p(X | W_c)}{\sum_{k=1}^{N} p(W_k)p(X | W_k)} \right) \quad (14)
\]

where \( W_k \) runs over all possible \( N \) class symbols, \( \log p(X | W_c) \) and \( \log p(X | W_k) \) are the log-likelihood scores of \( X \) on the correct lexical symbol and the \( k \)-th lexical symbol respectively. From Eq. (14),

\[
\log p(W_c | X) = \log \left( \frac{p(W_c, X)}{p(X)} \right) = I(W_c, X) + \log p(W_c),
\]

which relates \( I(W_c, X) \) to the posterior probability \( p(W_c | X) \).

In MMI training, the criterion of the classifier design and parameter estimation is to maximize the average mutual information \( I(W_c, X) \) on the training set. The relation between MMI and MCE is a very interesting topic and studied in [63, 64]. It is found that under certain conditions, direct comparisons can be made between these two approaches. In particular, when \( \eta = 1 \), the loss function used in MMI approach relates to the misclassification measure in MCE approach as follows [63]:

\[
I(W_c, X) = \log \left( \frac{1}{1 + e^{d_c(X) + \log(N-1)}} \right) + \log N, \quad (16)
\]

where \( d_c(X) \) is the misclassification measure in MCE for correct class \( W_c \).

From the form of the loss function used in MCE and MMI, it is easy to see that MMI approach is based on modeling the posterior distribution which relates to the optimal decision boundary of the classifier through the Bayes theory. On the other hand, MCE approach is based on an explicit modeling of the decision boundary which has a direct relation to the recognition error rate. This direct relation to recognition error rate in the MCE approach has several advantages. It is meaningful in the sense of minimizing the empirical recognition error rate of the classifier, and this property is not dependent on the parametric form of the discriminant function nor its relation to the form of the true class posterior distribution. If the true class posterior distributions are used as discriminant functions, the asymptotic behavior of the classifier will approximate the minimum Bayes risk.

The significance of the MCE approach in speech recognition is twofold. First, a classifier design based on direct minimization of the recognition error rate is a meaningful alternative to distribution estimation based approach. Second, the probability distributions (PDs) used in parametric modeling of speech are very limited comparing to the true PDs in the signal source, and the decision rule based on discriminant function approach is a reasonable alternative to “plug-in” MAP rule which is based on the model correctness assumption. More discussion and studies of these approaches were given in [20, 63].

4. EMBEDDED STRING MODEL BASED MCE APPROACH

In continuous speech recognition using sub-word model units, \( X \) is a concatenated string of observations belonging to different classes. For example, a sentence is a sequence of words, each of which is to be modeled by a distribution. The decoding process in continuous speech recognition is to compare (implicitly) all possible (word or subword) string models, and the word string whose string model has the highest likelihood score is chosen as the decoded string. The likelihood score of the word string is typically a combination of scores from various models, including scores from the acoustic model, language model, duration model, etc. The main reason of adopting this type of string model is simply that the basic speech recognition model units, which are used to form string models, can be estimated from a finite amount of available training data. On the other hand, the use of long term language model and context dependent acoustic model has extended the modeling dependencies beyond the level of individual words to phrase groups or even to the whole utterance level.

In [15], an MCE approach to classifier design based on string model was proposed and studied. The string model for a given word string \( S \) in a HMM based speech recognition system using continuous observation densities is given by

\[
S_Q = \arg \max_{S_Q} \log f(X, \Theta_{S_Q} | S_Q)
\]

where \( S_Q \) is a string model for word string \( S \), \( \Theta_{S_Q} \) is the optimal state sequence in the string model of \( S_Q \) and \( \log f(X, \Theta_{S_Q} | S_Q) \) is the log-likelihood score along the optimal state sequence \( \Theta_{S_Q} \). The discriminant functions, for \( k = 1, \ldots, N \), are

\[
g(X, S_k, \Lambda) = \log f(X, \Theta_{S_k}, S_k | \Lambda), \quad (18)
\]

where \( S_k \) is the \( k \)-th best string, \( \Lambda \) is the HMM set used in the \( N \)-best decoding, \( \Theta_k \) is the optimal path (state sequence) of the \( k \)-th string given the model set \( \Lambda \), and \( \log f(X, \Theta_{S_k}, S_k | \Lambda) \) is the related log-likelihood score on the optimal path of the \( k \)-th string.

For the correct string \( S_{\text{lex}} \), the discriminant function is given by

\[
g(X, S_{\text{lex}}, \Lambda) = \log f(X, \Theta_{S_{\text{lex}}}, S_{\text{lex}} | \Lambda), \quad (19)
\]

where \( S_{\text{lex}} \) is the correct string, \( \Theta_{S_{\text{lex}}} \) is the optimal alignment path and \( \log f(X, \Theta_{S_{\text{lex}}}, S_{\text{lex}} | \Lambda) \) is the corresponding log-likelihood score. These discriminant functions are
embedded in the MCE based loss function. The embedded string model based MCE approach is well suited for acoustic modeling using detailed context dependent models. It can describe various context dependencies such as triphone, quinphone, etc. In embedded string model based MCE approach, the long term dependencies are embedded in the basic speech recognition model units even their original context dependency definitions are not. It is observed in the experiments that many mono-phone based context independent model units obtained from the MCE approach exhibit speech recognition performance of context dependent model units[25]. The embedded string model based MCE approach found applications in various recognition tasks and significant error rate reduction were observed [15, 16, 25, 55, 33, 21].

5. COMBINED STRING MODEL BASED MCE APPROACH

In speech recognition, the final decision is based on the combination of scores from various knowledge sources represented by different models. Assuming independence of each model, the final score in the logarithm domain becomes a sum of log-likelihood scores from each individual model. In particular, in addition to the acoustic model, if a language model is used and its score is weighted by a weighting factor \( \lambda_L \), the final likelihood score of a candidate string is

\[
\log f(X \mid W) = \log f(X \mid W) + \lambda_L \log P(W).
\]

(20)

If the model correctness assumption is valid, the log-likelihood score should strictly follow Eq. (20) and the score weighting factor \( \lambda_L = 1 \). However, in speech recognition experiments and applications, it is found that a value of \( \lambda_L \) with \( \lambda_L \neq 1 \) demonstrates much better recognition performance[42], an indication that the true distribution of the signal source departs from the assumption made by the model. The combined string model is an MCE approach which is based on the discriminant function at the level of global combined string model. The combined string model can be done in the combination of the following two directions. One is horizontal, scores from multiple models and different knowledge sources are combined to form the final score, where each individual model may be estimated separately based on different estimation methods, including using different training data and constraints [9]. Another important direction for model combination is to estimate the individual model parameters in the combined string model as an integrated component of the final combined string model, in which the discriminant function is constructed at the combined string model level. The estimation of parameters at each individual model is achieved by tracing down the model combination tree to each of its leaf nodes following a chain rule like relationship [17, 9, 38, 73].

5.1. Discriminative Model Combination (DMC)

In speech recognition, multiple signal sources from multiple signal bands are also used in speech recognition [53]. Many of these models or knowledge sources may not be based on probabilities, and a discriminant function based approach is a suitable choice for model combination. Let \( \{\Lambda_1, \ldots, \Lambda_M\} \) be the individual model components in the model combination. We use the notation \( G(X \mid \Lambda_1, \ldots, \Lambda_M) \) to denote the combined string model given random input \( X \), where \( G \) is the function selected for model combination. If \( G \) is linear

\[
G(X \mid \Lambda_1, \ldots, \Lambda_M) = \sum_{k=1}^{M} \lambda_k g(X \mid \Lambda_k),
\]

(21)

where \( g(X \mid \Lambda_k) \) is the score from the \( k \)-th model. Discriminative model combination based on the MCE approach is to embed the combined string model based discriminant function \( G(X \mid \Lambda_1, \ldots, \Lambda_M) \) in the error rate based loss function and estimate the model combination factor \( \lambda_1, \ldots, \lambda_M \) as parameters in the combined string model. Many optimization methods can be applied to estimate \( \lambda_1, \ldots, \lambda_M \) which minimize the expected loss. The popular GPD algorithm has a very simple form in this case[9]. DMC is applied in many applications under the name of combined string model[17], discriminative model combination[9] and universal stochastic engine[29]. The MCE based discriminant function based approach provides a goodness criterion for estimating and adjusting those “tuning parameters” in speech recognition, especially when either a model correctness assumption is not valid or a unified framework is needed to combine knowledge sources that are different in origin or nature.

5.2. Discriminative Feature Extraction

The goal of discriminative feature extraction is to accomplish the speech recognition feature extraction from the standpoint of minimizing the recognition error rate for classification by machines. In place of the Bark scale from hearing, a new frequency scaling [62] can be derived. Feature extraction can be made discriminative based on the combined string model MCE paradigm. Since these speech recognition feature vectors are part of the combined string model, a goodness criterion of this approach can be derived from the relation of minimization of the expected loss and recognition error rate in the MCE formulation.

One application of discriminative feature extraction is in the design of cepstral lifters[10]. The justification of this type of lifter for cepstral feature vectors is given in [31]. In discriminative feature extraction, lifter design can be done according to the combined string model formulation. Instead of relying on human hearing capabilities for choosing the right lifter, the lifter parameters can be estimated using the MCE criterion to minimize the recognition error rate. Speech experiments using discriminative feature extraction based lifter were performed in several tasks [10, 38].

6. VERIFICATION AND IDENTIFICATION

Speaker verification and identification based on voice is an important area in speech research and has been studied
for several decades. In making a decision regarding the true identity of the speaker, two types of errors may occur. One is false rejection (type I error) and one is false acceptance (type II error). The problem can be conveniently formulated as a statistical hypothesis testing problem to test null hypothesis \( H_0 \) (i.e. \( X \) from source \( S_0 \)) and the alternative hypothesis \( H_1 \) (i.e. \( X \) from source \( S_1 \)). There are plenty of studies available in the statistical literature regarding the design of the optimal tests if \( P(X \mid H_0) \) and \( P(X \mid H_1) \) are known and fall into some specific distributions, such as the exponential family [43]. In practice, the test procedure is often based on a test statistic \( T(X) \) such that \( H_0 \) is rejected if \( T(X) \geq k \). According to the Neyman-Pearson Lemma, \( T(X) \) can be based on probability ratio test \( T(X) = \frac{P(X \mid H_1)}{P(X \mid H_0)} \) or likelihood ratio test \( T(X) = \frac{f(X \mid H_1)}{f(X \mid H_0)} \), and \( k \) is selected such that the level of the type I error \( P(T(X) \geq k \mid H_0) = \alpha \). However, in practical verification problems, we have no exact knowledge regarding the distributions of null and alternative hypotheses. Discriminant function based methods can be applied in designing the verifier.

6.1. Speaker Verification and Identification

In speaker verification and identification, the decision is based on the score of the test statistic \( T(X) \). For speaker identification problem, the speaker \( I \) is identified as the true speaker if

\[
I = \text{argmax}_q T(X, q). \tag{22}
\]

It can be seen that speaker identification can be formulated as an M class classification problem using associated test statistics. A suitable misclassification measure can be constructed according to the test statistics [48].

For speaker verification, we usually accept the claimed speaker identity \( I \) if the test statistic for \( I \)-th speaker

\[
T(X, I) > \tau. \tag{23}
\]

For speaker verification, the verification error can also be characterized by a mis-verification measure based on the test statistic \( T(X, I) \), similar to the misclassification measure in recognition. The MCE based discriminant function approach can be adapted to minimize the verification error [48, 41].

In addition to addressing the identification and verification problem as a special classification problem, methods of introducing the structure of statistical test in discriminant function based approach are also attempted. Minimum verification error (MVE) training is such an approach [51, 69]. The main difference to MCE approach is the use of two separate loss functions to model two types of errors in hypothesis testing. We describe the details of MVE approach below and exemplify the discriminant function approach to speaker verification and identification. The mis-verification measure, as opposed to the misclassification measure, for the \( q \)-th speaker is defined as follows:

\[
d_q(X, \Lambda) = \begin{cases} 
-g_d(X, \Lambda) + G_q(X, \Lambda) & \text{from } q \\
-G_q(X, \Lambda) + g_d(X, \Lambda) & \text{otherwise,} 
\end{cases} \tag{24}
\]

where \( g_d(X, \Lambda) \) is the log-likelihood score from the claimed \( q \)-th speaker model and \( G_q(X, \Lambda) \) is the score from the cohort speaker group \( C_q \). The mis-verification measure is embedded in a smooth sigmoid based loss function \( \ell(d(X, \Lambda)) \) as in the MCE formulation. Two separate loss functions are used to describe the type I and type II errors. The average loss for each type of errors approximates the empirical type I and type II error rate on the training samples.

\[
L_1(X, \Lambda) = \frac{1}{N_1} \sum_{i=1}^{N_1} \ell(d(X_i, \Lambda)) 1(X_i \in \text{Claimed Speaker}) \tag{25}
\]

and

\[
L_2(X, \Lambda) = \frac{1}{N_2} \sum_{i=1}^{N_2} \ell(d(X_i, \Lambda)) 1(X_i \notin \text{Claimed Speaker}), \tag{26}
\]

where \( 1() \) is the indicator function. The overall expected loss of the MVE is given by

\[
L(X, \Lambda) = \lambda_1 L_1(X, \Lambda) + \lambda_2 L_2(X, \Lambda), \tag{27}
\]

where \( \lambda_1 \) and \( \lambda_2 \) are design parameters which control the influence of type I and type II errors in the overall loss function. The model parameter estimation in MVE training is to minimize the expected loss of Eq. (27), which relates to the minimization of empirical error rate of type I and type II errors. Various speaker verification experiments are conducted and the discriminant function based approach demonstrates significant performance advantages over the distribution estimation based approach [48, 36, 41].

7. SUMMARY

In this paper, the discriminant function based MCE approach was introduced as an alternative to the distribution estimation based approach from Bayes decision theory. In MCE approach, the classifier design is to find a set of parameters which minimize the empirical recognition error rate. This is achieved through a special loss function where minimizing the expected loss relates to the reduction of the recognition error rate. The discriminant function based MCE approach applies to cases where the traditional distribution estimation based approach does not apply, especially when the family of the discriminant functions encountered in the classifier are not based on probability distributions. The goodness of this approach is justified without the model correctness assumption, and it applies to cases where the model correctness assumption is known to be invalid. Although significant progress has been made during the last ten years, many issues in discriminant function based MCE approach are still open and we are just beginning to look at new potentials of this approach in speech recognition.
8. REFERENCES


