EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

THE DET CURVE IN ASSESSMENT OF DETECTION TASK PERFORMANCE

A. Martin, G. Doddington#, T. Kamm+, M. Ordowski+, M. Przybocki**
*National Institute of Standards and Technology, Bldg. 225-Rm. A216, Gaithersburg, MD 20899, USA
#SRI International/Department of Defense, 1566 Forest Villa Lane, McLean, VA 22101, USA
+Department of Defense, Ft. Meade, MD 20755, USA

ABSTRACT

We introduce the DET Curve as a means of representing performance on detection tasks that involve a tradeoff of error types. We discuss why we prefer it to the traditional ROC Curve and offer several examples of its use in speaker recognition and language recognition. We explain why it is likely to produce approximately linear curves. We also note special points that may be included on these curves, how they are used with multiple targets, and possible further applications.

INTRODUCTION

Detection tasks can be viewed as involving a tradeoff between two error types: missed detections and false alarms. An example of a speech processing task is to recognize the person who is speaking, or to recognize the language being spoken. A recognition system may fail to detect a target speaker or language known to the system, or it may declare such a detection when the target is not present.

When there is a tradeoff of error types, a single performance number is inadequate to represent the capabilities of a system. Such a system has many operating points, and is best represented by a performance curve.

The ROC Curve traditionally has been used for this purpose. Here ROC has been taken to denote either the Receiver Operating Characteristic [2,3,4] or alternatively, the Relative Operating Characteristic [1]. Generally, false alarm rate is plotted on the horizontal axis, while correct detection rate is plotted on the vertical.

We have found it useful in speech applications to use a variant of this which we call the DET (Detection Error Tradeoff) Curve, described below. In the DET curve we plot error rates on both axes, giving uniform treatment to both types of error, and use a scale for both axes which spreads out the plot and better distinguishes different well performing systems and usually produces plots that are close to linear.

Figure 1 gives an example of DET curves, while Figure 2 contrasts this with traditional ROC type curves for the same data. Note the near linearity of the curves in the DET plot and how better spread out they are permitting easy observation of system contrasts.

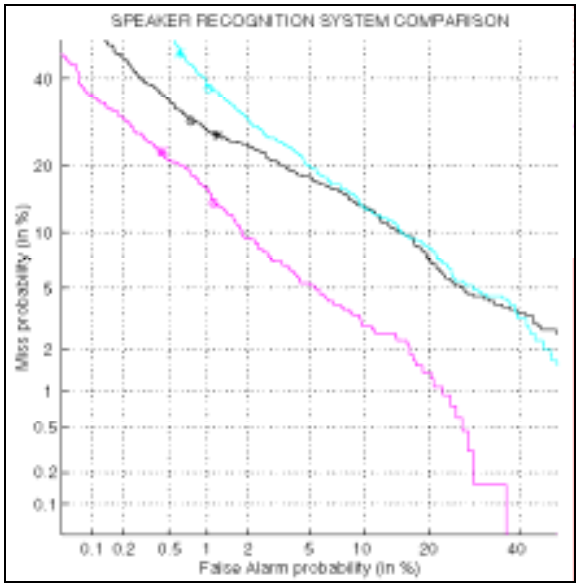

Figure 1: Plot of DET Curves for a speaker recognition evaluation.

GENERAL EVALUATION PROTOCOL

Our evaluations of speech processing systems are comparable to fundamental detection tasks. Participants are given a set of known targets (speakers or languages) for which their systems have trained models and a set of unknown speech segments. During the evaluation the speech processing system must determine whether or not the unknown segment is one of the known targets.

The system output is a likelihood that the segment is an instance of the target. The scale of the likelihood is arbitrary, but should be consistent across all decisions, with larger values indicating greater likelihood of being a target. These likelihoods are used to generate the performance curve displaying the range of possible operating characteristics.

Figure 2 shows a traditional ROC curve for a NIST coordinated speaker recognition evaluation task. The abscissa axis shows the false alarm rate while the ordinate axis shows the detection rate on linear scales. The optimal point is at the upper left of the plot, and the curves of well performing systems tend to bunch together near this

corner. (In the figures we omit the keys identifying the individual systems.)

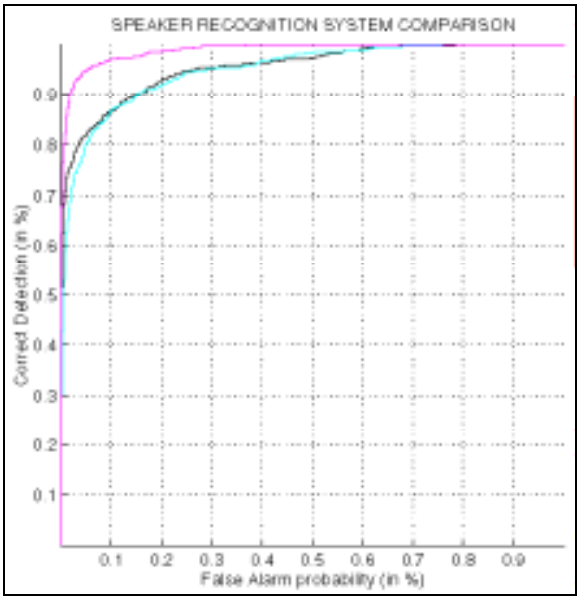

Figure 2: Plot of ROC Curves for the same evaluation data as in Figure 1.

NORMAL DEVIATE SCALE

Let us suppose that the likelihood distributions for non-targets and targets are both normally distributed with respective means u0 and u1. This is illustrated in Figure 3, where the variances of the distributions are taken to be equal.

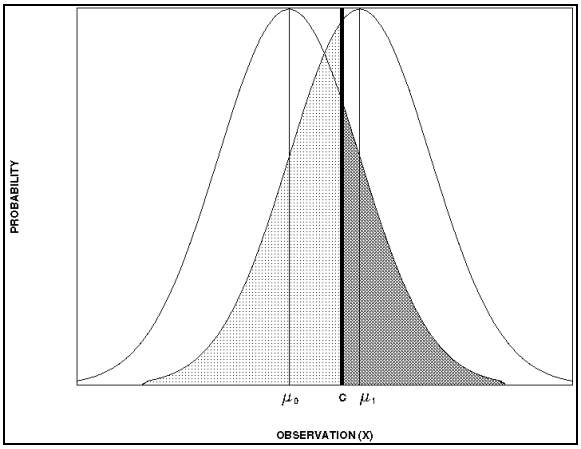

Figure 3: Normal Distributions.

The choice of an operating point c is shown by a bold line, and the two error types are represented by the areas of the shaded regions.

Now suppose that when we go to plot the miss versus the false alarm probabilities, rather than plotting the probabilities themselves, we plot instead the normal deviates that correspond to the probabilities. This is displayed in Figure 4.

In figure 4, we show probabilities on the bottom and left, and standard deviations on the top and right. The standard deviations are omitted from subsequent plots.

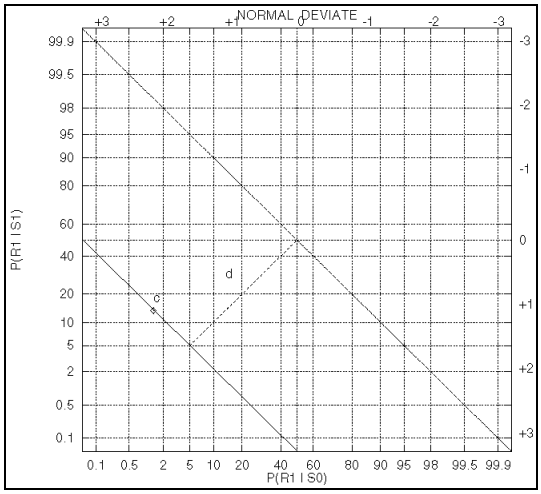

Figure 4: Normal Deviate Scale.

Note that the linearity of the plot is a result of the assumed normality of the likelihood distributions. The unit slope is a consequence of the equal variances. Also note that on the diagonal scale indicated we have

$$d = \sqrt{u1 - u0}$$

DET EXAMPLES

Figure 1 is a presentation of the DET curve for the same data as Figure 2. Note that the use of the normal deviate scale moves the curves away from the lower left when performance is high, making comparisons easier. We also see, as we typically do, that the resulting curves are approximately straight lines, corresponding to normal likelihood distributions, for at least a wide portion of their range.

There are two items to note about the DET curve. First, if the resulting curves are straight lines, then this provides a visual confirmation that the underlying likelihood distributions from the system are normal. Second, the diagonal y = -x on the normal deviate scale represents random performance.

If performance is reasonably good, we limit the curves to the lower left quadrant, as in Figure 1. We also somewhat arbitrarily limit the error rates plotted to 0.05%, or a bit over three standard deviations.

Figure 5 shows another set of typical DET Curves, in this case for a language recognition task. Once again, by

visual inspection of the DET curve we can verify that the underlying likelihood distributions are close to normal.

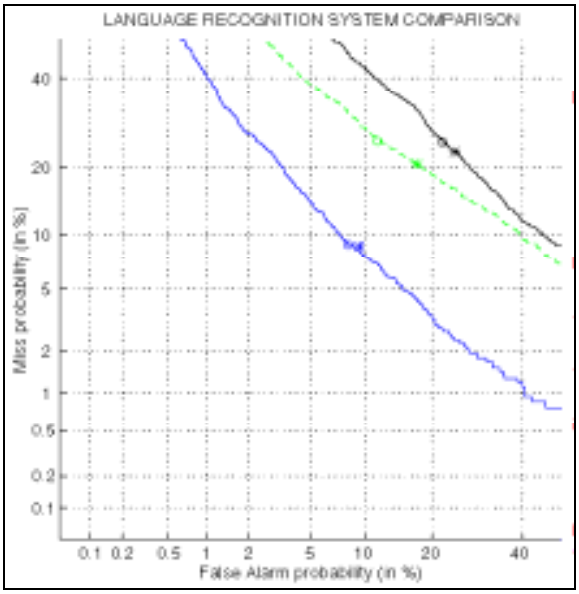

Figure 5: Plot of DET Curves from a language recognition evaluation.

Further examples of speaker recognition and language recognition DET Curves may be viewed at the NIST web site [7].

SPECIAL POINTS

A number of special points may be included on the DET curve. These points are not limited to speech processing tasks and can be applied to the fundamental detection task. For example, it may be of interest to designate points corresponding to a fixed false alarm rate or fixed missed detection rate, perhaps a performance objective for an evaluation. The grid lines on the example curves may be viewed this way. Confidence intervals, or a confidence box, around such points may also be included.

A weighted average of the missed detection and false alarm rates may be used as a kind of figure of merit or cost function. The point on the DET Curve where such an average is minimized may be indicated. In Figures 1 and 5, these points are indicated by °'s. (The error type weighting in figure 1 is 10:1. This corresponds to a cost of 10 for a missed detection and a cost of 1 for a false alarm. In figure 5, the error type weighting is 1:1.)

In our evaluations, the speech processing systems must also provide a hard yes or no decision as well as a likelihood score for each decision. The operating points of the hard decisions may be indicated on the curves. These are designated by *'s in Figures 1 and 5. The proximity of these points to the weighted average points described above is an indication of how appropriately the system implementers chose the hard decision operating points to optimize the chosen cost function.

AVERAGING ACROSS TARGETS

The DET curves presented all involved multiple targets, and required systems to provide likelihood scores on the same scale for all targets. For some applications, requiring a common scale may be considered undesirable. Furthermore, if all targets do not occur with about equal frequency, it is arguable that combining data from multiple targets may present a misleading indication of performance. The alternative is to generate separate curves for each target, and then generate an average curve across targets from these.

If the same non-targets are used with each target, then the ordinate values may be averaged for each abscissa value. This situation will not hold, however, if each target example also serves as a non-target example for each of the other targets. In this case, interpolation may be used to obtain a common set of abscissa values for the individual target curves which may then be averaged.

We prefer, however, to combine data from multiple targets directly. This requires systems to develop a common likelihood scale for all targets, which we believe desirable for many applications. We believe that with a large number of targets and roughly equal occurrences of all targets overall performance is effectively represented.

OTHER APPLICATIONS

The DET curve form of presentation is relevant to any detection task where a tradeoff of error types is involved. In previous years we have coordinated keyword and topic spotting evaluations involving such tasks.

We have also used the DET curve concept in large vocabulary speech recognition tasks where participants are asked to rate their confidence in the correctness of the words they hypothesize. A DET curve then shows the tradeoffs obtainable in the partial transcripts that result from setting thresholds on the confidence required to include hypothesized words. Figure 6 shows an example. Since performance at this task is poor at this point, all four quadrants are included in the curves.

CONCLUSION

The DET Curve has distinct advantages over the standard ROC type curve for presenting performance results where tradeoffs of two error types are involved. We have made it our standard way of presenting performance results of speaker and language recognition evaluations.

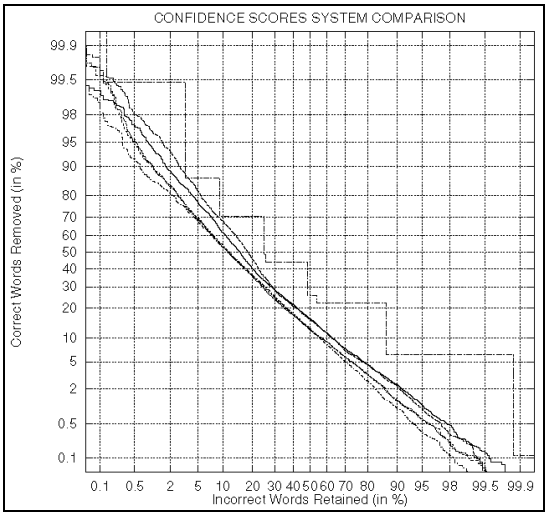

Figure 6: Plot of DET Curves from confidence scores in a large vocabulary speech recognition evaluation.

REFERENCES

[1] Swets, John A, "The Relative Operating Characteristic in Psychology", Science, Vol. 182, pp. 990-1000

[2] Swets, John A, ed., "Signal Detection and Recognition by Human Observers", John Wiley & Sons, Inc., pp. 611-648, 1964

[3] Green, David M. and Swets, John A., "Signal Detection Theory and Psychophysics", John Wiley and Sons, Inc., 1966

[4] Egan, James P., "Signal Detection Theory and ROC Analysis", Academic Press, 1975

[5] Speaker Recognition Workshop Notebook, Linthicum, MD, March 1996, unpublished

[6] Language Recognition Workshop Notebook, Linthicum, MD, June 1996, unpublished

[7] NIST web site - http://www.nist.gov/speech/