S1.2

On the Relations Between Modeling Approaches for Information Sources

Yariv Ephraim and Lawrence R. Rabiner

Speech Research Department
AT&T Bell Laboratories
Murray Hill, NJ 07974

ABSTRACT

We examine the relations between maximum likelihood (ML), maximum mutual information (MMI), and minimum discrimination information (MDI) modeling approaches, which have been applied to estimating acoustic word models in speech recognition systems. We show that all three approaches can be uniformly formulated as MDI modeling approaches for estimating the acoustic models for all words simultaneously. The three approaches differ in either the probability distribution (PD) attributed to the source being modeled or in the model effectively being used. None of the approaches, however, assumes model correctness, i.e., that the source has the PD of the model. A new modeling approach is proposed, which, in contrast with the other approaches considered here, directly aims at the minimization of the probability of error.

1. Introduction

Speech recognition could be optimally performed if the probability of any word[1] in the recognizer's vocabulary and the probability distribution (PD) of the corresponding acoustic signal were known. In this case the recognizer which is optimal in the sense of minimizing the probability of error is a maximum a-posteriori (MAP) estimator which chooses from all possible words in the vocabulary, the word which together with the acoustic input signal yields the highest joint probability. In practice, the word probability and the PD of the acoustic signal are not known and hence only suboptimal recognizers can be implemented.

The commonly used recognition approach is first to choose probabilistic models for the words and for the corresponding acoustic signals, and then to estimate the parameters of the models from appropriate word and acoustic training sequences. Once the parameters of the models have been estimated, the optimal MAP decision rule is applied to the estimated PD's, as if they were the true ones. The acoustic model is usually chosen to be a Markov source, or a hidden Markov model (HMM), and its parameters are estimated by the Maximum Likelihood (ML) estimation approach [1]. The word model, or the language model, is also often chosen to be Markovian.

Since no fundamental achievable recognition bounds (similar to Shannon bounds in coding theory) are known, and since the optimal recognizer cannot be implemented, it is not clear how good (or bad) the performance of current state of the art recognizers is as compared to the ultimate achievable performance. Hence, the assumptions upon which the suboptimal recognizer are based have been repeatedly challenged in an attempt to improve recognition accuracy. While there is common agreement among researchers on the validity of using Markovian models for words and acoustic signals, the best way to estimate the parameters of these models is still an open question. For example, it is not clear in what sense an HMM, which has been estimated by the so-called ML estimation approach, is optimal, since, in general, the acoustic signal is not a Markov source. Hence, the nice properties of the ML estimator, which are valid if the source and the model are the same, do not necessarily hold here. Another relevant question concerns the decision rule for recognition. The MAP decision rule is optimal if the probabilities of the words and the PD's of the acoustic signals are correct.

Recently, two new approaches for estimating the parameters of the acoustic HMM model were proposed [2], [3]. The first is optimal in the maximum mutual information (MMI) sense. This approach assumes that a language model is given and attempts to find the set of acoustic models which, together with the given language model, has the maximum possible mutual information with respect to the given training sequence. The second approach only considers the acoustic model and is optimal in the minimum discrimination information (MDI) sense. The objective here is to find the model which has the minimum discrimination information (or cross entropy) with respect to *all* sources which could have generated the measurements given in the training sequence from the actual source. The measurements are usually given in terms of moments of the acoustic signal, e.g., correlation vectors from the source.

The MMI and MDI approaches aim indirectly at reducing the error rate of the recognizer, since the objective of the two approaches is to improve the estimation of the PD of the acoustic signal. It is nontrivial, however, to show theoretically that either approach actually reduces the probability of error. Hence, the extent to which each approach really improves recognition accuracy, as opposed to modeling accuracy, is experimentally demonstrated and therefore is highly task dependent.

The purpose of this note is first to show some relations between the ML, MMI, and MDI modeling approaches. In fact we will show that the three approaches can be uniformly formulated as being MDI modeling approaches which differ in the PD's used for the source to be modeled and the model itself. This formulation is important since it provides a unique common basis for comparing the three modeling approaches. Second, it clearly shows the difference between the approaches in terms of the assumptions being made about the true PD of the source to be modeled and about the model itself.

After establishing the above relations, we shall present a new approach for doing the modeling which is more directly related to our main objective, i.e., minimization of the recognition error rate. The note does not aim at providing any specific algorithm for implementing the newly proposed approach but rather serves to enlighten the subject.

1. The term "word" is referred here to any language unit being modeled acoustically, i.e., subword units, physical words, phrases, etc.

CH2561-9/88/0000-0024 $1.00 © 1988 IEEE

2. ML, MMI, and MDI Modeling Approaches

Let P_{λ_m} be the PD of the acoustic model for the m–th word in the vocabulary, where λ_m is the parameter set of the model. Let P_μ be the PD of the word model, where μ is the parameter set of the model. We assume that there are L words in the vocabulary, and hence, L acoustic models and a single word model have to be designed. Let $Q_{Y|M}$ and Q_M, respectively, be the PD's attributed to the acoustic signal from a specific word and to the word itself, where Y is a random variable representing the acoustic signal and M is a random variable representing the word. These two PD's are estimated from appropriate acoustic and word training sequences. Let $y_T(m)=\{y_t(m),\ t=0,\cdots,\ T\}$ be the given acoustic training sequence for the m–th word, where $y_t(m)\in R^K$, the K dimensional Euclidean space. Finally, let $\{w_j,\ j=0,\cdots,\ J\}$ be the given word training sequence.

To simplify the discussion, we will consider in all subsections, but subsection 2.3, the case where the space Y on which P_{λ_m} and $Q_{Y|M}$ are defined is finite. From the practical point of view, this assumption is always met as our models and training sequences are stored in a digital computer. From the theoretical point of view, this assumption will allow us to present the main ideas in this section in a simple way without using measure theoretic arguments. The extension to the case where Y is infinite can be found in Csiszar and Tusnady [4]. For a finite space Y, the conditional PD's P_{λ_m} and $Q_{Y|M}$, respectively, have probability mass functions (pmf's) $p(y\mid m)=p_{\lambda_m}(y)$ and $q(y\mid m)$, and they are absolutely continuous with respect to each other. Similarly, the PD's P_μ and Q_M are absolutely continuous with respect to each other and have pmf's $p(m)=p_\mu(m)$ and $q(m)$, respectively.

Let $Q_{Y,M}$ be the joint PD of the acoustic signal and the word. Similarly, let $P_{Y,M}$ be the joint PD of the acoustic and word models. Let $q(y,m)$ and $p(y,m)$, respectively, be the pmf's corresponding to $Q_{Y,M}$ and $P_{Y,M}$. We have that $q(y,m)=q(y\mid m)q(m)$ and $p(y,m)=p_{\lambda_m}(y)p_\mu(m)$.

The discrimination information between $Q_{Y,M}$ and $P_{Y,M}$ will play a central role in this section. It is given by

$$D(Q_{Y,M}\|P_{Y,M}) = \sum_{m=1}^{L}\sum_{y\in Y} q(y,m)\ln\frac{q(y,m)}{p(y,m)} \tag{1}$$
$$= D(Q_M\|P_\mu) + \sum_{m=1}^{L} q(m)D(Q_{Y|M=m}\|P_{\lambda_m}),$$

where,

$$D(Q_M\|P_\mu) \triangleq \sum_{m=1}^{L} q(m)\ln\frac{q(m)}{p_\mu(m)} \tag{2}$$

is the discrimination information between the PD attributed to the word and the parametric word model, and

$$D(Q_{Y|M=m}\|P_{\lambda_m}) \triangleq \sum_{y\in Y} q(y\mid m)\ln\frac{q(y\mid m)}{p_{\lambda_m}(y)} \tag{3}$$

is the discrimination information between the PD attributed to the acoustic signal from the m–th word and the acoustic parametric model for that word. The second expression in (1) suggests that

$$\min_{\mu,\{\lambda_m\}_{m=1}^{L}} D(Q_{Y,M}\|P_{Y,M}) = \tag{4}$$
$$\min_{\mu}\left\{D(Q_M\|P_\mu) + \min_{\{\lambda_m\}_{m=1}^{L}}\sum_{m=1}^{L} q(m)D(Q_{Y|M=m}\|P_{\lambda_m})\right\}.$$

This means that if P_{λ_m} does not depend on μ, the word model parameter set, then jointly optimal word and acoustic modeling, in the sense of minimizing the discrimination information $D(Q_{Y,M}\|P_{Y,M})$, can be independently performed by minimizing $D(Q_M\|P_\mu)$ and $D(Q_{Y|M=m}\|P_{\lambda_m})$, respectively. As we shall see, this will be the case for the acoustic models used in the ML and MDI approach of [3], but not for the model used by the MMI approach.

In this correspondence we are only concerned with the estimation of the acoustic model. Hence, whenever necessary, we will assume knowledge of the word model. We now examine the three approaches for acoustic modeling mentioned in Section 1, namely, the ML, MMI, and MDI, and show that each approach effectively minimizes the discrimination information $D(Q_{Y,M}\|P_{Y,M})$ for a particular pair of $q(y\mid m)$ and p_{λ_m}.

2.1 ML Estimation

In the ML estimation approach we estimate the parameter set λ_m, given the training sequence $y_T(m)$, by

$$\max_{\lambda_m} \ln p_{\lambda_m}(y_T(m)). \tag{5}$$

Let $Q_{Y|M}$ be the empirical distribution of the m–th training sequence, i.e., the pmf $q(y\mid m)$ is given by

$$q(y\mid m) = \chi(y-y_T(m)) \triangleq \begin{cases} 1, & y=y_T(m) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $\chi(\cdot)$ is a probability measure which is concentrated on $y_T(m)$. On substituting (6) into (1) we have that

$$D(Q_{Y,M}\|P_{Y,M}) = D(Q_M\|P_\mu) - \sum_{m=1}^{L} q(m)\ln p_{\lambda_m}(y_T(m)), \tag{7}$$

where we have used

$$\sum_{y\in Y} q(y\mid m)\ln q(y\mid m) = \ln q(y_T(m)\mid m) = \ln 1 = 0. \tag{8}$$

Since the first term in (7) is independent of λ_m, we have

$$\operatorname*{argmin}_{\lambda_m} D(Q_{Y,M}\|P_{Y,M}) = \operatorname*{argmax}_{\lambda_m} \ln p_{\lambda_m}(y_T(m)) \tag{9}$$

which means that the standard ML estimation approach is an MDI modeling approach, for estimating all acoustic models simultaneously, when the PD attributed to the source is the probability measure which is concentrated in the acoustic training sequences for the different words in the vocabulary. Note that here we have attributed to the source a PD which is concentrated in the entire training sequence from each word, rather than in the individual vectors of that training sequence. The latter PD has been used in [4] when treating independent identically distributed (i.i.d) vector sources. The reason is that in our case the speech is not an i.i.d vector source nor is the commonly used hidden Markov acoustic model.

Note also that in the MDI derivation of the ML approach, the pmf $q(y\mid m)$ attributed to the source and the pmf p_{λ_m} of the model, are independently chosen. Hence, the ML estimation approach does not require any model correctness assumption, in which the PD attributed to the source is that of the model, as was suggested, for example, in [2], [5].

2.2 MMI Estimation

We now apply the same techniques to show that the MMI estimation approach is, in fact, an MDI modeling approach for estimating all acoustic models simultaneously using the empirical distribution of the acoustic signal from all words. The MMI modeling approach was first proposed in [6, p. 262], and was first applied to modeling speech signals in [2]. Let $I(Y\|M)$ be the mutual information between the two random variables Y and M. The random variable Y

obtains values in the finite space Y, and the random variable M can only have L values, the number of words in the vocabulary. The mutual information between Y and M is given by

$$I(Y\|M)=\sum_{m=1}^{L}\sum_{y\in Y} q^*(y\mid m)q^*(m)\ln\frac{q^*(y\mid m)}{\sum_{m'=1}^{L} q^*(y\mid m')q^*(m')}, \quad (10)$$

where $q^*(y\mid m)$ and $q^*(m)$ are the true pmf's of Y given M and of M, respectively. Since those pmf's are not known, the modeling approach proposed in [2], is to replace the pmf's in the argument of the information measure (i.e., the argument of the ln function) by the pmf's of the parametric models, and to calculate the expected value involved in (10) with respect to an estimate $q(y\mid m)$ of the true pmf $q^*(y\mid m)$ of the source. This results in

$$I(Y\|M)=\sum_{m=1}^{L}\sum_{y\in Y} q(y\mid m)q(m)\ln\frac{p_{\lambda_m}(y)}{\sum_{m'=1}^{L} p_{\lambda_{m'}}(y)p_\mu(m')}. \quad (11)$$

The estimate of $q(y\mid m)$ suggested in [2] is the empirical distribution (6). Substituting (6) into (11) gives

$$I(Y\|M)=\sum_{m=1}^{L} q(m)\ln\frac{p_{\lambda_m}(y_T(m))}{\sum_{m'=1}^{L} p_{\lambda_{m'}}(y_T(m))p_\mu(m')}, \quad (12)$$

which should be maximized over all $\{\lambda_m,\ m=1,\ \cdots,\ L\}$.

To formulate the MMI approach as an MDI approach, let $q(y\mid m)$ be as in (6), and consider the following acoustic model for the m-th word in the vocabulary

$$p(y\mid m)=\frac{p_{\lambda_m}(y)}{\sum_{m'=1}^{L} p_{\lambda_{m'}}(y)p_\mu(m')}. \quad (13)$$

On substituting (6) and (13) into (1) we obtain

$$D(Q_{Y,M}\|P_{Y,M})=D(Q_m\|P_\mu)-I(Y\|M), \quad (14)$$

which is minimized by the acoustic models that maximize $I(Y\|M)$. This demonstrates that MMI is an MDI modeling approach for the source (6) and the model (13). Note that in contrast with the ML case, where the minimization of $D(Q_{Y,M}\|P_{Y,M})$ over all λ_m could be done term by term, here the maximization of $I(Y\|M)$ over all models must be done simultaneously. This is, of course, due to the specific form of the model (13) which ties together all individual acoustic models.

2.3 MDI Estimation

We now review the principles of the MDI modeling approach proposed in [3], for estimating an individual acoustic model from a given training sequence from the source. We shall then generalize this approach for multiple model estimation. The major difference between this approach and the previous approaches considered here, is in the way the PD attributed to the source is estimated from the given training data. Rather than assuming that the PD of the source is concentrated in the training sequence, we consider the set of all PD's which could have resulted in the set of measurements taken from the source being modeled. From this set of PD's, we choose the PD which yields minimum discrimination information with respect to the parametric acoustic model. The resulting PD is called the MDI PD. The modeling is done by minimizing the discrimination information between the MDI PD, and the PD of the model, over all parameters of the model.

In [3] we considered the special case of Gaussian autoregressive (AR) hidden Markov modeling where the measurements from the source, at each time t, were the values of the elements of the covariance matrix of $y_t(m)$ within a given symmetric band of this matrix. If $R_T(m)\triangleq\{R_t(m),\ t=0,\ \cdots,\ T\}$ denotes the set of given partial covariance matrices for the m-th word, and $\Omega(R_T(m))$ is the set of all PD's which satisfy these measurements, then the MDI modeling approach chooses the parameter set λ_m as

$$\min_{\lambda_m}\ \min_{Q\in\Omega(R_T(m))} D(Q\|P_{\lambda_m}). \quad (15)$$

The minimization in (15) can be efficiently implemented, as we have shown in [3], through alternating minimization of $D(Q\|P_{\lambda_m})$ once over all $Q\in\Omega(R_T(m))$ assuming P_{λ_m} is known, and then over all HMM's of the given order assuming the MDI PD is known. For a given HMM P_{λ_m}, and partial covariance matrices $\{R_t(m)\}$ such that each $R_t(m)$ has a positive definite extension, there exists a unique PD which minimizes $D(Q\|P_{\lambda_m})$ over all $Q\in\Omega(R_T(m))$. The probability density function (pdf) of this PD is given by

$$q(y\mid m)=C\,p_{\lambda_m}(y)\exp\{-\frac{1}{2}\sum_{t=0}^{T} y_t^{\#}\Lambda_t(m)y_t\}, \quad (16)$$

where, C is a normalization constant which makes $\int dy\,q(y\mid m)=1$; # denotes vector transpose; and $\{\Lambda_t(m)\}$ is a set of symmetric matrices of Lagrange multipliers which are chosen so that

$$\int dy_t\,q(y_t\mid m)y_t y_t^{\#}=R_t(m)\quad\text{within the given band.} \quad (17)$$

The discrimination information between the MDI PD and the given model is given by

$$\min_{Q\in\Omega(R_T(m))} D(Q\|P_{\lambda_m})=-\ln C-\frac{1}{2}\mathrm{tr}\sum_{t=0}^{T} R_t(m)\Lambda_t(m). \quad (18)$$

The Lagrange multipliers can be found by a maximization of the right hand side of (18) over all $\{\Lambda_t(m)\}$ for which $q(y\mid m)$ is a pdf [3]. Due to the uniqueness of the solution of (17), the function in (18) has only one maximum point and hence the maximization can be carried over by any standard optimization procedure. Given the MDI PD, a new HMM can be found by a procedure similar to the Baum algorithm using the "Forward-Backward" formulas [3]. The algorithm is iterated in the above manner until some convergence criterion is satisfied.

The extension of the above MDI approach to multiple model design can be done in two different ways. First, note from (1) that $D(Q_{Y,M}\|P_{Y,M})$ can be minimized term by term over $Q\in\Omega(R_T(m))$ and λ_m. Hence, the approach of [3] can simply be viewed as an MDI approach for multiple acoustic model design. The second extension is done as follows. Let $R\triangleq\{R_T(m),\ m=1,\ \cdots,\ L\}$ be the given sequence of measurements corresponding to the L words in the vocabulary. Let $\lambda\triangleq\{\lambda_1,\ \cdots,\ \lambda_L\}$ be the parameter sets of all L models. Let $z\triangleq\{z_T(m),\ m=1,\ \cdots,\ L\}$, where, $z_T(m)\triangleq\{z_t(m),\ t=0,\ \cdots,\ T\}$ and $z_t(m)\in R^K$. Finally, let

$$p_\lambda(z)=\prod_{m=1}^{L} p_{\lambda_m}(z_T(m)) \quad (19)$$

be the acoustic model for the L words. If the PD which minimizes $D(Q\|P_\lambda)$ over all $Q\in\Omega(R)$ exists, then it has the following well known form.

$$\begin{aligned} q(z) &= C\,p_\lambda(z)\exp\{-\frac{1}{2}\sum_{m=1}^{L}\sum_{t=0}^{T} z_t^{\#}(m)\Lambda_t(m)z_t(m)\}, \quad (20)\\ &= \prod_{m=1}^{L} C_m\,p_{\lambda_m}(z_T(m))\exp\{-\frac{1}{2}\sum_{t=0}^{T} z_t^{\#}(m)\Lambda_t(m)z_t(m)\},\\ &\triangleq \prod_{m=1}^{L} q(z_T(m)\mid m). \end{aligned}$$

In this case we have that

$$\min_{Q\in\Omega(R)} D(Q\|P_\lambda) = \sum_{m=1}^{L} \min_{Q\in\Omega(R_T(m))} D(Q\|P_{\lambda_m}), \qquad (21)$$

and hence, the minimization of (21) over λ can be done term by term. This results in the MDI modeling approach of [3].

2.4 Discussion

The analysis presented in this section shows that the three modeling approaches considered here, ML, MMI, and MDI, are all optimal modeling approaches, in the MDI sense, for simultaneous estimation of the acoustic models for all words in the vocabulary. The approaches differ in the PD's attributed to the source and the model effectively being used. The ML and MMI approaches make precisely the same assumptions about the source being modeled. They both attribute to the source a PD which is concentrated in the individual acoustic training sequences corresponding to the different words in the vocabulary. The MDI approach attributes to the source a more robust PD estimate by considering all PD's which could have resulted in the given measurements from the source. Since in MDI modeling the PD attributed to the source and the model being used are independently chosen, none of the three approaches assumes model correctness.

The acoustic models used by the ML and the MDI approaches allow independent estimation of each individual model using the acoustic training sequence corresponding to the word being modeled. The acoustic models used by the MMI approach, however, are all tied together and hence all acoustic models must be simultaneously estimated. Furthermore, optimal joint estimation, in the MDI sense, of the acoustic and word models can be independently done for the acoustic models used in the ML and MDI approaches, since for those models the first term in (4) does not depend on the word model. This is not the case for the acoustic model used in the MMI approach. Hence, for the MMI acoustic model and a given word model to be jointly optimal in the MDI sense, both models should be simultaneously designed.

It is worthwhile mentioning that for the most popular acoustic models, namely HMM's, the ML and MDI approaches can be efficiently implemented by the well known Baum algorithm [1], and its generalization [3], respectively, while no efficient implementation is known for the MMI approach. This approach is usually implemented using general optimization procedures (e.g., descent methods).

Table 1 summarizes the three modeling approaches in terms of the PD attributed to the source and the model being used.

	ML	MMI	MDI
model	$p_{\lambda_m}(y)$	$\frac{p_{\lambda_m}(y)}{\sum_{m'=1}^{L} p_{\lambda_{m'}}(y)p(m')}$	$p_{\lambda_m}(y)$
source	$\chi(y-y_T(m))$	$\chi(y-y_T(m))$	$Cp_{\lambda_m}(y)\exp\{-\frac{1}{2}\sum_{i=0}^{T} y_i^t \Lambda_i(m) y_i\}$

3. Model design for minimum probability of error

In this section we propose a new approach for designing all L acoustic models simultaneously. This approach assumes that the decision rule is given in terms of the acoustic models $\{p_{\lambda_m}\}$ and the word model p_μ, and aims at estimating the parameters of those models so as to maximize the probability of correct recognition.

The probability of correct decision is given by

$$P_C = \sum_{m=1}^{L} \int_{\omega_m(\lambda,\mu)} q^*(y \mid m) q^*(m) dy \qquad (22)$$

where, $q^*(y \mid m)$ is the true pdf of the acoustic signal from the m-th word, $q^*(m)$ is the true probability of occurrence of the m-th word, and $\omega_m(\lambda,\mu)$ is the set of all y's for which the m-th word will be chosen. The decision rule assumed known here is the likelihood ratio for the estimated models and it is given by

$$\omega_m(\lambda,\mu) = \{y:\ \ln\frac{p_{\lambda_m}(y)p_\mu(m)}{p_{\lambda_l}(y)p_\mu(l)} > 0, \quad l\neq m\}. \qquad (23)$$

Define

$$g_{\lambda,\mu}(y,m) \triangleq \begin{cases} 1 & \ln\frac{p_{\lambda_m}(y)p_\mu(m)}{p_{\lambda_l}(y)p_\mu(l)} > 0 \quad , l\neq m \\ 0 & \text{otherwise} \end{cases}, \qquad (24)$$

and replace $q^*(y \mid m)$ and $q^*(m)$ in (22) by the empirical distributions obtained from the given acoustic and word training sequences, i.e.,

$$q^*(y \mid m) \rightarrow \chi(y-y_T(m)) \qquad (25)$$
$$q^*(m) \rightarrow \frac{1}{J+1}\sum_{j=0}^{J}\chi(m-w_j).$$

This results in

$$P_C = \frac{1}{J+1}\sum_{j=0}^{J} g_{\lambda,\mu}(y_T(w_j),w_j). \qquad (26)$$

In principle, the problem is solved since the task now is to maximize a well defined function over some given domain of $\{\lambda,\mu\}$, and by the physical nature of the problem, a maximum point must exist. Usually, however, the maximization procedure requires knowledge of at least the first order derivative of P_C with respect to $\{\lambda,\mu\}$. In our case $g_{\lambda,\mu}(y,m)$ is not differentiable but can be well approximated by a differentiable function, e.g., the sigmoid function.

The major difference between the newly proposed approach and the approaches discussed in Section 2 is that here the models are optimized for the *given* decision rule while in the other approaches the design of the models is independently done of the decision rule. Furthermore, the optimization here aims directly at the ultimate goal of minimizing the probability of error. The main disadvantage of this approach is that it lacks an efficient implementation. This is, however, also the case for other approaches which have been applied to speech recognition, e.g., the MMI, but which do not guarantee the minimization of the probability of error.

References

[1] L. E. Baum, T. Petrie, G. Soules, and N. Wiess, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.*, vol. 41, pp. 164-171, 1970.

[2] P. F. Brown, "*The Acoustic-Modeling problem in Automatic Speech Recognition*," Ph.D. Thesis, Department of Computer Science, Carnegie Mellon University, 1987.

[3] Y. Ephraim, A. Dembo, and L. R. Rabiner, "A minimum discrimination information approach for hidden Markov modeling," Submitted for publication.

[4] I. Csiszar and G. Tusnady, "Information geometry and alternating minimization procedure," Preprint no. 35, 1983, Math-Inst. Hungar, ACAD SCI., Budapest.

[5] J. E. Shore, "On a relation between maximum likelihood classification and minimum relative-entropy classification," *IEEE Trans. Inform. Theory*, vol. IT-30, No. 6, Nov. 1984.

[6] P. A. Devijver and J. Kittler, *Pattern Recognition-A Statistical Approach*, Printice-Hall International, Inc., London, 1982.