ON LINEAR DETs

Jiří Navrátil [1] *and David Klusáček* [2]

[1]IBM T.J. Watson Research Center, Human Language Technologies, Yorktown Heights, NY 10598, USA

[2]Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

e-mail: `jiri@us.ibm.com, klusacek@atrey.karlin.mff.cuni.cz`

ABSTRACT

This paper investigates the properties of a popular ROC variant - the Detection Error Trade-Off plot (DET). In particular, we derive a set of conditions on the underlying probability distributions to produce linear DET plots in a generalized setting. We show that the linear DETs on a normal deviate scale are not exclusively produced by normal distributions, however, that normal distributions do play an unique role in the threshold behavior as one moves along the DET line. An interesting connection between linear DETs and the Kullback-Leibler divergence is also discussed.

Index Terms— DET, ROC, Detection, Kullback-Leibler Divergence

1. INTRODUCTION

In detection systems, the Receiver Operating Characteristic (ROC) analysis involves a trade-off between the False Alarm and Miss error rates, and belongs to a traditional portfolio of methods for performance assessment. A particular ROC variant termed the *Detection Error Trade-Off* (DET), originally described by Swets [1], was introduced into the speaker recognition community by Martin et al. [2]. where it has found a wide-spread, virtually exclusive use, expanding to related detection tasks, particularly due to its popularization through the NIST speech technology evaluations [3].

Formally, an ROC chart is a set of operating points of a detection system obtained by plotting the False Alarm Rate, $P_{FA} \in [0,1]$ (on the abscissa), and the Miss Rate, $P_M \in [0,1]$ (on the ordinate).

The DET is an ROC plotted on non-linearly warped coordinates, such that systems with normally distributed detection scores will produce straight lines. The DET plot generally offers a better viewability and assessment of systems with close-to-normal score distributions [2]. The axes are warped by a *normal deviate* function ϕ^{-1}, which is an inverse of

$$\phi(P) = \int_{-\infty}^{P} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

This work was in part supported by the Czech Grant Agency under contract no. 401/03/H047

It is known that if the two score populations, i.e. the targets and the impostors, are normally distributed, the DET curve will be a straight line. Suppose the impostor population is from a Gaussian with μ_1, σ_1, and the target population is from a Gaussian with μ_2, σ_2. Then, the two error types as functions of a threshold t are

$$P_M(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2}(x-\mu_2)^2/\sigma_2^2} = \phi\left(\frac{t-\mu_2}{\sigma_2}\right)$$

$$P_{FA}(t) = \int_{t}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2}(x-\mu_1)^2/\sigma_1^2} = \phi\left(\frac{\mu_1-t}{\sigma_1}\right)$$

Taking the inverse

$$\frac{t-\mu_2}{\sigma_2} = \phi^{-1}(P_M) \quad \text{and} \quad \frac{\mu_1-t}{\sigma_1} = \phi^{-1}(P_{FA})$$

and equating by t

$$\phi^{-1}(P_M) = -\frac{\sigma_1}{\sigma_2}\phi^{-1}(P_{FA}) + \frac{\mu_1-\mu_2}{\sigma_2} \qquad (1)$$

we obtain a linear relationship on the DET chart (whose axes are warped by ϕ^{-1} as stated above).

Linear or close-to-linear plots are observed in practice and the normal assumption is often appropriate. However, an observed DET linearity tempts one to falsely conclude that the underlying score distributions must be normal. Even the classic reference paper [2] does not resist to state: *"If the resulting curves are straight lines, then this provides visual confirmation that the underlying likelihood distributions from the system are normal."*. In this paper we present a set of general results regarding line-producing distributions and, as a corollary, show that in fact a variety of non-Gaussian distributions will, too, produce linear DETs. However, Gaussianity, as will be shown, does play a unique role in linear DETs: it is the sole distribution attaining a linear threshold behavior as one moves along the DET line.

The paper is organized as follows: Based on the definitions given in Section 2, we state general conditions on line-producing distributions in Section 3. Section 4 is concerned with the linear behavior of the errors as a function of the detection threshold. Finally, an interesting relationship between the linear DETs and the symmetric Kullback-Leibler divergence is discussed in Section 5.

2. PRELIMINARIES

Notation Let $\mathbb{R}^\star = \mathbb{R} \cup \{-\infty, +\infty\}$.
Definition A *probability distribution* P is a function satisfying the following properties: a) $P : \mathbb{R}^\star \longrightarrow [0,1]$ b) is non-decreasing, c) is continuous on $\mathbb{R}^\star$, d) has continuous derivative on $\mathbb{R} \setminus N$ with N a finite (possibly empty) set of points, e) $P(-\infty) = 0$ and $P(+\infty) = 1$.

Remark The distribution function defined above is often referred to as cumulative density function.

Furthermore, let $P_I(t)$ denote the probability distribution of an impostor score being $< t$, and $P_T(t)$ that of a target score being $< t$. Hence, the probability of a miss error is $P_M(t) = P_T(t)$, and of false alarm error is $P_{FA}(t) = 1 - P_I(t)$, with t being the detection threshold.

Definition F is an *axis warping function* if it satisfies the following properties: a) F is a one-to-one mapping $F : [0,1] \longrightarrow \mathbb{R}^\star$, b) increasing on $[0,1]$, c) has continuous derivative on $(0,1) \setminus N$, where N is a finite set d) $\forall x \in (0,1) : F(x) \in \mathbb{R}$, e) $F(1-x) = -F(x)$, f) $F(1) = \infty, F(0) = -\infty$, g) is continuous.

Remark It is easy to show that f) and g) follow from a)-e).

It is easy to show that any valid F has an *inverse* F^{-1} with the properties as follows: a) it is a continuous one-to-one mapping $F^{-1} : \mathbb{R}^\star \longrightarrow [0,1]$, b) increasing on $\mathbb{R}^\star$, c) with derivative continuous on $\mathbb{R} \setminus N$, where N is finite, d) $\forall x \in \mathbb{R} : F^{-1}(x) \in (0,1)$ e) $1 - F^{-1}(p) = F^{-1}(-p)$ f) $F^{-1}(-\infty) = 0$ and $F^{-1}(\infty) = 1$.

Definition An *F-axis system* is an ROC coordinate system in which both axes are identically warped by a scaling function F.

Keeping the axis warping function general allows for deriving results for a *variety* of axis systems, noting that for the standard DET chart, i.e. the normal deviate scaling: $F \equiv \phi^{-1}$, and for the traditional ROC chart: $F(x) = x$.

3. LINE-PRODUCING DISTRIBUTIONS

Proposition 3.1 *In a given F-axis system, if P_I is any arbitrary impostor distribution function and a linear DET curve is produced, then the corresponding target function is given by*

$$P_T(t) = F^{-1}(aF(P_I(t)) + b), \quad \forall t \in \mathbb{R}^\star,\ a > 0,\ a, b \in \mathbb{R} \tag{2}$$

Moreover, P_T defined this way is a valid probability distribution whenever P_I is.

Proof If the DET should be a line the following must hold

$$F(P_M(t)) = -aF(P_{FA}(t)) + b \quad \forall t \in \mathbb{R}^\star$$

or

$$P_T(t) = F^{-1}(-aF(1 - P_I(t)) + b)$$

and due to symmetry

$$P_T(t) = F^{-1}(aF(P_I(t)) + b).$$

thus obtaining Eq. 2. To prove that for any given valid P_I (and F), a P_T is a valid distribution we consider that
a) $\mathrm{Rng}(F^{-1}) = [0,1]$ and $\mathrm{Dom}(P_I) = \mathbb{R}^\star$.
b) P_T is non-decreasing due to P_I being non-decreasing, F and F^{-1} being increasing, and $a > 0$.
c) composition of continuous functions on $\mathbb{R}$ is continuous, and for $\pm\infty$ F^{-1} must go to 0 or 1.
d) derivative of P_T is a multiplication of terms consisting of zeroth and first derivative of P_I, F and F^{-1}, which are up to a finite number of points continuous.
e) $P_T(\infty) = F^{-1}(aF(P_I(\infty)) + b) = F^{-1}(aF(1) + b) = F^{-1}(a\infty + b) = F^{-1}(a\infty + b) = F^{-1}(\infty) = 1$. Similarly for $-\infty$.
Hence, P_T is a valid distribution function. $\blacksquare$

The Proposition 3.1 shows that there is a multitude of distributions producing linear DETs. We only need to choose any valid P_I to obtain a corresponding P_T that will yield a linear DET. Note that the resulting functions P_I and P_T may not necessarily be from the same family (see example in Section 3.1).

Proposition 3.2 *For any $c_1, c_2, d_1, d_2 \in \mathbb{R}$ and $P_I(t) = F^{-1}(c_1 t + d_1)$ the following holds: $P_T(t) = F^{-1}(c_2 t + d_2)$ iff the DET plot is a line.*

Proof DET plot is a line iff $P_T(t) = F^{-1}(aF(P_I(t)) + b)$ iff $P_T(t) = F^{-1}(aF(F^{-1}(c_1 t + d_1)) + b)$ iff $P_T(t) = F^{-1}(ac_1 t + ad_1 + b)$, where $c_2 = ac_1$ and $d_2 = ad_1 + b$. $\blacksquare$

Remark In propositions 3.1 and 3.2 P_I and P_T can be interchanged.

Proposition 3.2 tells us that if one population is known to distribute as the inverse of the scaling function F (e.g. normally for $F \equiv \phi^{-1}$), and we observe a linear curve, then it follows that the other population must also distribute as F^{-1}.

3.1. Example of different distributions yielding a DET line

In a first example, we choose the *inverse sigmoid function* as our warping system (by this we replace the Gaussian case, in which, unfortunately, the function ϕ^{-1} has no closed form)

$$F(P) = \ln \frac{P}{1-P} \tag{3}$$

$$F^{-1}(t) = \frac{1}{1+e^{-t}} \tag{4}$$

Suppose $1 - P_{FA} = P_I \sim linear$:

$$P_I(t) = \begin{cases} \frac{t - t_{min}}{t_{max} - t_{min}} \ , \ t_{min} \leq t \leq t_{max} \\ 0 \ , \ t < t_{min} \\ 1 \ , \ t > t_{max} \end{cases} \qquad (5)$$

then it follows from Eq. 2, that

$$\begin{aligned} P_M(t) &= F^{-1}\left(a \ln \frac{P_I(t)}{1 - P_I(t)} + b\right) \\ &= \frac{1}{1 + \left(\frac{t - t_{min}}{t_{max} - t}\right)^{-a} / e^b} \ , \ t_{min} < t < t_{max} \end{aligned}$$

and $P_M(t) = 0$ for $t \leq t_{min}$ and $P_M(t) = 1$ for $t \geq t_{max}$. For example with $a = 2$ and $b = 0$ we obtain the rational function

$$P_M(t) = \frac{t^2 - 2t t_{min} + {t_{min}}^2}{2t^2 - 2t(t_{max} + t_{min}) + {t_{max}}^2 + {t_{min}}^2}$$

which is in general nonlinear for $t_{min} < t < t_{max}$, i.e. P_I and P_M are different.

To illustrate the Proposition 3.2 in this axis system, take

$$P_I(t) = \frac{1}{1 + e^{-\left(\frac{t - \mu_1}{\sigma_1}\right)}} \sim sigmoid$$

then according to Eq. 2

$$\begin{aligned} P_M(t) &= F^{-1}\left(a \ln \frac{P_I(t)}{1 - P_I(t)} + b\right) \\ &= \frac{1}{1 + e^{\left[-a'\left(\frac{t - \mu_1}{\sigma_1}\right) + b'\right]}} \sim sigmoid \qquad (6) \end{aligned}$$

In a second example, we return to the normal deviate scale $F = \phi^{-1}$. By choosing $P_I \sim linear$ as defined in Eq. 5, the line-producing function P_T is numerically approximated. We then randomly generate positive and negative samples from P_T and P_I, respectively (sample size 10^5). The resulting DET plot along with histograms for two different settings of a, b (from Eq. 2) is shown in Figure 1.

4. LINEAR THRESHOLD BEHAVIOR

By a *linear threshold behavior (LTB)* we understand that as we continuously vary the threshold t we move along the (DET) curve at a constant rate, i.e.

$$\frac{\partial F\left(P_{FA}(t)\right)}{\partial t} = const \qquad (7)$$

$$\frac{\partial F\left(P_M(t)\right)}{\partial t} = const \qquad (8)$$

Obviously on linear plots either condition implies the other. The LTB can be of interest in practice as the sensitivity of a

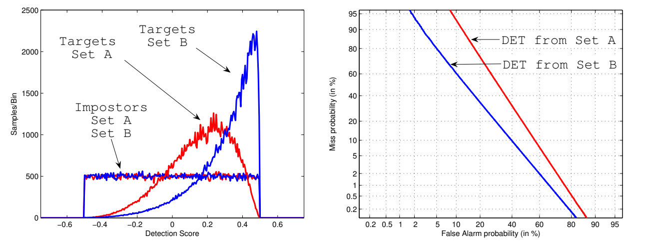

Fig. 1. Example of non-Gaussian and different target and impostor distributions producing a linear DET (right) on the standard deviate scale. Histograms (left) of the synthetically generated impostors were drawn from a uniform source (with a linear distribution function as in Eq. 5) and targets from a density function obtained via numerical approximation of $\phi(a\phi^{-1}(P_I) + b)$. Set A has $a = 1.9, b = -0.7$ and Set B has $a = 1.5, b = -1.2$.

detection system to threshold adjustments stays uniform along the DET curve.

Proposition 4.1 *For an F-scale system, the LTB is attained iff both populations distribute as F^{-1} (up to domain scale and shift).*

Proof "⇒": From Eq. 8

$$\begin{aligned} \frac{\partial F\left(P_M(t)\right)}{\partial t} &= k_1 \\ P_M(t) &= F^{-1}\left(k_1 t + k_0\right) \end{aligned}$$

Similarly for P_{FA} we have $P_{FA}(t) = F^{-1}\left(k_1' t + k_0'\right)$. Using the symmetry of F we get $P_I(t) = 1 - P_{FA}(t) =$ $F^{-1}\left(-k_1' t - k_0'\right)$, hence both populations are distributed the same way, up to the scale and shift.

"⇐": If we had $P_M(t) = F^{-1}\left(k_1 t + k_0\right)$ and $P_I(t) =$ $F^{-1}\left(k_1' t + k_0'\right)$, then obviously

$$\frac{\partial F\left(P_M(t)\right)}{\partial t} = const \quad \text{and} \quad \frac{\partial F\left(P_{FA}(t)\right)}{\partial t} = const$$

∎

Proposition 4.1 lends the normal distribution a unique role within the DET ϕ^{-1} axis system, namely that of exclusively attaining the LTB along any linear DET curve.

4.1. Example of LTB in the $sigmoid^{-1}$-axis system

Let F be again defined as in (3)

$$F(P) = \ln \frac{P}{1 - P}$$

and we are again considering a linear plot, i.e.

$$F\left(P_M(t)\right) = -a\, F\left(P_{FA}(t)\right) + b = -a \ln \frac{P_{FA}(t)}{1 - P_{FA}(t)} + b$$

Now,

$$\begin{aligned}\frac{\partial F\left(P_M(t)\right)}{\partial t} &= -a\frac{\partial F\left(P_{FA}\right)}{\partial P_{FA}}\frac{\partial P_{FA}(t)}{\partial t}\\ &= -a\frac{1}{P_{FA}(t)\left(1-P_{FA}(t)\right)}\frac{\partial P_{FA}(t)}{\partial t}\\ &= const\end{aligned}$$

the last equality is required as per Eq. 8. Hence

$$\frac{\partial P_{FA}(t)}{\partial t} = a'\, P_{FA}(t)\left(1-P_{FA}(t)\right) \quad (9)$$

which is a first-order differential equation and has a unique functional solution

$$P_{FA}(t) = \alpha_1 \frac{1}{1+\alpha_2 e^{\alpha_3 t}} = 1 - P_I(t)$$

with α_1, α_2, α_3 constants. In order for P_I to be a distribution function, α_1 must equal 1, therefore

$$P_I(t) \sim sigmoid$$

but since we know from Proposition 3.2 that if one of the populations in a $sigmoid^{-1}$-scale system is $sigmoid$-distributed, the second population must necessarily be also $sigmoid$-distributed to produce a linear curve (see Eq. 6) . Therefore only $sigmoid$-distributed populations can, exclusively, attain linear threshold behavior in the $sigmoid^{-1}$-scale system, in accordance with Proposition 4.1.

4.2. Traditional ROC

It is interesting to point out that the traditional ROC chart, defined by a linear axis system $F(x) = x$, has a linear plot obtainable for any type of distribution if and only if $P_T \equiv P_I$. This follows from the fact that for $F(x) = x$ the Eq. 2 becomes

$$P_T(t) = aP_I(t) + b$$

but since $P_I(-\infty) = 0$ and $P_I(+\infty) = 1$ it must hold that $a = 1$ and $b = 0$ for any P_I in order for P_T to be a valid distribution, i.e. $P_T \equiv P_I$. In other words a linear ROC plot is observed only with systems working at an absolute chance level.

5. DET RELATION TO THE SYMMETRIC KULLBACK-LEIBLER (SKL) DIVERGENCE

We return to normally distributed populations. The SKL divergence between two densities $p_1(x)$ and $p_2(x)$ is defined as

$$D(p_1||p_2) = \int_X p_1(x)\log\frac{p_1(x)}{p_2(x)} + p_2(x)\log\frac{p_2(x)}{p_1(x)}dx \quad (10)$$

and specifically for two Gaussians $p_1 \sim \mathcal{N}(\mu_1, \sigma_1)$, and $p_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ the SKL has the following form

$$\begin{aligned} & D(p_1||p_2) = && (11)\\ =\ & \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2}+\frac{\sigma_2^2}{\sigma_1^2}+\frac{(\mu_1-\mu_2)^2}{\sigma_2^2}+\frac{(\mu_1-\mu_2)^2}{\sigma_1^2}\right)-1 \\ =\ & \frac{1}{2}\left(R^2+R^{-2}+\Delta^2+(\Delta/R)^2\right)-1 && (12)\end{aligned}$$

with

$$\begin{aligned} R &= -\frac{\sigma_1}{\sigma_2} && (13)\\ \Delta &= \frac{\mu_1-\mu_2}{\sigma_2} && (14)\end{aligned}$$

representing the slope (R) and the bias (Δ) of a DET line as per Eq. 1.

Comparing to the Eq. 1, we note that the Eq. 12 contains the squared linear coefficients of *two DET lines* (in the ϕ^{-1}-axis system) that are mutually inverse, the first defined as in Eq. 1

$$\begin{aligned}\phi^{-1}\left(P_M\right) &= -\frac{\sigma_1}{\sigma_2}\phi^{-1}\left(P_{FA}\right)+\frac{\mu_1-\mu_2}{\sigma_2}\\ &= R\phi^{-1}\left(P_{FA}\right)+\Delta\end{aligned}$$

and the second defined as

$$\begin{aligned}\phi^{-1}\left(P_M\right) &= -\frac{\sigma_2}{\sigma_1}\phi^{-1}\left(P_{FA}\right)+\frac{\mu_1-\mu_2}{\sigma_1}\\ &= R^{-1}\phi^{-1}\left(P_{FA}\right)+\Delta/R\end{aligned}$$

which is its (linear) inverse.

From this observation it becomes clear that any system optimization maximizing the SKL is essentially minimizing the DET curve bias term, Δ, while keeping the slope unchanged (due to the $R^2 + R^{-2}$ term). We refer the interested reader to [4] for a study of a criterion derived from Eq. 1 specifically aiming at selective optimization of either the slope or the bias term (or both) of the DET line to optimize the detection system for specific operating regions.

6. REFERENCES

[1] J.A Swets, "The relative operating characteristic in psychology," *Science*, vol. 182, pp. 990–1000, December 1973.

[2] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. of the European Conference on Speech Communication and Technology (EUROSPEECH)*, Rhodes, Greece, September 1997, pp. 1895–8.

[3] (URL), "http://www.nist.gov/speech/tests/index.htm," .

[4] J. Navrátil and G.N. Ramaswamy, "DETAC - a discriminative criterion for speaker verification," in *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, Denver, CO, September 2002.