PERFORMANCE EVALUATION OF AUTOMATIC SPEAKER RECOGNITION SCHEMES

D. Venugopal and V.V.S. Sarma

Indian Institute of Science
Bangalore 560012

ABSTRACT

A mathematical formulation of an automatic speaker verification scheme as a two class pattern recognition problem is presented. Expressions for the expected values and the variance of the design-set and the test set error rates are derived. The bound on the performance of an automatic speaker identification system as a cascade of independent verification systems is derived. The implications of these results in the design of an automatic sp eaker recognition system are discussed.

I INTRODUCTION

The problem of performance evaluation of any automatic speaker verification system (ASVS) is yet to be satisfactorily solved. In general pattern recognition literature, the performance estimation has received considerable notice recenty and the importance of these results in the design of an ASVS is discussed in a recent note of the authors[1].

In this paper, an ASVS is analysed as a 2-class pattern recognition problem to bring out explicitly the effect of the various parameters on the performance estimation following the mathematical model of a pattern verification system provided by Dixon[2]. The expected error rates for both the design-set and the test-set are derived as a function of the number of samples per speaker (N), the number of features (L), the number of customers (M), the number of design impostors (K) and the Mahalanobis distance (Δ) between the classes under Gaussian assumptions. The variance of the design-set error rate is derived bringing out the importance of choosing sufficient number of impostors at the system design stage. The expected error rates for an automatic speaker identification system (ASIS) as a cascade of M independent ASVS are derived. The importance of these results in the design of an automatic speaker recognition system is pointed out.

II MATHEMATICAL MODEL FOR ASVS

In an ASVS, a given speaker not necessarily belonging to the system, utters a predetermined phrase. He also presents a label claiming that he is a particular "customer" belonging to the system. In the system, a predetermined set of features, possibly depending on the label entered, is extracted from the utterance and the speaker is either accepted or rejected.

Let the ASVS be designed for a set of M known customers $S = \{S_i,\ i=1,\ldots,M\}$ and a single alien group, S_o of all unknown speakers (rest of the world!). This inclusion of an alien group S_o of speakers distinguishes an ASVS from an ASIS and is essential as there is always a chance of a person not belonging to the set S trying to impersonate as one of S. Corresponding to each member of S, there is a label L_j ($j=1,\ldots,M$). If any speaker wants to be verified under the label L_j, then we define two classes: the C_j class consisting of the speaker S_j and the imposter class $\bar{C}_j$ consisting of the remaining (M-1) speakers and the alien group S_o.

$$C_j = \{S_j\};\quad \bar{C}_j = \{S_o, S_i,\ i=1,\ldots,M,\ i\neq j\},\quad j=1,\ldots,M.$$

The system on the basis of the feature vector X of dimension L will assign the speakers to C_j or $\bar{C}_j$. The accept/reject rule using the optimal Bayesian classifier is to to accept the claim of a particular speaker as valid if

$$p(C_j/L_j,X) > p(\bar{C}_j/L_j,X) \qquad (1)$$

and reject it otherwise.

The aposteriori probabilities are given by

$$p(C_j/L_j,X) = p(X/L_j,C_j)\,p(L_j/C_j)\,p(C_j)\ /p(L_j,X)$$

$$p(\bar{C}_j/L_j,X) = p(X/L_j,\bar{C}_j)\,p(L_j/\bar{C}_j)\,p(\bar{C}_j)\ /p(L_j,X) \qquad (2b)$$

where the probabilities have the usual meanings and $p(L_j/C_j)$ is the probability of speaker S_j wanting to be verified under his own label L_j and $p(L_j/\bar{C}_j)$ is the probability of an "impostor" trying to present label L_j. While the actual values of the various probabilities of (2a) and (2b) depend upon the conditions in a particular environment, we may assume in most cases

$p(L_j/C_j) \gg p(L_j/\bar{C}_j)$ and $p(\bar{C}_j) \gg p(C_j)$, assuming equal a priori probabilities for all speakers. A reasonable assumption considering the above inequalities is $p(L_j/C_j)p(C_j) = p(L_j/\bar{C}_j)p(\bar{C}_j)$. Again, it may be postulated that in the presence of the knowledge of class C_j the feature vector is independent of the label L_j, $p(X/L_j,C_j)=p(X/C_j)$. The decision rule (1) can now be rewritten as

$$\text{Accept if } p(X/L_j,C_j) > p(X/L_j,\bar{C}_j) \qquad (3)$$

and reject otherwise.

The class-conditional densities in (3) are given by

$$p(X/C_j)=p(X/S_j) \quad j = 1,\ldots,M \qquad (4)$$

$$p(X/\bar{C}_j)= \sum_{\substack{i=0 \\ i\neq j}}^{M} p(X/S_i)p(S_i)$$

where $p(X/S_i)$, $i=0,1,\ldots,M$ are speaker-conditional densities of the feature vector X.

If the distributions $p(X/S_i)$, $i=0,..M$ are completely known, then the error rate can be calculated exactly. Otherwise $p(X/S_i)$ $i=1,\ldots,M$ are to be estimated from N labelled samples of each speaker of set S where as $p(X/S_o)$ is to be estimated from a set of "impostor references" (of speakers of S_o). Just as the number of training samples per class is finite, the number of impostors (K) that can be considered to represent the group S_o is also finite.

Nature of the Class-Conditional Densities:

Assumption: $p(X/C_j) \sim N(\mu_j,\Sigma_j)$ and $p(X/\bar{C}_j) \sim N(\bar{\mu}_j, \bar{\Sigma}_j)$, where $\Sigma_j=\Sigma, \bar{\Sigma}_j=\beta\Sigma$ and $\beta=M+K-1$ are the known covariance matrices of the two classes C_j and $\bar{C}_j$ and the means μ_j and $\bar{\mu}_j$ are to be estimated from N design samples of class C_j and βN design samples of class $\bar{C}_j$.

Remark: It may not be unreasonable to assume $p(X/S_i)=p(X/C_j)$ to be Gaussian. Then $p(X/\bar{C}_j)$ is a finite mixture of Gaussians. This will be a multimodal distribution for small M and K. Again it may not be unreasonable, for large M and K, to fit a Gaussian distribution to samples of class C_j.

Classifier : For notational convenience, we denote C_j by C_1 and $\bar{C}_j$ by C_2 and $p(X/C_1) \sim N(\mu_1,\Sigma)$ and $p(X/C_2) \sim N(\mu_2,\beta\Sigma)$. For this case of unequal means and covariance matrices the minimum probability of error classifier is the one using quadratic discriminant function. In this paper, the minimax linear discriminant with equal error rate is used for further analysis[3]. We define a linear discriminant

$$d = (\Sigma^*)^{-1}(\mu_1^*-\mu_2^*) \qquad (5)$$

where $\Sigma^*= t+(1-t)\Sigma$ where t 1 is a parameter of our choice and $\mu_1^* =(1/N)\ \Sigma_{j=1}^{N} X_{ij}$ and $u_2^* = (1/\beta N)\Sigma_{j=1}^{\beta N} X_{2j}$ where X_{ij} is the jth labelled sample of class i. For equal error rate, the threshold θ is given by

$$\theta=d'(\beta^{1/2}\mu_1^* + \mu_2^*)/(\beta^{1/2}+1) \qquad (6)$$

Therefore, a sample X is assigned to class C_1 if $d'X \geq \theta$, i.e.

$$(\mu_1^*-\mu_2^*)'(\Sigma^*)^{-1}\{X-(\beta^{1/2}\mu_1^*+\mu_2^*)/(\beta^{1/2}+1)\} \geq 0 \qquad (7)$$

and to class C_2, otherwise.

III. TEST AND DESIGN-SET ERROR RATES FOR AN ASVS

The ASVS may be tested either by new utterances of the speakers belonging to S and S_o (test set) or by the sample utterances used for designing the system (design set).

Test-set error rate: The expected test-set error rate ($\bar{\varepsilon}_T$) may be written from (7) as

$$\bar{\varepsilon}_T=pr\left[(\mu_2^*-\mu_1^*)'(\Sigma^*)^{-1}\left\{X-\frac{\beta^{1/2}\mu_1^*+\mu_2^*}{\beta^{1/2}+1}\right\}>0\right] \qquad (8)$$

where X is the feature vector corresponding to an arbitrary new utterance from class C_1.

Proposition 1 : $\bar{\varepsilon}_T$ in (8) may be expressed as the probability of the ratio of two non-central χ^2 variates ω_1 and ω_2 being greater than the quantity $(1-\rho_1)/(1+\rho_1)$.

$$\bar{\varepsilon}_T=pr\left[\omega_1/\omega_2 > (1-\rho_1)/(1+\rho_1)\right] \qquad (9)$$

where ω_1 and ω_2 are distributed as $\chi^2(L,\lambda_1)$ and $\chi^2(L,\lambda_2)$.

$$\lambda_1=[2(1+\rho_1)]^{-1}\beta N\left\{(\beta+1)^{-\frac{1}{2}}-[1+\beta^2+(\beta^{\frac{1}{2}}+1)^2N]^{-\frac{1}{2}}\right\}^2\Delta^2$$

$$\lambda_2=[2(1-\rho_1)]^{-1}\beta N\left\{(\beta+1)^{-\frac{1}{2}}+[1+\beta^2+(\beta^{\frac{1}{2}}+1)^2N]^{-\frac{1}{2}}\right\}^2\Delta^2$$

Δ^2 = Mahalanobis squared distance between the two populations = $(\mu_2^*-\mu_1^*)'(\Sigma^*)^{-1}(\mu_2^*-\mu_1^*)$

$\rho_1 = (\beta^{3/2}-1)\{(\beta+1)(1+\beta^2+(\beta^{\frac{1}{2}}+1)^2N)\}^{-\frac{1}{2}}$.

Proof : (The proof follows that of Moron[4]. We define two random vectors u and v such that

$$u=(\beta N/\beta+1)^{\frac{1}{2}}(\Sigma^*)^{-\frac{1}{2}}(\mu_2^*-\mu_1^*) \text{ and}$$

$$v=\frac{\beta N}{1+\beta^2+\beta(\beta^{\frac{1}{2}}+1)^2N)^{\frac{1}{2}}}(\beta^{\frac{1}{2}}+1)(\Sigma^*)^{-\frac{1}{2}}\left\{x-\frac{\beta^{\frac{1}{2}}u_1^*+u_2^*}{\beta^{\frac{1}{2}}+1}\right\}$$

where $(\Sigma^*)^{\frac{1}{2}}(\Sigma^*)^{\frac{1}{2}}=(\Sigma^*)$.

Then $\bar{\varepsilon}_T$ can be written as

$$\bar{\varepsilon}_T=pr(u'v>0)=pr\left\{(u+v)'(u+v)-(u-v)'(u-v)>0\right\}$$

(u+v) and (u-v) are distributed independently with dispersion matrices $2(1+\rho_1)I_L$ and $2(1-\rho_1)I_L$, where ρ_1 is the correlation coefficient between corresponding pair of elements of u and v and I_L is the LXL unit matrix. Thus if, $\omega_1=(\frac{1}{2})(u+v)'(u+v)/(1+\rho_1)$ and $\omega_2=\frac{1}{2}(u-v)'(u-v)/(1-\rho_1)$, eqn.(9) follows. Detailed proof is given in reference 5.

Proposition 2 : It is also possible to express the expected error rate $\bar{\varepsilon}_T$ in closed form expression[4].

$$_T = Q(L,\lambda_1,\lambda_2,\rho_1), \text{ where} \qquad (10)$$

$$Q(L,\lambda_1,\lambda_2,\rho_1)=1-C(\theta_1^{\frac{1}{2}},\theta_2^{\frac{1}{2}})-\exp(-\frac{\theta_1+\theta_2}{2})x \sum_{m=1-\frac{1}{2}L}^{\frac{1}{2}L-1}(\theta_1/\theta_2)^{\frac{1}{2}m} I_m(\theta_1\theta_2)^{\frac{1}{2}}\left\{B_{\mp\frac{1}{2}(1\mp\rho_1)}(\tfrac{1}{2}L+m,\tfrac{1}{2}L-m)-\delta_{m0}\right\} \quad (11)$$

where $\theta_1=\frac{1}{2}(1+\rho_1)\lambda_1$, $\theta_2=\frac{1}{2}(1-\rho_1)\lambda_2$, $C(\theta_1^{\frac{1}{2}},\theta_2^{\frac{1}{2}})$ is the circular coverage function, $I_m(z)$ the modified Bessel function of first kind $B_x(p,q)$ the incomplete β-function. The upper sign is for $m \geq 0$ and the lower sign for $m<0$.

Design-set error rate : The expected design set error rate ($\bar{\varepsilon}_D$) may be written from eqn.(7) as

$$\bar{\varepsilon}_D=pr\left[(\mu_2^*-\mu_1^*)(\Sigma^*)^{-1}\left\{X_{ij}-\frac{\beta^{\frac{1}{2}}\mu_1^*+\mu_2^*}{\beta^{\frac{1}{2}}+1}\right\}>0\right] \quad (12)$$

where X_{ij} is the feature vector corresponding to an arbitrary utterance from the design set of class C_1.

Proposition 3: $\bar{\varepsilon}_D$ in (12) may be expressed as the probability of the ratio of two non-central χ^2 variates ω_3 and ω_4 being greater than the quantity $(1-\rho_2)/(1+\rho_2)$

$$\bar{\varepsilon}_D=pr\left[\omega_3/\omega_4 > (1-\rho_2)/(1+\rho_2)\right] \quad (13)$$

where ω_3 and ω_4 are distributed as $\chi^2(L,\lambda_3)$ and $\chi^2(L,\lambda_4)$

$$\lambda_3=\left[2(1+\rho_2)\right]^{-1}\beta N\left\{(\beta+1)^{-\frac{1}{2}}-\left[1-\beta^{3/2}(\beta^{\frac{1}{2}}+2)+\beta(\beta^{\frac{1}{2}}+1)^2N\right]^{-\frac{1}{2}}\right\}^2\Delta^2$$

$$\lambda_4=\left[2(1-\rho_2)\right]^{-1}\beta N\left\{(\beta+1)^{-\frac{1}{2}}-\left[1-\beta^{3/2}(\beta^{\frac{1}{2}}+2)+\beta(\beta^{\frac{1}{2}}+1)^2N\right]^{-\frac{1}{2}}\right\}^2\Delta^2$$

$$2=-(\beta+1)^{\frac{1}{2}}\left\{1-\beta^{3/2}(\beta^{3/2}+2)+(\beta^{\frac{1}{2}}+1)^2N\right\}^{-\frac{1}{2}} .$$

Proof : The proof is similar to that of proposition 1.

Proposition 4: It is also possible to express the expected error rate $\bar{\varepsilon}_D$ in a closed form expression[4].

$$\bar{\varepsilon}_D = Q(L,\lambda_3,\lambda_4,\rho_2), \text{ where} \quad (14)$$

$Q(L,\lambda_3,\lambda_4,\rho_2)$ is the same function as defined in (10) with λ_1 and λ_2 and ρ_1 being replaced by λ_3, λ_4 and ρ_2 respectively.

Proposition 5: The variance σ^2 of the random variable $\bar{\varepsilon}_D$ is given by

$$\sigma^2 = \bar{\varepsilon}_D(1-\bar{\varepsilon}_D)/(\beta+1)N \quad (15)$$

The proof follows that of Foley[6] and is given in reference 5.

In Fig. 1 and 2 the values of $\bar{\varepsilon}_D$ and $\bar{\varepsilon}_T$ are plotted as functions of N/L and β. Fig. 1 gives the nature of the biases that creep into the estimates for small N/L. The test-set error rate is pessimistically biased and the design-set error rate is optimistically biased. Fig. 2 shows that for an ASVS the expected error rates become independent of population size for large population. It may be seen from (15) that the variance of the design-set error rate is inversely proportional to the number of recorded sample utterances for speaker and the total number of speakers including the design impostors. Fig. 2 and (15) show that for small number of customers (M) if sufficiently large number of design impostors (K) are not used, both $\bar{\varepsilon}_D$ and $\bar{\varepsilon}_T$ are optimistically biased and are not reliable. However, for large M, a large K is not essential.

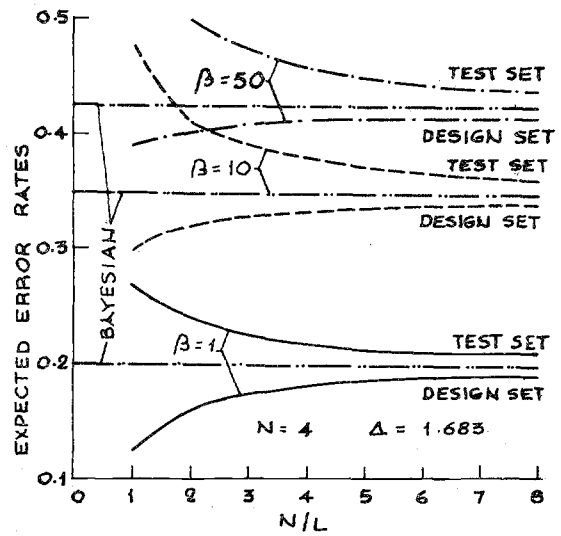

Fig.1-Expected Error Rates as Function of N/L

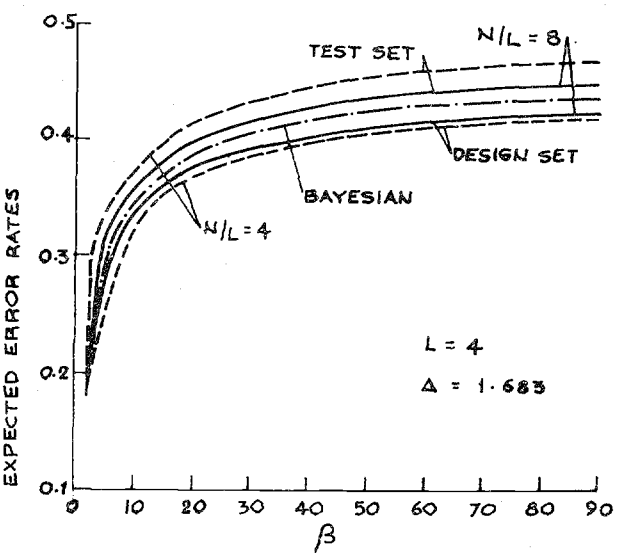

Fig.2-Expected Error Rates as Functions of Population Size

IV PERFORMANCE EVALUATION OF ASIS

As ASIS can be realized as a cascade of (M-1) or M ASVS's as shown in Fig.3. All the ASVS's are assumed to have identifical performance. If there is a reject option at Mth stage also the possibility of (M+1) classes corresponding to M customers and an alien class (as not belonging to the system) can be introduced in an ASIS as well. On the other hand, the decision can be terminated at (M-1)th stage and speaker S_M can be accepted.

Let p be the probability of error and q the probability of correct decision of an ASVS. Assuming that the jth speaker has tested the system, we can draw the decision tree as shown in Fig.4. If D_i, i = 1,..,M is the decision taken by the system at the ith stage that the speaker is S_i, then we can write the probability of correct decision as

$$P_C = \sum_{j=1}^{M} P(S_j, D_j) = \sum_{j=1}^{M} P(D_j/S_j)P(S_j) \qquad (16)$$

Assuming equal apriori probabilities for all speakers belonging to S and from Fig. 4, we can write

$$P_C = (1/M)\left(\sum_{j=1}^{M} q^j\right) = (q/M)(1-q^M)/(1-q) \qquad (17)$$

Equation (19) shows the effect of population size on the performance of an ASIS, thus corroborating Doddington's results [7].

V DISCUSSION OF RESULTS

The design of an ASVS proceeds in three steps: (i) Data base preparation, (ii) feature selection and extraction and (iii) statistical classification and performance evaluation. All the stages are, of course, interrelated. The results of section III provide information on: (i) Preparation of data set (number of design sample utterances per speaker (N)), (ii) The dimension of the feature vector (L). If N/L ratio is small there will be wide disparities in performance estimates that will be obtained if the system is tested on the design set or on an independent test set. (iii) The discriminating ability of a feature depends on the appropriate distance between the classes concerned. If the underlying distributions are Gaussian the distance between classes itself provides an estimate of error. It should be kept in mind, however, that the distance estimate from a finite number of samples per class is a biased estimate of the true distance between the populations, (iv) The error rates ε_D and ε_T are functions of β (Fig.2). Even if an ASVS is to be designed for a small number of customers M, a sufficiently large number K of impostors should be considered in the design set so as to make the performance estimates reliable. For large M this may not be so important. The design of a verification system is thus equally complex for small or large M because of the presence of the alien class.

REFERENCES

1. V.V.S. Sarma and D. Venugopal, "Performance evaluation of automatic speaker verification systems", IEEE Trans. Acoustics, Speech, Signal Processing (to be published).

2. R.C.Dixon and P.E. Bourdeau, "Mathematical model for pattern verification", IBM J. of R and D, Vol. 13, pp 717-721, Nov. 1969.

3. T.W.Anderson and R.R.Bahadur, "Classification into two multivariate normal distributions with different covariance matrices", Ann. of Math. Statist., Vol.33 pp 420-431, June 1962.

4. M.A.Moron, "On the expectation of errors of allocation associated with a linear discriminant function", Biometrika, Vol. 62, pp 141-148, Apr. 1975.

5. V.V.S.Sarma and D. Venugopal, "Statistical problems in performance assessment of Automatic Speaker Recognition Systems" CIP Report No.61, Dept. of ECE, Indian Inst. of Science, India, Jan. 1977.

6. D.H. Foley, "Considerations of sample and feature size", IEEE Trans. Information Theory, Vol.IT-18, pp 618-626, Sept.1972.

7. A.E.Rosenberg, "Automatic speaker verification: a review", Proceedings IEEE, Vol.64, pp 475-487, Apr. 1976.

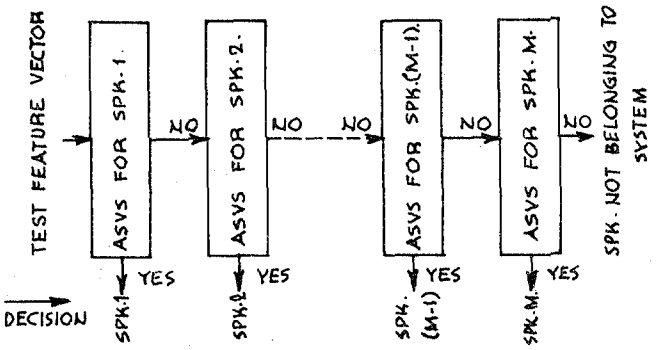

Fig.3 - ASIS as a Cascade of ASVS

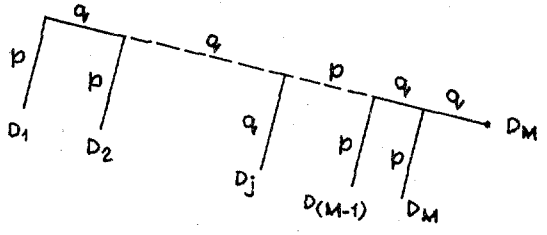

Fig.4 - Decision Tree for Speaker j