A CHERNOFF–BASED APPROACH TO THE ESTIMATION OF TRANSFORMATION MATRICES FOR BINARY HYPOTHESIS TESTING

F.D. Lorenzo-García[*], A.G. Ravelo-García[†], J.L. Navarro-Mesa[†], S.I. Martín-González[†], P.J. Quintana-Morales[†], E. Hernández-Pérez[†]
Departamento de Ingeniería Telemática[*]. Departamento de Señales y Comunicaciones[†]
*Universidad de Las Palmas de Gran Canaria. Spain. *fdlorenzo@dit.ulpgc.es, †jnavarro@dsc.ulpgc.es*

ABSTRACT

We present a new method for improving the classificacation score in the problem of binary hypothesis testing where the classes are modeled by a Gaussian mixture. We define a cost function which is based on the Chernoff distance and from it a transformation matrix is estimated that maximizes the separation between the classes. Once defined the cost function we derive an iterative method for which we give a simplified version where one mixture component per class is previously selected to participate in the estimation. The initialization of the method is studied and we give two possibilities for this. One is based on the Bhattacharyya distance and the other is based on the average divergence measure. The experiments are carried out over a database of speech with and without pathology and show that our approach represents an improvement in classification scores over other methods also based on matrix transformation.

1. INTRODUCTION

The problem of class separability is well known and has been studied in several papers. The main problem is that the classification scores degrade significantly when the classes are highly confusable. Our main objective is to tackle this problem in order to achieve low confusability in binary hypothesis tests. Various discriminative methods have been proposed in the literature to deal with the problem of confusability. Some methods use an optimization criteria based on mutual information [1] or on minimum classification error [2] to estimate model parameters such that the separation between the competing classes was maximized. In [3] an average divergence measure is used as criterion for finding a transformation matrix which maps the original features into a more discrimitave subspace to improve class separability performance. This and other approaches [4,3,6] use subspace projections that map the feature space into a new subspace by maximizing an appropriately chosen class-separability criterion. We have previously worked in this idea [7] relying on the concept of divergence as a measure of separation between competing states in binary hypothesis where each class is modelled by means of Hidden Markov Models. The paper presented here is not an incremental work based on [7]. Rather it offers a new vision in the design of transformation matrices.

A key idea behind our approach to the design of transformation matrices is the Chernoff distance. This distance can be used to determine an upper bound of the probability of error. Based on this distance, we start by defining a cost function which through its maximization let to address a minimization of the classification error. Thus, maximizing this cost function is proposed as a criterion for estimating the matrix that maps the feature space into a new subspace. A similar idea of using the Chernoff distance as a criterion is presented in [8] where the distance is defined using the notion of directed distance matrices.

This matrix is automatically trained using a cost function based on the Chernoff distance. A selected set of Gaussians from each class participates in the process and the transformation matrix is expected to separate the Gaussians in the new subspace. For this purpose we develop an iterative method based on the steepest ascent method that aims at finding the maximum separation between the Gaussians that represent the classes. In iterative methods the initialization is of major importance because affects the evolution to a global or a local maximum. Two different initializations are studied that use a definition of distance from the mean vector and the covariance matrix. One definition aims at maximizing the Bhattacharyya distance and the other aims at maximizing an average divergence.

Once defined the cost function and the maximization method we center our experiments in pathological speech classification. The vectors in the original space are formed by Mel-warped log-filterbank energies (MFE) features. Hence, the aim of the experiments is to obtain a transformation matrix to a new space where speech with and without pathology can be better separated than with other methods. A reference transformation matrix is the one based in the discrete cosinus transform which is classic in the speech recognition literature.

A set of experiments are carried out over a pathological speech database. A variety of methods to estimate transformation matrices is applied for comparison purposes

and the results show that our method shows promising results that outperform the other methods.

2. DEFINITION OF A COST FUNCTION BASED ON THE CHERNOFF DISTANCE

The Chernoff distance is a measure of similarity between two probability density functions (pdf). For example, each pdf may define the probability of pertaining to a given class, and therefore signifies how similar or different the two classes are. This distance can be defined as follows [5]:

$$\overline{D} = \max_{0\le\alpha\le 1}\left\{-Ln\left(\int P_1^{\alpha} p_1^{\alpha}(X) P_2^{1-\alpha} p_2^{1-\alpha}(X)dx\right)\right\} \qquad (2.1)$$

where $p_i(X)= p(X/w_i)$ $\{i=1,2\}$ and $P_i=P(w_i)$ are the conditional probability density and the a priori probability of class *'i'*, respectively. Obviating the maximization in (2.1) with respect to α and without loss of generality it is assumed to be constant hereafter and equal to ½ in the experiments. A maximum similarity between densities is seen as a tendency of $\overline{D}$ to 0 while a minimum similarity is seen as tendency of $\overline{D}$ to infinity. Note that $\overline{D}$ = 0 when $p_i(X) = p_j(X)$ or in the trivial cases $\alpha=\{0,1\}$, and $\overline{D}$ = ∞ when the two classes are absolutely separable and their pdf's do not overlap. The important fact for our purposes is that the larger the distance $\overline{D}$ is between the two distributions, the smaller the probability of misclassification between classes. In fact, the expression in (2.1) is often used for obtaining an upper bound on the probability of misclassification such that the bigger the distance the smaller that probability. Thus, maximizing the Chernoff distance is used here as a key idea that will be further used for finding a transformation matrix from an original space to a transformed one where the classes are maximally separated.

Now let $Y=\{(x^1,y^1),\ldots,(x^N,y^N)\}$ be a finite set of training instances, where each instance x^t corresponds to a label $y^t=\{1,2\}$. Let *A (k x m)* be a linear matrix which maps an original observation x^t into a transformed one as $v^t = A^T x^t$, where x^t is a *k*-dimensional vector, v^t is an m-dimensional vector and $m\le k$. The convex nature of (2.1) due to the application *–Ln[.]* over the integral is not appropriate for obtaining the transformation matrix *'A'* that maximizes the separation between classes. Instead, we propose a cost function based the Chernoff distance as follows:

$$D = \max_{\alpha,A}\tanh\left\{-s\cdot Ln\left[\sum_{t=1}^{N} P_1^{\alpha} p_1^{\alpha}(v^t) P_2^{1-\alpha} p_2^{1-\alpha}(v^t)\right]\right\} \qquad (2.3)$$

where the integral has been substituted by a summation because the training set is finite, the hyperbolic tangent has been introduced for convenience and $s>0$ is a constant factor that controls the dynamic range of the argument. Notice that now when the pdf's of the two classes tend to overlap both $\overline{D}$ and D tend to zero but when the overlap decreases then D tends to one. The resulting function is concave increasing and facilitates the search for minimum overlapping (confusability) between classes by searching for the maximum of (2.3). This search will be done with a steepest ascent method that is presented in the next section.

3. MAXIMITATION OF THE COST FUNCTION

Let's particularize for the case in which each class $j=\{1,2\}$ probability $p_j(X)$ is characterized by a mixture of *M* Gaussian components with means μ_j^i, covariance matrixes Σ_j^i, weighting factors ω_j^i and $\{i=1,\ldots,M\}$. The objective is to obtain the matrix *A* such that the cost function is maximized since it is associated to a minimum classification error. To do that we start by taking the partial derivatives of (2.3) with respect to *A*. Applying the chain rule the derivative first becomes

$$\frac{\partial D}{\partial A} = s\left[1-\tanh^2\left(-s\cdot Ln\left(\sum_{t=1}^{N} p_1^{\alpha}(v^t) p_2^{1-\alpha}(v^t)\right)\right)\right]\left[\frac{\partial}{\partial A}\left(-Ln\left(\sum_{t=1}^{N} p_1^{\alpha}(v^t) p_2^{1-\alpha}(v^t)\right)\right)\right] \qquad (3.1)$$

Now let's develop more in detail the partial derivative with respect to *A* in the right hand side of (3.1). Taking again the chain rule and derivating the logarithm in this way $\partial Ln[g(x)]/\partial x = g'(x)\,\partial Ln[g(x)]/\partial[g(x)]$, the second factor in square brackets of the right hand side of equation (3.1) can be made equal to

$$\frac{\partial}{\partial A}\left(-Ln\left(\sum_{t=1}^{N} p_1^{\alpha}(v^t) p_2^{1-\alpha}(v^t)\right)\right) = \sum_{t=1}^{N}\left[\frac{\partial/\partial A\left(p_1^{\alpha}(v^t) p_2^{1-\alpha}(v^t)\right)}{\sum_{t=1}^{N} p_1^{\alpha}(v^t) p_2^{1-\alpha}(v^t)}\right] \qquad (3.2)$$

where the denominator in the right hand side of equation (3.2) acts as a normalization weight of the contribution to the derivative from each training vector x^t. The numerator has the partial derivatives of both mixtures with respect to matrix *'A'*. Since these derivatives include several terms and some of them are numerically negligible we will make the simplification that only the most important component from each mixture is taken into consideration. Thus, only single mixture components from each class are considered in (3.2).

The expression of the single component *'i'* from class *'j'* in the transformed space can be expressed as:

$$p_j^i(x^t, A) = \frac{1}{(2\pi)^{m/2}\left|A^T\Sigma_j^i A\right|^{1/2}}\exp\left\{-\frac{1}{2}\left(x^t-\mu_j^i\right)^T A\left(A^T\Sigma_j^i A\right)^{-1}A^T\left(x^t-\mu_j^i\right)\right\} \qquad (3.3)$$

For example, in the particular case of class $j=1$, the partial derivative with respect to the transformation matrix '*A*' of a given component *'i'* in (3.2) is:

$$\frac{\partial {p_1^i}^{\alpha}(A)}{\partial A} = {p_1^i}^{\alpha}(A)\left[\alpha \cdot B_1^i A - \alpha \cdot \Sigma_1^i A\left(A^T \Sigma_1^i A\right)^{-1}\left(A^T B_1^i A\right) + \Sigma_1^i A\left(A^T \Sigma_1^i A\right)^{-1}\right] \quad (3.4)$$

where $B_j^i = A^T\left(x^t - \mu_j^i\right)\left(x^t - \mu_j^i\right)^T A$. Note that (3.4) is estimated for each training vector x^t and therefore each one of them will directly influence the estimation of A.

Using equation (3.4) and making some mathematical manipulations, the derivative of the product in (3.2) and rearranging the results from the derivatives, equation (3.2) can be rewritten as

$$\frac{\partial}{\partial A}\left(-Ln\left(\sum_{t=1}^{N} p_i^{\alpha}(v^t) p_j^{1-\alpha}(v^t)\right)\right) = \sum_{t=1}^{N}\left[\frac{p_1^{\alpha}\left(v^t\right) p_2^{1-\alpha}\left(v^t\right)}{\sum_{t=1}^{N} p_1^{\alpha}\left(v^t\right) p_2^{1-\alpha}\left(v^t\right)}\right] * \cdots \quad (3.5)$$

$$\left[\begin{array}{l}\left(\alpha \cdot B_1^i A - \alpha \cdot \Sigma_1^i A\left(A^T \Sigma_1^i A\right)^{-1}\left(A^T B_1^i A\right) + \Sigma_1^i A\left(A^T \Sigma_1^i A\right)^{-1}\right) \\ + \left((1-\alpha) \cdot B_2^l A - (1-\alpha)\Sigma_2^l A\left(A^T \Sigma_2^l A\right)^{-1}\left(A^T B_2^l A\right) + \Sigma_2^l A\left(A^T \Sigma_2^l A\right)^{-1}\right)\end{array}\right]$$

where components *'i'* and *'l'* have been taken from class *1* and *2*, respectively.

The transformation matrix could be obtained by solving equation (3.1) equating to cero, which is difficult. Instead, we will apply a steepest ascent method for finding matrix A, i.e., given an initial matrix A^0 and update matrix $A^{(l)}$ in the following manner:

$$A^{(l+1)} = A^{(l)} + \gamma_l \left.\frac{\partial D}{\partial A}\right|_{A=A^{(l)}} \quad (3.6)$$

for $\{l=1,2,..., N_l\}$ where *'l'* is the iteration index, N_l is the number of iterations, $\gamma_l = \gamma_0 (1-l/N_l)$ is a step size that depends on the step and γ_0 is a small initialization constant that has been set to 0'1 in the experiments. We call our method Iterative Chernoff Maximization (ICM).

Now we can make some considerations. The factor $(1 - \tanh^2[.])$ in (3.1) plays an interesting role because it is a reflection of the way in which the cost function progresses to its maximum when a new iteration is performed. Thus, this factor tends to zero making the derivative smaller step by step. The product between the smoothing factor *'s'* and the step size γ_l act as a prevention from a fast progress of (2.1) to the maximum.

The way in which the initialization matrix A^0 is estimated is open. In this paper we will study two possible initializations. In the first, we use the Bhattacharyya distance as defined in [6, 5] which is coherent with the Chernoff distance when $\alpha=1/2$. Then, A^0 can be estimated as folllows. Let $U = [u_1,...,u_n]$ be the eigenvector matrix, and let $\Lambda=diag(\lambda_1,..., \lambda_n)$ be the eigenvalue matrix of $\Sigma_2^{-1}\Sigma_1$ where the super indexes identifying the components from each class mixture have been omitted. Hence in the original space the Bhattacharyya can be written as

$$B(1,2) = \sum_{i=1}^{n}\left[\frac{1}{4}\frac{\left\{u_i^T(\mu_2 - \mu_1)\right\}^2}{1+\lambda_i} + \frac{1}{4}\left\{\ln\left(\lambda_i + \frac{1}{\lambda_i} + 2\right) - \ln 4\right\}\right] \quad (3.7)$$

Since we want to maximize the B(1,2) distance, the columns of the matrix $A^{(0)}$ can be initialized with the m eigenvector of $\Sigma_2^{-1}\Sigma_1$ corresponding to de m largest [.] terms in (3.7).

And in the second initialization, we use the transformation matrix that maximizes the average divergence measure (ADM) as defined in [3]. The matrix A that maximizes the average divergence can be formed by selecting the m eigenvectors of the matrix $(V^{-1}*M)$ corresponding to the m largest eigenvalues, where V is a common covariance matrix among all classes and M is a matrix that can be defined as follows

$$M = \sum_{i=1}^{K}\sum_{j=1}^{K} P_i P_j \left(\mu_i - \mu_j\right)\left(\mu_i - \mu_j\right)^T \quad (3.8)$$

4. EXPERIMENTS AND RESULTS

For the classification experiments we used speech with and without pathology. The database is composed of 54 speakers without pathology and 608 speakers with pathology from the Disordered Voice Database Model 4337. Recordings consisted in sustained vowel /ah/. A half of the database was used for training and a half for testing. The original features were MFE obtained by applying m=20 triangular filters to the magnitude spectrum and the dimensionality in the transformed space is k=10. The speech waveforms are sampled at 25 KHz, and are blocked into 1500 samples from 30 msec. frames with 20 msec. of overlap between adjacent blocks. Each frame is passed through pre-emphasis filter and a Hamming window. Then, a 2048 points FFT is applied to the frame to produce a 1024-point power spectrum. The power spectra are combined using a weighted sum, shaped by the triangular filter, to obtain the filter output. Logarithms of the 20 outputs are then calculated arriving at 20 MFE for each frame. The whole set of training vectors is used to characterize each class by a mixture of Gaussian probability densities with M=4 or 6 components. For comparison purposes we have made experiments in the original space (MFE) and in the transformed space with MFCC, ADM and ICM with the two initializations introduced in section 3. We have experimentally found that a good choice of the control factor in (2.3) is s=1/32. Note that the sub index in ICM stems from the method for initializing the transformation matrix where sub index B indicates Bhattacharyya method and L indicates the transformation matrix obtained from the ADM method. In table 1 we show the classification scores for the four methods mentioned in the previous sections. As we can see, both versions the method we propose, ICM, outperform the MFE, MFCC and ADM in all experiments.

It is interesting to observe that the MFCC transformation does not improve the classification scores in comparison with MFE in contrast to what one could expect from the literature. The scores obtained from the other transformations are really improved. Both $ICM_{(B)}$ and $ICM_{(L)}$ give the best results.

Method	M	Scores
MFE	*6*	*85.49%*
MFCC	*6*	*84.29%*
ADM	*6*	*92.45%*
$ICM_{(B)}$	*6*	*93.05%*
$ICM_{(L)}$	*6*	*94.86%*
MFE	*4*	*84.59%*
MFCC	*4*	*78.85%*
ADM	*4*	*92.75%*
$ICM_{(B)}$	*4*	*94.26%*
$ICM_{(L)}$	*4*	*93.05%*

TABLE 1. Classification scores for the different methods

Since the amount of examples for the non pathologic speech is small, we show in tables 2 and 3 the confusion matrixes of the different methods between pathologic (PS) and normal (NS) speech.

In table 2 the number of component per mixture is M = 6 while in table 3 the number of component per mixture is M=4. From the table we can see that again de ICM method offer the lowest confusion score even for the NS class that has such a small amount of examples compared with the PS class.

Method	Pathology	NS	PS
MFE	*NS*	*77.78%*	*22.22%*
MFE	*PS*	*13.82%*	*86.18%*
MFCC	*NS*	*88.89%*	*11.11%*
MFCC	*PS*	*16.12%*	*83.88%*
ADM	*NS*	*44.44%*	*55.56%*
ADM	*PS*	*3.29%*	*96.71%*
$ICM_{(B)}$	*NS*	*81.48%*	*18.52%*
$ICM_{(B)}$	*PS*	*5.92%*	*94.08%*
$ICM_{(L)}$	*NS*	*85.19%*	*14.81%*
$ICM_{(L)}$	*PS*	*4.28*	*95.72%*

TABLE 2. Confusion Matrixes for the different methods with M=6

Method	Pathology	NS	PS
MFE	*NS*	*81.48%*	*18.52%*
MFE	*PS*	*15.13%*	*84.87%*
MFCC	*NS*	*88.88%*	*11.12%*
MFCC	*PS*	*22.04%*	*77.96%*
ADM	*NS*	*81.48%*	*18.52%*
ADM	*PS*	*6.25%*	*93.75%*
$ICM_{(B)}$	*NS*	*85.18%*	*14.82%*
$ICM_{(B)}$	*PS*	*4.93%*	*95.07%*
$ICM_{(L)}$	*NS*	*81.48%*	*18.52%*
$ICM_{(L)}$	*PS*	*5.92%*	*94.08%*

TABLE 3. Confusion Matrixes for the different methods with M=4

5. CONCLUSIONS

Up to our knowledge no previous work has addressed the problem of class separability by maximizing the Chernoff distance through a matrix transformation in the same way we do in this paper. We have formulated a cost function based on this distance and we have given the formulae for an iterative maximization. The simplified version presented here is computationally less demanding than the original one while give good classification scores. Obviating simplifications, from a theoretical point of view, in this paper we open the scope of possibilities to obtain transformation matrices, but also all the parameters that parameterize the mixture components of each class. This and other related subjects will be a matter of future work in binary and M-ary hypothesis testing with a special emphasis in extending the formulation to hidden Markov models.

6. REFERENCES

[1] L.R. Bahl, P.F. Brown, P.V. Souza, & R.L. Mercer, "Maximum mutual information estimation of hidden Markov models for speech recognition". Proc ICASSP'86, pp 49-52, 1986.

[2] P.C. Chang & B-H. Juang, "Discriminative training of dynamic programming based speech recognizers". IEEE Trans. Speech Audio Processing, Volume 1, number 2, pp 135-143, 1993.

[3] P.C. Loizou, & A.S. Spanias, "Improved speech recognition using a subspace projection approach". IEEE Transactions on Speech and Audio Processing, Volume: 7, Issue: 3, pp 343–345, 1999.

[4] R. Haeb-Umbach & H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition". Proc ICASSP'92, pp 13-16, 1992.

[5] Fukunaga, K, "*Introduction to Statistical Pattern Recognition*". Academic Press, Inc. 2nd Edition, 1990.

[6] P.C. Loizou, & A.S. Spanias, "High-performance alphabet recognition". IEEE Transactions on Speech and Audio Processing, Volume: 4, pp 430-445, 1996.

[7] F.D. Lorenzo-García & J.L.Navarro-Mesa, "Improved Binary Hypothesis Classification Using a Matrix Transformation Approach". Sixth IMA-IEE International Conference on Mathematics in Signal Processing, pp 147-150, 2004.

[8] M Loog & R. Duin. "Linear Dimensionality Reduction via Heterocedastic Extension of LDA: The Chernoff Criterion". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 38, Nº 6, June 2004, pp 732-739.