An Introduction to Application-Independent Evaluation of Speaker Recognition Systems

David A. van Leeuwen[1] and Niko Brümmer[2]

[1] TNO Human Factors,
Soesterberg, The Netherlands
david.vanleeuwen@tno.nl
[2] Spescom DataVoice,
Stellenbosch, South Africa
nbrummer@za.spescom.com

Abstract. In the evaluation of speaker recognition systems, the trade-off between missed speakers and false alarms has always been an important diagnostic tool. The NIST series of Speaker Recognition Evaluations has formalized this tool in the well-known DET-plot [1]. NIST has further defined the task of *speaker detection* with the associated *Detection Cost Function* (DCF) to evaluate performance. Since the first evaluation in 1996, these evaluation tools have been embraced by the research community and research groups have accordingly been optimizing their systems to minimize the DCF. Although it is an excellent evaluation tool, the DCF has the limitation that it has parameters that imply a particular *application* of the speaker detection technology.

In this chapter we introduce an evaluation measure that instead *integrates* detection performance over a range of application parameters. This metric, C_{llr}, was introduced in 2004 by one of the authors [2], and has been described extensively in a larger paper in 2006 [3], where various properties and interpretations of the measure are discussed at length. Here we introduce the subject with a minimum of mathematical detail, concentrating instead on the various interpretations of C_{llr} and its practical application.

We will emphasize the difference between *discrimination* abilities of a speaker detector ('the position/shape of the DET-curve'), and the *calibration* of the detector ('how well was the threshold set'). We will show that if speaker detectors can be built to output well-calibrated log-likelihood-ratio scores, users of such systems can define their own application parameters and still make minimum-expected-cost decisions by applying standard thresholds. Such detectors can be said to have *application-independent* calibration. The proposed metric C_{llr} can properly evaluate the discrimination abilities of the log-likelihood-ratio scores of the detector, as well as the quality of the calibration. Finally, we present a new graphical representation, which forms an analysis of some of the properties of C_{llr}. This representation, called an *Applied Probability of Error* (APE)-curve, is complementary to the traditional DET-curve.

1 Introduction

Formal evaluations have played a major role in the development of speech technology in the past decades. The paradigm of formal evaluation was established in speech technology by the National Institute of Standards and Technology (NIST) in the USA. By providing the research community with a number of essential ingredients, such as new speech data, tasks and rules, and a concluding workshop, these regular evaluations have led to significant improvements in all these evaluated technologies. It is therefore not strange that the evaluation paradigm has been adopted by other research and standards organizations around the world in various technology areas.

One of the most regularly held evaluations in the area of speech research is that of *text-independent speaker recognition*. This Speaker Recognition Evaluation (SRE) series has been organized yearly since 1996 by NIST, and has had its 11th edition in the first quarter of 2006. Despite the many factors that have varied along the various editions, a few key aspects have remained essentially constant. One of these is the primary evaluation measure, namely the *detection cost function* (DCF). It is specified in terms of the cost of misses and the cost of false alarms, as well as the prior probability for the target speaker hypothesis. In addition to the DCF, NIST compares the discrimination abilities of systems in *Detection Error Trade-off*[3] (DET)-curves [1], which researchers have embraced almost emotionally. In retrospect it can be concluded that it was quite an important insight of NIST to define DCF and the presentation of the error trade-off curves as they did, for it has become the standard in speaker recognition and is also gradually finding its way into other areas of research.

In the workshop concluding the most recent (2006) NIST SRE, an exciting new development became apparent. It was announced that NIST would in future employ a new primary evaluation measure. This measure, which we call C_{llr}, is the subject of this chapter. It was proposed in a conference paper in 2004 [2] and followed in 2006 by an extended journal paper [3]. The purpose of this chapter is to be a more accessible tutorial introduction to the topic. (Apart from the two above references, interested readers may want to see various other papers which have since appeared on the same or closely related topics [4–8])

In the following, we will first review the problem of speaker detection and the traditional evaluation techniques. This will be followed by motivation for and introduction to some aspects of the new C_{llr} evaluation methodology and the analysis thereof.

1.1 Recognition, verification, detection, identification

In the past, researchers have studied various forms of speaker recognition problems. Most notably, the problem of *speaker identification* has been studied extensively. It seems quite intuitive to see speaker recognition as an identification task, because that appears the way humans perceive the problem. When you hear

[3] Originally termed PROC in the 1996 evaluation plan

the voice of somebody familiar, you might immediately recognize the identity of the speaker. However, if we try to measure the performance of an automatic speaker identification system, we find a number of questions hard to answer. How many speakers should we consider in my evaluation? What is the distribution of speakers in the test? If we think about it deeper, we can see that performance measures such as identification accuracy will depend on the choice of these numbers in the evaluation. What if a speaker identification system is exposed to an 'unknown' speaker in the test? People have introduced 'open set identification' as alternative to 'closed set identification,' but really the latter is quite an unrealistic situation.

The solution to these undesirable questions lies in the proper statement of the speaker recognition task: in terms of *speaker detection*. Formally, the question is: *Given two recordings of speech, each uttered by a single speaker, do both speech excerpts originate from the same speaker or not?*[4] By developing technology that can answer this question for a broad range of speakers, many different applications are possible. Speaker verification is a direct implementation of the detection task, while open or closed set identification problems can be formulated as repeated application of the detection task.

The succinct statement of the speaker recognition problem in terms of *detection* has several advantages. The analysis of the evaluation can be performed in a standard way, which is the subject of Sect. 2. The evaluation measures do not intrinsically depend on the number of speakers or the distribution of so-called target and non-target trials. The true answer of the detection task can, if the evaluation data collection is carefully supervised, be known by the evaluator with very high confidence. Patrick Kenny summarized these positive aspects of the detection approach by saying: "I've never come across a cleaner problem [in speech research]".[5]

2 The traditional approach of the evaluation of speaker recognition systems

2.1 The errors in detection

In order to evaluate a speaker detection system, we can subject the system to two different kinds of *trial*. In each trial, the system is given two recordings of speech, originating either from the *same* speaker or from two *different* speakers. The former situation is called a *target trial* and the latter a *non-target trial*. The evaluator has a truth reference to tell the two types of trial apart, but the system under evaluation has only the speech recordings as input. It is therefore the purpose of the speaker detector to distinguish target trials from non-target

[4] One might call this a one-speaker open set identification task

[5] This is how the statement is recalled as perceived by the authors in a salsa-bar during the week of the 2006 Speaker Odyssey Workshop. However, the extremely high noise levels made proper human perception very hard, which is indicative of the fact that Automatic *Speech* Recognition cannot be stated as such a clean problem.

trials. In classifying the trials, there are two possible errors a system can make, namely

- false positives, or *false alarms*, classifying a non-target trial as a target trial, and
- false negatives, or *misses*, classifying a target trial as a non-target trial.

We observe that the speaker detection problem gives rise to *two* types of error, the rates of occurrence of which are to be measured in an evaluation. Having two different error-rates complicates things because it makes it hard to compare the performance of one system with another, or to observe an improvement in one system when it is adjusted. Since comparison is the essential goal of evaluation, it is important to find a way to do this. It is therefore the purpose of this chapter to examine the question: how do we combine these two error-rates into a single performance measure that is representative of a wide range of applications?

2.2 The DET-plot: A measure of discrimination

In order to continue, we need to introduce some of the basic concepts of how speaker detectors work. There are many sources of variability in speech signals and therefore a speaker detection system cannot be based on exact matching of two patterns. Rather, it works with (statistical) models, and it calculates some form of *score*[6] which represent the degree of support for the target speaker hypothesis rather than the non-target hypothesis. The higher (more positive) the score, the more the target hypothesis is supported and the lower (more negative) the score, the more the non-target hypothesis is supported. It can be shown that all the information which is relevant to making decisions between the two hypotheses and which can be extracted from the two speech inputs of a trial, can be distilled into a single real-valued score. Decisions as to which hypothesis is true can now be based on whether or not the score exceeds a well chosen threshold. Setting this threshold (a process known as *calibration*) is the next challenge.

If we now look at the scores that a speaker detector typically yields for the two types of trials, target and non-target trials, we may plot score distributions as in Fig. 1. These score distributions, obtained from a real speaker detector evaluated on NIST SRE 2006 data, has typical behaviour: the distributions overlap, the target scores having higher values on average than non-target scores, and the variance of the distributions is different. The threshold-based decision leads to the error-rates P_{FA} and P_{miss}, that can be read from the figure as the proportion of the non-target scores exceeding the threshold and the proportion of target scores below the threshold. From the figure you may also appreciate the fact that if the threshold were chosen differently, the values of P_{FA} and P_{miss} would change. More specifically, they would change in opposite directions. Thus, there is an inherent trade-off between lowering P_{FA} against lowering P_{miss}.

[6] often called a *likelihood ratio*, but we will not use this term for reasons that will become clear later

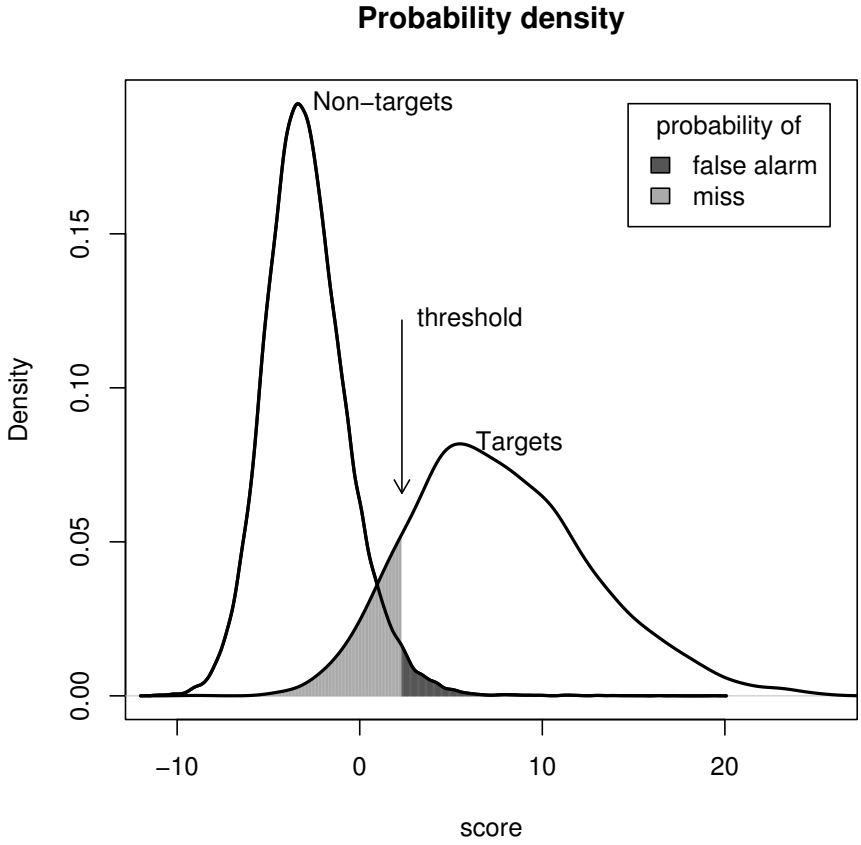

Fig. 1. The score distributions for non-target (left) and target (right) trials. The grey areas left and right of the threshold represent P_{miss} and P_{FA}, respectively.

This trade-off is most spectacularly shown in a graph that is known as the Detection Error Trade-off or *DET-plot* [1], where a parametric plot of P_{miss} versus P_{FA} is made, an example is shown in Fig. 2. The axes of a DET-plot are warped according to the *quantile function of the normal distribution*, or using another name, the probit function,

$$Q(p) = \text{probit}(p) = \sqrt{2}\,\text{erf}^{-1}(2p-1). \tag{1}$$

where p is P_{FA} or P_{miss}, and 'erf^{-1}' is the inverse of the *error function*. There are several effects of the warping of axes. Firstly, if the target and non-target score are distributed normally, the detection error trade-off will be a straight line,[7] with a slope $-\sigma_{\text{non}}/\sigma_{\text{tar}}$, where $\sigma_{\text{tar,non}}$ are the standard deviations of the target and non-target distributions, respectively [9, 10]. Secondly, the warping has the advantage that several curves plotted in the same graph gives rise to less clutter than if the probability axes were linear, as in ROC-curves (Receiver Operating Characteristic, which is the traditional way of plotting false alarms versus misses, or hits).

The DET-plot shows what happens as the decision threshold is swept across its whole range, but on the curve one can also indicate a fixed *operating point* as obtained when making decisions at a fixed threshold. It has been customary in NIST evaluations to require not only scores, but also hard decisions. The P_{miss} and P_{FA} measured for these hard decisions correspond to such an operating point on the curve.[8] It is good practice to draw a box around this point, indicating the 95 % confidence intervals of P_{FA} and P_{miss}, assuming trial independence and binomial statistics [11].

The DET-plot very clearly shows how the two error types can be traded off against each other. For a given DET-performance the false alarm rate can be reduced to an almost arbitrary low level by setting the detection threshold high enough, if one is prepared to accept a high miss rate. And vice versa; it all depends on the application of the system: if the costs of a false alarm are very high, or the prior probability of a target event is very low, we set the threshold high and we 'operate' in the upper-left corner of the plot. If the application sets different demands, we can operate at the opposite end. This trade-off is not new, a theory of signal detection was developed for radar signals midway the 20th century, and later used by psychophysicists to model human perception of stimuli in the sixties [12, 13]. We experience the same trade-off in everyday life, such as in trying to separate spam e-mails from serious messages, and in trying to create laws in society that can convict criminals while guaranteeing freedom for citizens. In fact, in understanding the DET or ROC curves it becomes apparent that striving for 'zero tolerance' or any other form of perfect filtering will backfire immediately by resulting in unreasonable high costs at the flip side of the coin.

[7] The reverse is not true, however. Note, that even though the underlying distributions deviate noticeably from normal distributions (see Fig. 1), the DET-curve is straight over a reasonably large range of probabilities.

[8] provided these hard decisions were indeed made by thresholding the same score that was used to generate the DET-plot

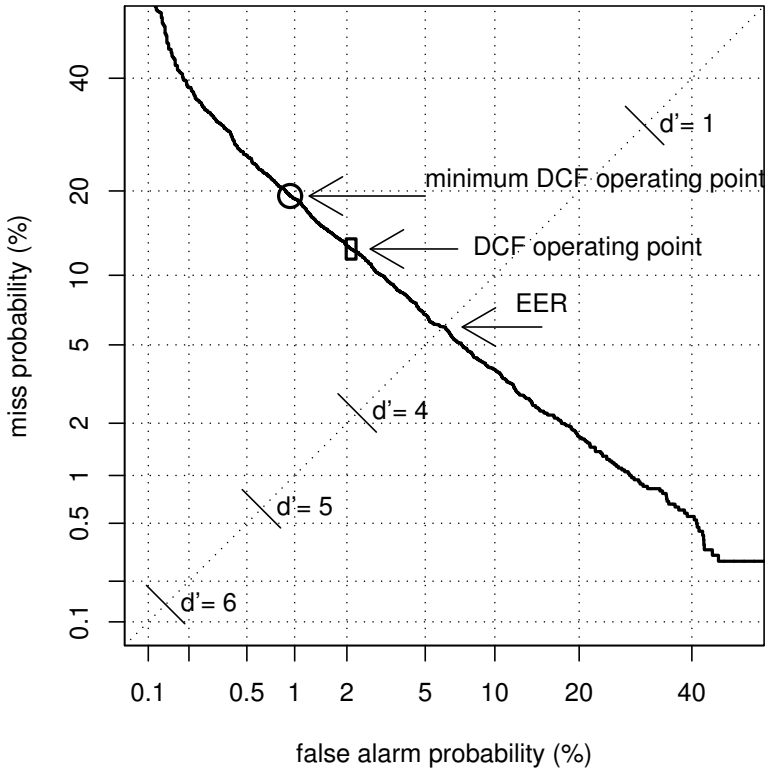

Fig. 2. A DET-plot, obtained from the distributions shown in Fig. 1. The line shows how the false alarm probability is traded-off against miss probability as the threshold increases from the lower-right to upper-left corner. The rectangle indicates the operating point of the decisions made, the co-ordinates correspond to the surface of the grey areas in Fig. 1. Further, the Equal Error Rate (EER) and the *operating points* for the DCF and the 'minimum DCF' (see Sect. 2.3) are indicated in the figure. Not part of a normal DET-plot, we have indicated, using little diagonal lines, the position of DET-curves originating from two Gaussian score distributions of equal variance σ^2, with means separated by $d'\sigma$, for several values of d' (see Sect. 2.2).

Returning now to speaker recognition, researchers have grown very fond of DET-curves because they indicate the discrimination potential of their system at a glance. DET-curves more towards the lower-left indicate better discrimination ability between the target and non-target trials, and hence better algorithms. Tiny improvements in the detector will show noticeable displacement in the DET-curve, which stimulates the researcher to think of even more clever things. A DET-plot is a great diagnostic tool: if the curve deviates far from a straight line, or shows unexpected cusps or bends, this is usually an indication that there is something wrong in the detector or in the evaluation data or its truth reference. As a final goody, plotting a DET-curve does not require setting a threshold.

The equal error rate. We went from decisions and P_{FA} and P_{miss} to no decisions and a whole *curve* that characterizes our detector. Can we somehow summarize the DET-curve as a single value? Yes, we can, in several ways.

Firstly, noticing P_{FA} and P_{miss} move in opposite directions if the threshold is changed, there always is a point where $P_{\text{FA}} = P_{\text{miss}}$. This joint value of the error rates is called the *Equal Error Rate* or EER. In the DET-plot it can be found as the intersection of the DET-curve and the diagonal. The EER is a concise summary of the discrimination capability of the detector.[9] As such it is a very powerful indicator of the *discrimination* ability of the detector, across a wide range of applications. However, it does *not* measure calibration, the ability to set good decision thresholds.

It may be interesting to compare the EER to a related measure from signal detection theory. Here the task is to detect a signal in Gaussian noise, and hence the two distributions to be separated are normal and have equal variance. In this case, the DET-curve is completely characterized by the single parameter 'd-prime,' the distance between the means of the distributions measured in units of the standard deviation: $d' = (\mu_{\text{tar}} - \mu_{\text{non}})/\sigma$. In Table 1 the relation between d' and the EER is shown, in order to give an idea what the separation of the target and non-target distributions means in terms of EER. Another way of seeing d' is in the DET-plot (see Fig. 2), where it represents straight lines of slope -1. The value of d' determines where the diagonal is crossed, starting at the upper-right corner for $d' = 0$ moving down linearly to the lower-left corner where $d' \approx 6$.

Table 1. Relation between d', the separation of distribution in terms of standard deviations, and the EER.

d'	0	1	2	3	4	5
EER (%)	50.0	30.9	15.8	6.7	2.27	0.62

[9] It can be shown [14, 3] that if decision thresholds are always set optimally, then the EER is the *upper bound* of the average error-rate of the detector as P_{tar} is varied. By average error-rate, we mean $P_{\text{tar}}P_{\text{miss}} + (1 - P_{\text{tar}})P_{\text{FA}}$, where P_{tar} is the prior probability of a target event.

2.3 The Detection Cost Function: simultaneous measure of discrimination and calibration

In calculating the DET-plot and EER, the evaluator effectively chooses optimal decision thresholds, with reference to the truth. These evaluation procedures therefore do not measure the actual decision-making ability of the detector on unseen data. The canonical solution is a direct one—simply require the detector to make decisions and then count the errors. Now how do we now combine these error counts (of two types of error) into a scalar measure of goodness of decision-making ability?

At a first glance, one could simply use the total number of errors as a performance measure. Indeed, this solution is routinely practised by the machine learning research community. However, reflecting on real applications there are at least two important complications:

- The proportion of targets and non-targets may be different from the proportions in the evaluation database.
- The two types of errors may not have equally grave consequences. For example, for a fraud detection application the costs of a missed target (cross customers) can be higher than the cost of a false alarm (a fraudulent action not observed), while for access control the cost of a false alarm (security breach) may outweigh the cost of a miss (annoyed personnel).

It therefore makes sense to weight the two normalized error-rates with (i) the prior probability of targets in the envisaged application and (ii) the estimated *costs* of the two error types. Applying these weightings, one then arrives at a scalar performance measure, namely the *expected cost of detection errors*,

$$C_{\text{det}}(P_{\text{miss}}, P_{\text{FA}}) = C_{\text{miss}} P_{\text{miss}} P_{\text{tar}} + C_{\text{FA}} P_{\text{FA}} (1 - P_{\text{tar}}). \tag{2}$$

This function has become known as the *detection cost function*. Here the normalized error-rates P_{miss} and P_{FA} are determined by the evaluator by counting errors. The application dependent cost parameters C_{miss} and C_{FA} are discussed above, and the parameter P_{tar} is the prior probability that a target speaker event occurs in the application. This prior must be assigned to correspond to some envisaged application of the speaker detector.

Given prescribed values for the parameters of C_{det}, the onus now rests on the designer of a speaker recognition system under evaluation, to choose a score decision threshold that minimizes C_{det}. For this purpose the evaluee may use a quantity of development data with a known truth reference. Minimizing C_{det} on the development data may or may not give a C_{det} that is close to optimal on new unseen evaluation data. This is an important part of the art of designing a speaker detector: to calculate scores that are well-normalized so that thresholds set on development data still work well on unseen data.

In summary, the three application-dependent parameters C_{miss}, C_{FA} and P_{tar}, form the detection cost function $C_{\text{det}}(P_{\text{miss}}, P_{\text{FA}})$, which gives a single scalar performance measure of a speaker detection system.

The detection cost function is a simultaneous measure of *discrimination* and *calibration*. This error measure of a detector will have a low value provided that both (i) EER is low *and* (ii) the threshold has been set well.

C_{det} has been used since the first NIST speaker recognition evaluation in 1996 as the primary evaluation measure, and with it, the three application-dependent cost parameters have been assigned values $C_{\text{miss}} = 10$, $C_{\text{FA}} = 1$ and $P_{\text{tar}} = 1\,\%$. These values have never changed in the evaluations, and occasionally a researcher wonders how these values were chosen. The long tradition and fixed research goals have caused these choices to fade from our collective memory, but in a recent publication [11] an example of an application with these cost parameters is given.

'Minimum Detection Cost.' Minimum C_{det} is similar, but not identical to EER. It is a measure of discrimination, but not of calibration. It is defined as the optimal value of C_{det} obtained by adjustment of the detection threshold, given access to the truth reference. Unlike EER it is dependent on the particular application-dependent parameters of C_{det}.

In the context of the NIST SRE, it is customary to indicate $C_{\text{det}}^{\text{min}}$ on DET-curves, as is shown by the circle in Fig. 2. Note that this circle does not show the numerical value of $C_{\text{det}}^{\text{min}}$, rather it shows the values of P_{miss} and P_{FA} at which C_{det} is minimized. This is in contrast to the APE-curve, which we introduce below, which does directly show the numerical value of $C_{\text{det}}^{\text{min}}$.

Discussion. So we've found two more performance metrics, EER and $C_{\text{det}}^{\text{min}}$, that each summarize the DET-plot in their own way. Both are used extensively in literature, the former in a 'general application' context and the latter in a 'NIST evaluation' context. They are very important performance metrics, but they circumvent one major issue: setting the threshold. In fact, EER and $C_{\text{det}}^{\text{min}}$ are *after the fact* error measures. They imply that the threshold can not be set until all trials have been processed and, moreover, the truth about the trials is known. Summarizing, EER and $C_{\text{det}}^{\text{min}}$ are great for indicating the discrimination potential, but they do not fully measure the capability of making hard decisions.

Is this really a problem? For many researchers it is not. Setting the threshold, as is necessary for submitting results to a NIST evaluation, is simply based on last year's evaluation data, for which the truth reference has been released.[10] This usually results in a C_{det} that is not too much above $C_{\text{det}}^{\text{min}}$, and everything is fine. Sometimes, the evaluation data collection paradigm has changed or the recruitment of new speakers has been carried out in a different way, and the calibration turns out wrong. A real shame, but usually most participating systems 'get hurt' in the same way, and there is always a next year to do better.

So let us recapitulate our quest for a single, application independent performance measure for speaker recognition systems. We started with a clear and

[10] Often, the calibration happens just before the results are due. The present authors are in this respect not different from other researchers.

unambiguous statement of the task of a speaker recognition system. This lead to two types of error which are interrelated by means of a trade-off. By using a cost function C_{det}, we could reduce the two error measures to a single metric, at the cost of having to define application-dependent parameters. Postponing the setting of a threshold gave us a beautiful DET-plot and a powerful EER summary, at the cost of not measuring calibration.

3 A new approach to speaker recognition evaluation

In the previous section we have introduced several measures characterizing the performance of a speaker recognition system. Although they each have their merits and their use is quite widespread, we will show in this section that we can demand more information from a speaker detector than just a score and a decision, and that there exists a metric that says how good this information is. It combines the concept of expected costs, like C_{det} does, with soft decisions and application-independence, like the DET-curve suggests. Before we introduce it, we are going to have a closer look at the interpretation of scores.

3.1 The log-likelihood-ratio

So far, we have learnt that a speaker detection system produces a score for every trial. The only thing we have required of the score is that a higher score means that the speech segments are more alike. A set of scores is sufficient to produce a DET-curve, and with an additional threshold we can also calculate C_{det}. But there is a *lot* of freedom in the values of the scores. First, there is an arbitrary offset that can be added to all scores (and the threshold) and nothing in the evaluation will change. Or the score can be scaled; in fact, the whole score-axis can be warped by any monotonic rising function, and everything in the DET-plot will stay exactly the same. There is no meaning in the scores, other than an ordering.

We can use this freedom in score values to fix the problem of application dependence. To see how this works, we examine how a score s for a given trial can be used to make an optimal decision for that trial. The expected cost of making an *accept* decision is $(1 - P(\text{target trial}|s))C_{\text{FA}}$, while the expected cost of making a *reject* decision is $P(\text{target trial}|s)C_{\text{miss}}$. Here $P(\text{target trial}|s)$ is the *posterior* probability for a target trial, given the score s. The minimum-expected-cost decision is known as a Bayes decision.[11] To make a Bayes decision, we need the posterior, which may be expressed, via Bayes' rule, as

$$\text{logit}\, P(\text{target trial}|s) = \mathcal{L}(s) + \text{logit}(P_{\text{tar}}) \tag{3}$$

[11] It is easily shown that if one makes a Bayes decision for every trial, this will also optimize the expected error-rate over all the trials, which is just our evaluation objective C_{det}.

where[12]

$$\mathcal{L}(s) = \log \frac{P(s|\text{target trial})}{P(s|\text{non-target trial})} \tag{4}$$

is known as the *log-likelihood-ratio* of the score. Putting this all together, we get a concise decision rule:

$$\text{decision}(s, \theta) = \begin{cases} \text{accept} & \text{if } \mathcal{L}(s) \geq -\theta, \\ \text{reject} & \text{if } \mathcal{L}(s) < -\theta, \end{cases} \tag{5}$$

where the decision threshold θ is a function of the application-dependent cost and prior parameters,

$$\theta = \log \left(\frac{P_{\text{tar}}}{1 - P_{\text{tar}}} \frac{C_{\text{miss}}}{C_{\text{FA}}} \right) \tag{6}$$

Equation (5) forms a neat separation between $\mathcal{L}(s)$ and θ. The purpose of the score, s, is to extract relevant information from the given speech data of the trial. The purpose of $\mathcal{L}(s)$ is to shape, or *calibrate*, this information into a form that can be used in a standard way to make good decisions. The information, $\mathcal{L}(s)$, extracted from the speech data is application-independent, because all the application-dependent parameters have been separated and encapsulated into the single *application parameter* θ.

Notice that $\mathcal{L}(s)$ may also be called a *score*. It has the same look and feel[13] as s, where more negative scores favour the non-target hypothesis and more positive scores favour the target hypothesis. The difference is that $\mathcal{L}(s)$ is calibrated so that minimum-expected-cost decisions may be made with the standard threshold θ.

In fact $\mathcal{L}(s)$ may be interpreted as expressing the degree of support that the raw score s gives to one or the other hypothesis. When $\mathcal{L}(s)$ is close to zero, the score does not strongly support either hypothesis, but as the absolute value of $\mathcal{L}(s)$ grows there is more support for one or the other hypothesis. The hypothesis that is favoured is indicated by the sign of $\mathcal{L}(s)$.

If a speaker detector can produce $\mathcal{L}(s)$ instead of the raw s, this has obvious advantages for users. The *same* system can now be used by different users having *different* applications (i.e., different θ), and still the calibration is right. The user does not have to ask the system developer: "My application parameters have changed. Could you please re-calibrate your detector?" Now the user can easily calculate the threshold θ and indeed change it at will as circumstances dictate.

So what is new here? Nothing in fact. The theory of making Bayes decisions has been known for a long time. The catch is that even if your DET-curve is good it may also be difficult to calculate well-calibrated soft decisions in log-likelihood-ratio form, just like it used to be difficult to set good hard decision

[12] We use the function: $\text{logit}\, p = \log \frac{p}{1-p}$, which re-parametrizes probabilities as *log odds*, because for binary hypotheses, it transforms Bayes' rule to the elegant additive form of (3).

[13] This is why we prefer to work with a log-likelihood-ratio, rather than a likelihood-ratio. The (non-negative) likelihood-ratio has the uncomfortable asymmetry where smaller scores are compressed against 0.

thresholds for C_{det}. The key to this problem is that until quite recently it has not been known in the speaker recognition community how to *evaluate the quality* of detection log-likelihood-ratios. The purpose of this chapter is therefore to introduce the reader to how this may be done. Once we know how to measure, half the battle towards improving performance has been won.

3.2 Log-likelihood-ratio cost function

At a first glance, evaluation of log-likelihood-ratio scores may be accomplished by a small adjustment of the NIST SRE protocol:

> Instead of having evaluees submit hard decisions for evaluation via C_{det}, they are now required to submit soft decisions in log-likelihood-ratio form. Then instead, the *evaluator* makes the decisions by setting the threshold at $-\theta$. These decisions may then be plugged into C_{det} as before, to get a final evaluation result.

In principle this is a very good plan, but it has the flaw of not really changing anything. If the value of θ is known to participants, then they may calibrate their scores to work well only at the specific point on the log-likelihood-ratio axis that is 'sampled' by evaluation at θ. Intuitively, sampling the log-likelihood-ratio at a single point can show that scores have been shifted to have log-likelihood-ratio interpretation, but it still leaves the scale of the evaluated scores completely arbitrary.

Once we have realized that a single sampling point is the problem, it is conceptually easy to fix: just sample the decision-making ability of the log-likelihood-ratio scores under evaluation at more than one value of θ. The evaluator may now calculate a C_{det} at each of these operating points. This leaves the questions of (i) how many points do we need to sample, (ii) which points do we choose and (iii) how do we combine the different C_{det} results over these points in order to get a single metric?

Of course there are many good answers to these questions. Here we discuss the particular solution which has been motivated in detail in [3]. This solution proposes to sample C_{det} over an infinite 'spectrum' of operating points and to then simply integrate over them, thus:

$$C_{\text{llr}} = C_0 \int_{-\infty}^{\infty} C_{\text{det}}\big(P_{\text{miss}}(\theta), P_{\text{FA}}(\theta), \theta\big)\ d\theta \tag{7}$$

where C_{llr} is the new metric, which we call the *log-likelihood-ratio cost function* and where $C_0 > 0$ is a normalization constant. Some notes are in order:

- The error-rates P_{miss} and P_{FA} are now functions of θ, because $-\theta$ is just the decision threshold. By sweeping the decision threshold, the evaluator is effectively sweeping the whole DET-curve of the system under evaluation. This effectively turns C_{llr} into a summary of *discrimination* ability over the whole DET-curve, somewhat similar to EER.

– Equally important is the fact we have now also made C_{det} dependent on θ. Since C_{det} implies making actual decisions, we are also incorporating the evaluation of calibration into our metric. Moreover, since C_{det} varies with θ, we are also measuring calibration over the whole θ-spectrum. Recall from (2) that C_{det} is parameterized by the triplet $(P_{\text{tar}}, C_{\text{miss}}, C_{\text{FA}})$. We may parametrize C_{det} equivalently[14] by $(\tilde{P}_{\text{tar}}, \tilde{C}_{\text{miss}} = 1, \tilde{C}_{\text{FA}} = 1)$, where $\tilde{P}_{\text{tar}}$ 'incorporates' the cost parameters. This single parameter $\tilde{P}_{\text{tar}}$ can be expressed in terms of θ,

$$\begin{aligned}\tilde{P}_{\text{tar}} &= \frac{P_{\text{tar}}C_{\text{miss}}}{P_{\text{tar}}C_{\text{miss}} + (1 - P_{\text{tar}})C_{\text{FA}}} \\ &= \frac{1}{1 + e^{-\theta}} = \text{logit}^{-1}\,\theta \end{aligned} \tag{8}$$

If we parameterize like this, then $\theta = \text{logit}(\tilde{P}_{\text{tar}})$ has the interpretation of *prior log-odds*. The interested reader may consult [3] for further motivation of this parametrization. In short, although specifying cost and prior are necessary when making decisions in real applications, having both costs and prior as evaluation parameters is redundant. Since the cost and prior multiply to form the parameter θ, we may arbitrarily assign fixed costs and parametrize the entire spectrum of applications by the single parameter $\tilde{P}_{\text{tar}}$, or equivalently by θ. By assigning unity costs we gain the advantage that now C_{llr} may be interpreted as an integral over *error-rates*. Finally, since we are making actual decisions and evaluating them via C_{det}, we are not only measuring discrimination, but we are also at the same time measuring *calibration*.

Realizing that the new measure C_{llr} is a measure of both *discrimination* and *calibration*, we see that C_{llr} for a detector will be good provided that both (i) EER is low *and* (ii) $\mathcal{L}(s)$ is reasonably well calibrated over all operating points of the θ-spectrum.

To recapitulate, C_{det} is a measure of discrimination and calibration suitable for evaluating *hard* (application dependent) detection decisions, while C_{llr} is a measure of discrimination and calibration suitable for evaluating *soft* (application-independent) detection decisions in log-likelihood-ratio form.

Practical calculation Equation (7) is a derivation and an interpretation of our new metric C_{llr} but how do we practically calculate this integral? The good news is that it has an analytical closed-form solution:

$$C_{\text{llr}}\big(\{\mathcal{L}'_t\}\big) = \frac{1}{2\log 2}\left(\frac{1}{N_{\text{tar}}}\sum_{t\in\text{tar}}\log(1 + e^{-\mathcal{L}'_t}) + \frac{1}{N_{\text{non}}}\sum_{t\in\text{non}}\log(1 + e^{\mathcal{L}'_t})\right). \tag{9}$$

where $\mathcal{L}'_t$ is the attempt of the system under evaluation to calculate the log-likelihood-ratio (of (4)) for trial t; and where 'tar' is a set of N_{tar} target trials

[14] By *equivalent*, we mean that identical decisions, DET-curves and comparisons between systems are made. The DCF itself is scaled down by a factor $P_{\text{tar}}C_{\text{miss}} + (1 - P_{\text{tar}})C_{\text{FA}}$, which is 1.09 for the NIST parameters.

and 'non' is a set of N_{non} non-target trials. The two normalized summation terms respectively represent expectations of 'log costs' for target trials (left-hand term) and for non-target trials (right-hand term).

Let us look more closely at these log costs. For a target trial the cost is $C_{\text{tar}} = \log(1 + e^{-\mathcal{L}'_t})$. If the detector correctly gives a high degree of support for the target hypothesis, $\mathcal{L}'_t \gg 1$, then the cost is low: $C_{\text{tar}} \approx 0$; but if it incorrectly gives a high degree of support for the non-target hypothesis, $\mathcal{L}'_t \ll -1$, then the cost is high[15]: $C_{\text{tar}} \approx |\mathcal{L}'_t|$. Conversely, the cost for non-target trials, $C_{\text{non}} = \log(1 + e^{\mathcal{L}'_t})$, behaves the other way round.

We have seen that extremely strong support for either hypothesis can have high cost, but what is the cost of a neutral log-likelihood-ratio? When $\mathcal{L}'_t = 0$, then $C_{\text{tar}} = C_{\text{non}} = \log 2$. This means that *the reference detector*, which does not process speech and which just outputs $\mathcal{L}'_t = 0$ for every trial, will earn itself a reference value of $C_{\text{llr}} = 1$. This is of course no coincidence, but is a consequence of the normalization factor in (9).

3.3 Discrimination/Calibration decomposition: The PAV algorithm

So far we have shown how the new cost measure C_{llr} generalizes C_{det}—but can we also find an analogy for $C_{\text{det}}^{\text{min}}$, the minimum achievable C_{det} if calibration were right? Again, the answer is affirmative. Just like a miscalibrated threshold can be fixed, post hoc, by choosing a different threshold that minimizes C_{det}, it is possible to find a *monotonic rising* warping function w, which, when applied applied to $\mathcal{L}'_t$ for every trial t, will minimize C_{llr} as measured on the warped log-likelihood-ratios $\mathcal{L}''_t = w(\mathcal{L}'_t)$. As before the minimization is performed given the truth reference for the evaluation, but note that it involves finding the whole warping function w rather than just a single threshold value. The warping function is constrained to be monotonic rising for several reasons:

- It is consistent with applying a single decision threshold to both $\mathcal{L}'_t$ and $\mathcal{L}''_t$.
- A monotonic rising function is invertible and therefore information-preserving. The warping function should correct only the *form* (calibration) of the output, but not the *content* (discriminative ability) of the score.
- The DET-curve (and therefore also the EER) is invariant under monotonic rising warping.
- If there were no constraint, C_{llr} would trivially be optimized to zero, which is a useless result.

[15] When degree of support is expressed as log-likelihood-ratio, then the behaviour of the log-cost is intuitively pleasing: if the detector output has the wrong sign, there is a cost which increases with the magnitude of the error. But if degree of support is instead expressed as a posterior probability, then a posterior of exactly 0 corresponds to $\mathcal{L}'_t = -\infty$ and then $C_{\text{tar}} = \infty$ (likewise, for a non-target trial, a posterior of 1 gives $C_{\text{non}} = \infty$). This is not a flaw of the C_{llr} metric. Rather it shows that a posterior of 0 or 1 is an unreasonable output to give in a pattern recognition problem where there can never be complete certainty about the answer. Working with system outputs (of moderate magnitude) in *log*-likelihood-ratio form, rather than likelihood-ratio form or posterior probability form naturally guards against this problem.

How do we find w? Note first that since monotonicity is the only constraint, every value of w can be optimized independently for every trial, in a *non-parametric* way. There is a remarkable algorithm known as the *Pool Adjacent Violators* (PAV) algorithm[16] which can be employed to do this constrained non-parametric optimization. The input is the system-supplied log-likelihood-ratio scores for every trial as well as the truth reference. The output is a set of optimized log-likelihood-ratio values for these trials, where the sorted ordering of input and output scores remains the same, because of the monotonicity. With these optimally calibrated log-likelihood-ratios $w(\mathcal{L}'_t)$ we can apply (9) to find the *minimum* C_{llr}

$$C_{\text{llr}}^{\text{min}} = C_{\text{llr}}\big(\{w(\mathcal{L}'_t)\}\big). \tag{10}$$

It is beyond the scope of this chapter to go into the details of the PAV algorithm (details are available in [3] and references therein), but it may be instructive to see what the warping function $w(\mathcal{L})$ typically looks like. Let us take the system that produced the score distributions in Fig. 1 and the DET-curve shown in Fig. 2. We plot the warping function $w(\mathcal{L})$ for this system, as found by the PAV algorithm, in Fig. 3. The PAV warping function has a stepped nature, which is a consequence of the 'pooling' of monotonicity violators. This system shows an average slope of 1 over a reasonable range of $\mathcal{L}$, but there is an offset. The log-likelihood-ratios given by this system are too optimistic towards target speakers. One can further observe a non-linear flattening of the curve at the extremes, indicating that the system-supplied log-likelihood-ratio tended to be over-optimistic in those regions.

Note that the PAV algorithm can also be used as the basis for calibration. Just like a detector can be calibrated for a single application-type by choosing a threshold that minimizes C_{det} on some development test data, it is possible to calibrate log-likelihood-ratio scores by applying the PAV algorithm to development test data scores s, to minimize C_{llr} for that data. The warping function $w(s)$ can then be interpreted as a *score to log-likelihood-ratio* function $\mathcal{L}(s)$. Having said this, we leave the subject of calibration methods, since it is not a topic of this chapter. Rather, this is the story how to *measure* calibration.

Recall that C_{llr} is a measure of *both discrimination and calibration*. But since $C_{\text{llr}}^{\text{min}}$ has any calibration mismatch optimized away, it is a now pure measure of *discrimination*. This now allows us to decompose[17] C_{llr} to also obtain a pure measure of calibration. Because of the logarithmic nature of C_{llr}, it turns out that it is appropriate to form an additive decomposition: Our measure of calibration now becomes just $C_{\text{llr}} - C_{\text{llr}}^{\text{min}}$. This difference is non-negative, is close to zero for well-calibrated systems, and grows without bounds as the system under calibration becomes increasingly miscalibrated. In summary, this PAV-based

[16] It is also known as *isotonic regression*.

[17] In this chapter, we use the term *discrimination/calibration* decomposition. This is similar in spirit, but not in form, to the *refinement/calibration* decomposition which was introduced by De Groot two decades ago [15] and again recently examined for speaker detection in ref. [6]

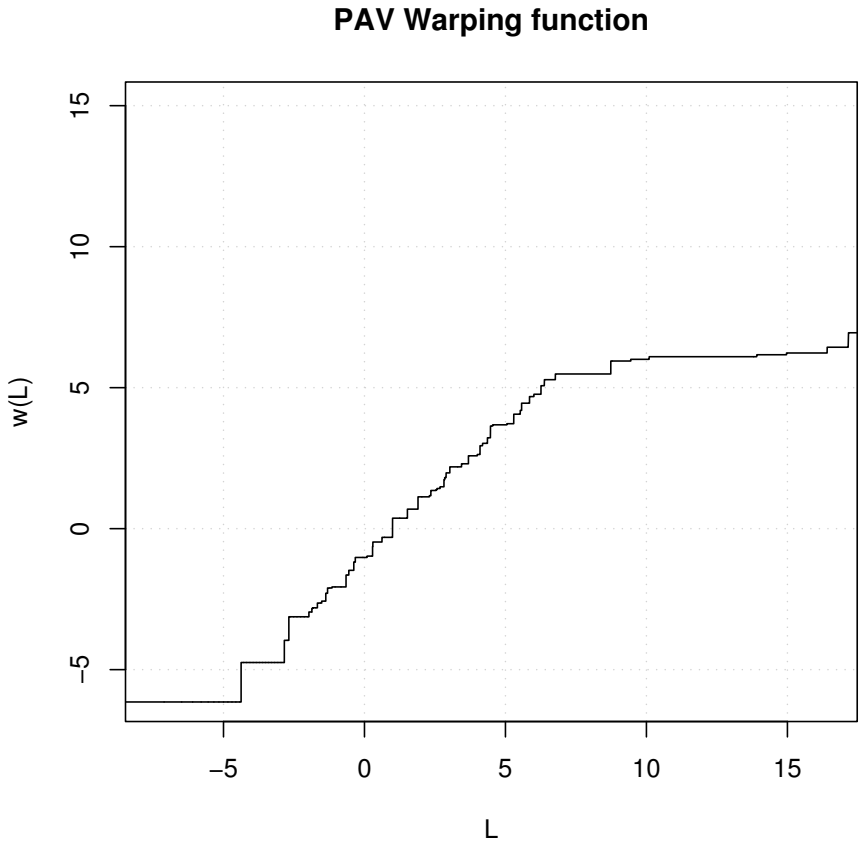

Fig. 3. The result of the PAV algorithm applied to the log-likelihood-ratio scores for which the score distributions were shown in Fig. 1.

procedure forms the application-independent generalization of the traditional measures $C_{\text{det}}^{\text{min}}$ and $C_{\text{det}} - C_{\text{det}}^{\text{min}}$.

As we shall further demonstrate with APE-curves below, the ability to do this discrimination/calibration decomposition is an important feature of the C_{llr} methodology. The ability to separate these aspects of detector performance empowers the designer of speaker detection systems to follow a divide-and-conquer strategy: First concentrate on building a detector with good discriminative ability, without having to worry about calibration issues. Then when you want to move on to practical applications, concentrate on also getting the calibration sorted out.

3.4 The APE-curve: Graph of the C_{llr} integral

The C_{llr}-integral, (7), is the integral of $C_{\text{det}}(\theta)$ over the application parameter θ. We will now show that this integral can be visualized in a powerful graph. The essential part of the integrand of (7) is the error probability

$$P_e(\theta) = \tilde{P}_{\text{tar}}(\theta)P_{\text{miss}}(\theta) + (1 - \tilde{P}_{\text{tar}}(\theta))P_{\text{FA}}(\theta). \tag{11}$$

Note that all of P_e, $\tilde{P}_{\text{tar}}$, P_{miss} and P_{FA} are functions of θ. The graph of P_e against θ forms the basis of the *Applied Probability of Error* (APE)-plot.

In Fig. 4 we show the APE-plot for our example system. Along the horizontal axis we have θ, which as explained before can be called the 'prior log odds'. Note that the horizontal axis of the APE-plot is the whole real line, but that we plot[18] only the interesting interval close to $\theta = 0$. The vertical axis is the error-rate axis, which takes values between 0 and 1. On these axes, we plot three curves: solid, dashed and dotted, which are respectively error-rates of the actual, PAV-optimized and reference systems. From these plots we can read a wealth of information:

The solid curve is $P_e(\theta)$ of (11). It shows the error-rate obtained (at each θ) when minimum-expected cost decisions are made with the log-likelihood-ratio scores $\mathcal{L}'_t$ as output by the system under evaluation. Note:

- The area[19] under the solid curve is proportional to C_{llr}, which can be interpreted as the *total actual error* over the spectrum of applications.
- The vertical dashed line at $\theta = -\log 9.9$ represents the traditional NIST DCF parameters, so that the solid curve at this point gives[20] the traditional *actual* C_{det}.

[18] Recall that both of the axes in DET-curves are also infinite and that there too, we plot only a selected region.

[19] The area is the analytically derived definite integral over the whole infinite θ-axis and not just the area under the visible part of the curve.

[20] The value of the solid curve is an *error-rate*, which is a scaled version of the *cost*, C_{det}, where the scaling factor is 1.09, as derived in footnote 14.

- The error-rate goes to zero for large $|\theta|$, in such a way that the C_{llr} integral exists (has a finite value).[21]

The dashed curve shows $P_e(\theta)$, but with scores $\mathcal{L}'_t$ replaced by $w(\mathcal{L}'_t)$ as found by the PAV algorithm.

- The area under the dashed line is proportional to $C_{\text{llr}}^{\text{min}}$, which can be interpreted as the *total discrimination error* over the whole spectrum of applications.
- The area between the solid and dashed curves represents the *total calibration error*.
- At the vertical line representing the NIST DCF parameter settings, $C_{\text{det}}^{\text{min}}$ can be read[22] from the dashed curve.
- The dashed curve has a unique global maximum, which is the equal-error-rate (EER). This maximum is typically located close to $\theta = 0$.

The dotted curve represents the probability of error for the reference detector, which does not use the speech input, basing its decisions only on the prior $\tilde{P}_{\text{tar}}$. As noted above, the reference detector outputs $\mathcal{L}'_t = 0$ for every trial. The error-rate of the reference detector is $P_e(\theta) = \min(\tilde{P}_{\text{tar}}(\theta), 1 - \tilde{P}_{\text{tar}}(\theta))$. Note here:

- The APE-plot scale does not show the maximum at $P_e = 0.5$.
- The area under the dotted curve is proportional to one (with the same scale factor as the areas under the other curves), and therefore represents the C_{llr}-value of the reference system.
- For $|\theta| \gg 1$, P_e goes to zero rapidly.
- For large negative θ we can observe that our example system performs *worse* than the reference detector!

The APE-curve is complementary to the traditional DET-curve. There is information, like the EER, that is duplicated in both curves, while some information displays better on the DET-curve, and other information better on the APE-curve. As a general rule, the DET-curve is a good tool for examining details of discriminative ability, while the APE-curve a a good tool for examining details of calibration. In addition, both curves have value as educational resources: As we know, the DET-curve demonstrates the error-tradeoff. The APE-curve demonstrates:

- The derivation of C_{llr} as an integral of error-rate over the spectrum of applications.
- The importance of the EER as an application-independent indicator of discriminative ability.
- As discussed in more detail below, C_{llr} has the information-theoretic interpretation of being the amount of information that is lost between the input speech and the final decisions. The APE-curve is therefore a graphical

[21] This holds, provided that $|\mathcal{L}'_t| < \infty$, for every trial t. If however the system does output even a single log-likelihood-ratio of infinite magnitude having the wrong sign, then the C_{llr} integral will evaluate to infinity.

[22] again subject to the scaling factor of 1.09.

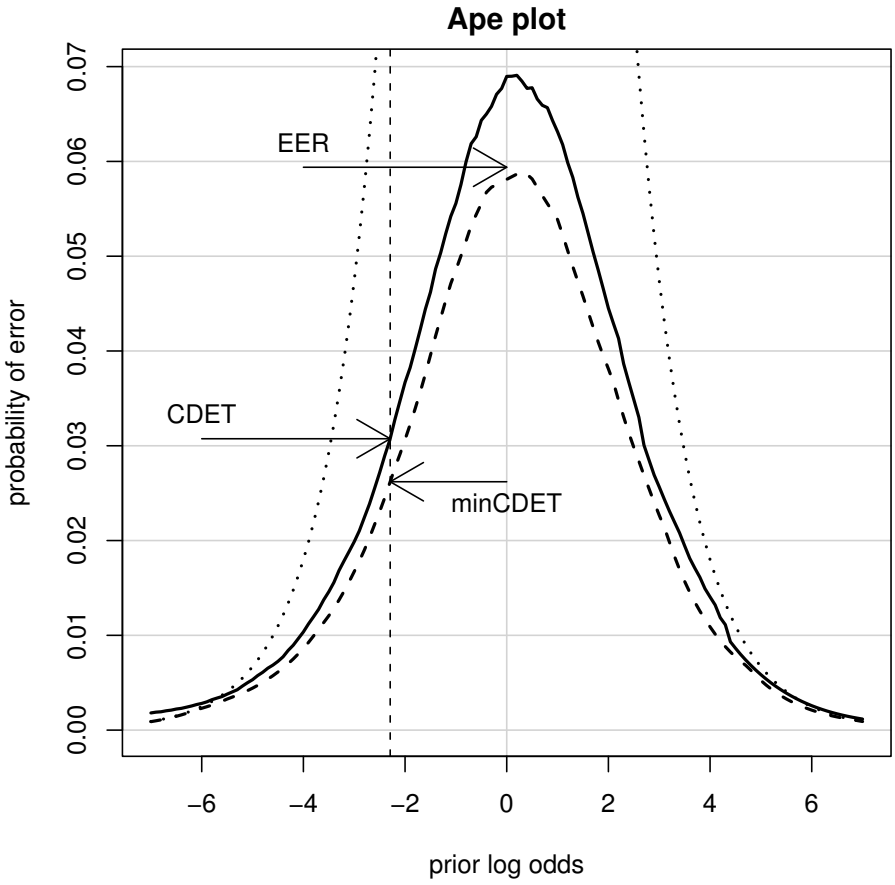

Fig. 4. APE-plot for our example system. Indicated are: $P_e(\theta)$ for observed $\mathcal{L}$ (solid curve), optimally calibrated $w(\mathcal{L})$ (dashed curve) and a reference detector (dotted curve).

demonstration of a relationship between information and error-rates—the more information you extract from the speech, the lower the error-rates will be.

Discussion. There is something interesting going on in the APE-curve around $\theta = 0$. On the one hand we see that P_e gives the biggest contribution to C_{llr} in this region. That would suggest that the task of the detector is hardest for $\theta \approx 0$, including the task of calibration. On the other hand, the benefit with respect to the reference detector is also the biggest in this region. Another way of phrasing this is that it seems that the information can be extracted from the speech signal most effectively when $\tilde{P}_{\text{tar}} \approx 0.5$. For $|\theta| \gg 1$ there is already a lot of information in the prior, and it is difficult to add something useful by analyzing the speech signal, even though the probability of error is lower.

There is a further concern: it is also more difficult to accurately estimate error-rates when $|\theta| \gg 1$, because the absolute number of errors in these regions becomes small and eventually vanishes. So it seems the extreme regions of the APE-curve are regions where our detectors probably won't work so well, but also where we cannot estimate their performance accurately. In our APE-plots, we ignore these regions by not plotting them. This is just the same as is done with DET-curves. The horizontal and vertical axes of the DET-plot are infinite, but we always plot just a finite interesting region of this plot. Outside of this plot, the DET-curve becomes increasingly jagged, which is an indication of poor error-rate estimates.

The saving grace is that there are real-life effects that force reasonable applications to lie close to $\theta = 0$. There may certainly be applications where the prior P_{tar} becomes very small. But when things become scarce, their value generally increases. This means the cost of missing scarce events increases as the prior becomes smaller. Now recall (6) and note that a decrease in P_{tar} will be compensated for by an increase in C_{miss}, leaving θ approximately unchanged. Conversely, a similar argument shows that when $1 - P_{\text{tar}}$ becomes small, then C_{FA} would increase to compensate, again tending to keep θ roughly constant. It does therefore seem to make sense to concentrate our efforts to the benign central region of the APE-curve (or the corresponding region of the DET-curve).

3.5 Information-theoretic interpretation of C_{llr}

We have introduced C_{llr} as an integral of C_{det} over the spectrum of applications, but as hinted above, C_{llr} can be also be interpreted as a measure of *loss of information* [3].

Again, we will not do a rigorous information-theoretic derivation, but rather show informally how $1 - C_{\text{llr}}$ can be interpreted as the average information per trial (in bits of Shannon's entropy) that is gained by applying the detector. The information extracted by the detector from the speech is dependent on what is already known before considering the speech. This *prior knowledge* is encapsulated in the prior, P_{tar}. When $P_{\text{tar}} = 0$, or $P_{\text{tar}} = 1$, then there is

already certainty about the speaker hypothesis and the detector cannot change this—the posterior will also be 0 or 1. However, values of P_{tar} between these extremes leaves a degree of prior uncertainty, up to a maximum of 1 bit where $P_{\text{tar}} = 0.5$. This maximum prior uncertainty is the reference level against which C_{llr} measures the information that the detector can extract from the speech. The information extracted from the speech by the detector, namely $1 - C_{\text{llr}}$ bits per trial, behaves in the following way:

- A (theoretically) perfect detector has $C_{\text{llr}} = 0$ and therefore $1 - C_{\text{llr}} = 1$, so it extracts *all* the information for every trial, transforming the prior uncertainty to posterior *certainty* in every case.
- A good, well-calibrated, real-life detector has $0 < C_{\text{llr}} < 1$, extracting an amount of information somewhere between 0 and 1 bit per trial.
- The reference detector which does not process the input speech has $C_{\text{llr}} = 1$ and therefore extracts 0 bits of information from every trial.
- A very badly calibrated[23] detector can do worse than this, having $C_{\text{llr}} > 1$, therefore extracting a *negative* amount of information. The negative sign indicates that on average over the APE-curve, the detector under evaluation has a higher error-rate than the reference detector. In this case it is therefore detrimental to use the detector and it is obviously better not to use (or at least to go and re-calibrate) the detector, because one could do better by just using the reference detector.

3.6 Comparison of systems: DETs and APEs

Let us end this chapter with an example of the use of C_{llr} and APE-plots for comparing systems or conditions. This, in the end, is one of the key reasons to perform evaluations. To this purpose we use the data of two systems under evaluation of NIST SRE 2006 [4] which both may be called state of the art. The first system (which we have seen in earlier figures) consists of a single detector, the second system consists of the fusion of 10 separate detectors, of which the first system is one.

We further compare two evaluation conditions. The first condition includes trials with speech spoken in several languages, while the second condition has the subset of the trials where both speech segments are English.

We first look qualitatively at the DET-plot of three system/conditions in Fig. 5. Note how the DET warping of axes separates the three curves comfortably in the plot[24].

[23] It is only calibration problems that can cause $C_{\text{llr}} > 1$. If we remove calibration effects, considering only the discriminative ability of the detector, we find $0 \leq C_{\text{llr}}^{\text{min}} \leq 1$.

[24] With many different systems or conditions, the number of curves in a DET-plot is more often than not limited by the number of colours and/or line types. Also notice that the legend in the plot enumerates the curves in the same top-to-bottom order as the curves appear in the plot, i.e., according to the EER. (This practice is unfortunately not followed by all authors.)

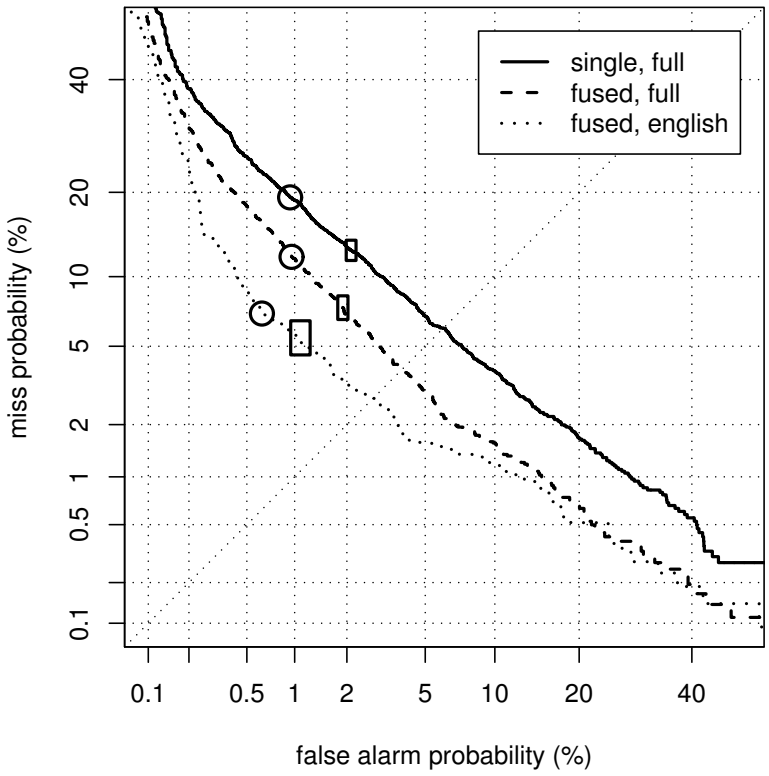

Fig. 5. A DET-plot for three system/conditions. From top to bottom: Single system, all trials; Fused system, all trials; and fused system, English trials. Notice that the upper and lower curve should not be compared with each other.

If we now inspect the curves more closely, we see that in terms of discrimination ability, the fused system performs favourably compared to the single system. Similarly we can conclude that, for the fused system, the English only trials were easier to discriminate than the whole collection of trials including several languages. (It does not really make sense to compare the upper and the lower curve, since both system and condition are different.) As for calibration, we can only conclude that for the NIST DCF the calibration was reasonable, and possibly better for the English only condition. We can finally observe that the lowest curve gets a bit noisy because a relatively low number of errors are made. For the English-only condition we have less than 30 target trial errors around $P_{\text{miss}} < 1.4\,\%$, so that if we apply George Doddington's 'rule of 30' [16] we find that for these low miss probabilities we are less than 90 % confident that the true P_{miss} is within 30 % of the observed P_{miss}.

We next look at the same systems evaluated on the same data, but depicted in APE-plots in Fig. 6. Here we have included a bar-graph of the C_{llr} and its decomposition into discrimination and calibration loss, expressed in bits. The scales of the figures are the same, so that values can be compared visually. We can observe that although the fused system has much better discrimination power than the single system, the calibration error is roughly the same. Similarly, restricting trials to only English has a bigger effect on the discrimination than on the calibration. From the APE-curves we can learn that there is still quite some calibration performance to be gained for the fused system, especially at $\theta = 0$. All systems/conditions seem to suffer from being 'worse than the reference system' at very low θ.

One difference between DET and APE is the way that inaccuracies due to the limited number of trials show up. The curve in a DET-plot usually becomes ragged at the ends due to the low number of errors involved, showing that at each end, respectively P_{miss} or P_{FA} is poorly estimated. The fact that this effect is visible on the plot is a consequence of the magnification of small probabilities by the probit scale used in the DET-curve. In the APE-curve we do not see these effects, because when either P_{miss} or P_{FA} is poorly estimated, their value on the vertical axis is also small. Since C_{llr} is the area under the APE-curve, we see that fortunately these inaccuracies contribute relatively little to the total C_{llr} integral. Having said this, we must also remark that the proportions of the numbers of target and non-target trials in a NIST evaluation typically is 1:10, which leads to almost optimum accuracy at the operating point defined by C_{det}—this may be observed from the roughly equal 95 %-confidence intervals in the DET-plot around C_{det}. This 1:10 ratio has the effect that the left-hand side of the APE-plot is somewhat less noisy than the right-hand side.

4 Conclusion

We reviewed and appreciated the traditional measures that the speaker recognition community uses to assess the quality of automatic speaker recognition systems. The detection cost function C_{det} measures the application-readiness

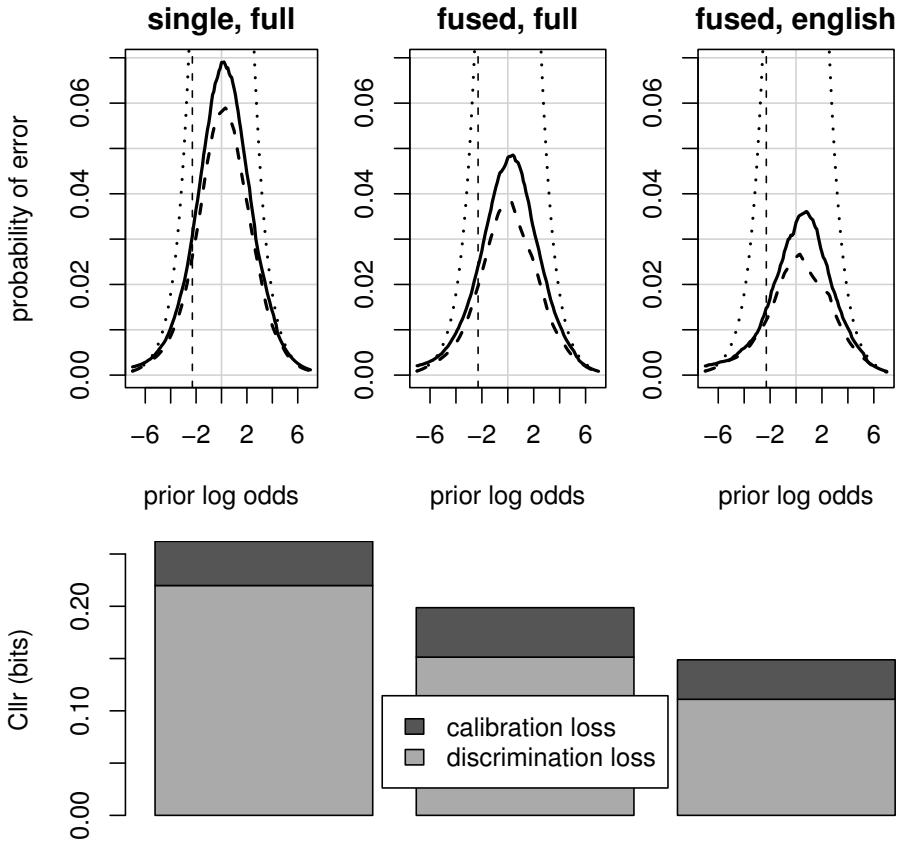

Fig. 6. APE-plots of the systems shown in Fig. 5. Note, that the graphs left and middle compare two systems, while the graphs middle and right compare two conditions.

of a system for a particular application-type as defined by the parameters P_{tar}, C_{miss} and C_{FA}. NIST deserves credit for defining the task and evaluation measure and the progress that this has stimulated in the field. In particular, concentrating on detection rather than identification; and using expected cost, rather than error-rate for evaluation have had far-reaching effects. Moreover, the DET-curve, with its warped axes, show very well the trade-off between P_{FA} and P_{miss}, and allow for direct comparison of discrimination ability of many different systems or conditions in a single graph. Again, NIST deserves credit for introducing this type of analysis in the community—indeed, gradually DET-plots are being applied in other disciplines. Finally, when calibration is not an issue, the traditional EER remains a good single-valued summary of the discriminative capability of a detector. The utility of the EER as summary of discriminative ability can be appreciated in different ways in the DET and APE-plots.

We have further shown the limitations of C_{det} and $C_{\text{det}}^{\text{min}}$, in the sense that although they do measure calibration, they do so only in an application-dependent way. Of course, the DET-plot and the EER do not measure calibration.

Next, we reviewed the advantages of working with *log-likelihood-ratios* instead of merely with scores. Perhaps the most important advantage is that users can then set their own decision thresholds, where the thresholds are dependent only on properties of the application and not on the properties of the speaker detector. Despite these obvious and well-known advantages, the use of log-likelihood-ratio outputs in speaker recognition has not been common, presumably because such likelihood-ratio outputs are in practice subject to calibration problems, and without being able to measure these calibration problems, researchers had no good way to even start tackling this problem.

Our most important contribution in this chapter is therefore the introduction of a methodology to *measure the quality* of log-likelihood-ratios via C_{llr}. Moreover, we paid special attention to the issue of calibration, by forming a discrimination/calibration decomposition of C_{llr}. The practical calculation of C_{llr} via (9) is no more complex[25] than the traditional P_{miss} and P_{FA} calculations. The calculation of $C_{\text{llr}}^{\text{min}}$ is somewhat more complex, because it involves the PAV algorithm, but fortunately implementations are available to researchers, see e.g. [3].

Finally, we showed that the new metric C_{llr} has the interpretation not only as an integral of error-rates over the spectrum of applications, but also as the average information loss between speech input and decisions. This relationship is graphically demonstrated by the APE-plot, which indeed, for analysis of calibration, forms a useful complement to traditional DET-plots.

In conclusion, looking towards the future, it was announced at the June 2006 workshop of the NIST Speaker Recognition Evaluation that NIST intended to include the new measure C_{llr} as the primary evaluation measure in future evaluations. We hope this will stimulate more research on the subject of calibration, which is an important factor of the design of speaker recognition systems.

[25] with due respect for some numerical accuracy issues

References

1. Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M.: The DET curve in assessment of detection task performance. In: Proc. Eurospeech 1997, Rhodes, Greece (1997) 1895–1898
2. Brümmer, N.: Application-independent evaluation of speaker detection. In: Proc. Odyssey 2004 Speaker and Language recognition workshop, ISCA (2004) 33–40
3. Brümmer, N., de Preez, J.: Application-independent evaluation of speaker detection. Computer Speech and Language **20** (2006) 230–275
4. : The NIST year 2006 Speaker Recognition Evaluation Plan. `http://www.nist.gov/speech/tests/spk/2006/index.htm` (2006)
5. et al., W.M.C.: Estimating and evaluating confidence for forensic speaker recognition. In: Proc. ICASSP. (2005)
6. et al., W.M.C.: Understanding scores in forensic speaker recognition. In: Proc. Odyssey 2006 Speaker and Language Recognition Workshop. (2006)
7. D. Ramos-Castro, J.G.R., Ortega-Garcia, J.: Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework. In: Proc. Odyssey 2006 Speaker and Language Recognition Workshop. (2006)
8. Brümmer, N., van Leeuwen, D.A.: On calibration of language recognition scores. In: Proc. Speaker Odyssey. (2006) submitted.
9. Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score normalization for text-independetn speaker verification systems. Digital Signal Processing **10** (2000) 42–54
10. Navrátil, J., Ramsawamy, G.N.: The awe and mistery of t-norm. In: Proc. Eurospeech. (2003) 2009–2012
11. van Leeuwen, D.A., Martin, A.F., Przybocki, M.A., Bouten, J.S.: NIST and TNO-NFI evaluations of automatic speaker recognition. Computer Speech and Language **20** (2006) 128–158
12. Swets, J.A.: Signal detection and recognition by human observers; contemporary readings. Wiley, New York (1964)
13. Green, D.M., Swets, J.A.: Signal Detection Theory and Psychophysics. Wiley, New York (1966)
14. Bernardo, J.M., Smith, A.F.M.: Bayesian Theory. Wiley, New York (1994)
15. DeGroot, M., Fienberg, S.: The comparison and evaluation of forecasters. The Statistician (1983) 12–22
16. Doddington, G.R., Przybocki, M.A., Martin, A.F., Reynolds, D.A.: The NIST speaker recognition evaluation—Overview, methodology, systems, results, perspective. Speech Communication **31** (2000) 225–254