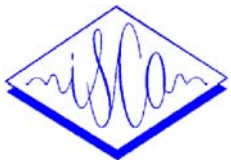

ODYSSEY04 - The Speaker and Language Recognition Workshop
Toledo, Spain
May 31 - June 3, 2004

Application-Independent Evaluation of Speaker Detection

Niko Brümmer

Spescom DataVoice, Stellenbosch, South Africa
nbrummer@za.spescom.com

Abstract

We present a Bayesian analysis of the evaluation of speaker detection performance. We use expectation of utility to confirm that likelihood-ratio is both an optimum and application-independent form of output for speaker detection systems. We point out that the problem of likelihood-ratio calculation is equivalent to the problem of optimization of decision thresholds. It is shown that the decision cost that is used in the existing NIST evaluations effectively forms a utility (a *proper scoring rule*) for the evaluation of the quality of likelihood-ratio presentation. As an alternative, a logarithmic utility (a *strictly* proper scoring rule) is proposed. Finally, an information-theoretic interpretation of the expected logarithmic utility is given. It is hoped that this analysis and the proposed evaluation method will promote the use of likelihood-ratio detector output rather than decision output.

1. Introduction

The goal of this paper is to motivate for a new evaluation methodology of speaker detection systems that will unify application of speaker detection for different uses. In particular we seek to unify the design and evaluation goals of speaker detection for *forensic* use and for *decisional* use. The NIST-type evaluation methodology is appropriate for decisional use, but is lacking for forensic applications, where the requirement exists to present suitably normalized likelihood-ratios rather than decisions [3][4][20]. We show how to satisfy both simultaneously.

2. Probability theory

Most of the material in this paper could be presented with the "orthodox" language and interpretation of probability theory that is customary in most of the speech engineering literature. But, we shall instead make use of a relatively unknown Bayesian interpretation of probability theory, the use of which is, once understood, compellingly attractive for applications such as speaker recognition.

The basic rules of probability are just the *product* and *sum* rules[1]. But there are different interpretations of probability which all share these *same* rules. This unfortunately often leads to confusion between different interpretations. Probability interpretations include:

- *Frequentist* (orthodox) statistics, where probabilities are taken as limiting relative frequencies in infinite repetitions of similar cases. Here prior and posterior probabilities are not used, only sampling probabilities (also called likelihoods.)
- Subjective (De Finetti school) *Bayesian* statistics, where probabilities are taken as personal belief. See e.g. [1].
- *Probability theory as logic*. This interpretation is also Bayesian, but probability is viewed as an objective representation of knowledge [2][22].

The main cause of these differences are the conceptual difficulties associated with specifying prior probabilities, leading in the one extreme to rejection of the use of priors.

It is the opinion of the author that in most of the speaker recognition literature, the frequentist interpretation is used. In this literature prior probabilities *are* put to limited use, but then most often in the frequentist sense where they can be estimated from data. Exceptions include some works in forensic speaker recognition [3][4][20] and also [5][18].

Here we motivate that for speaker recognition applications, *probability theory as logic* is most applicable. For a typical application, probability as relative frequency, particularly where priors are concerned, is problematic. Then also, we are building machines: We do not want the machines to have subjective beliefs. Our chosen interpretation is summarized below:

2.1. Probability theory as logic

All users of probability in speech processing are strongly encouraged to study the excellent book on this subject by Jaynes [2]. (It is also a good background for appreciating this work.) (For a tutorial overview see [22].) A short summary follows:

Probability has a much wider application than just frequencies in repetitive random situations: It can be used as a tool of *quantitative inductive reasoning,* which is reasoning in the face of *uncertainty,* where the uncertainty is *due to lack of knowledge*.

Cox [6] first showed that using the rules of probability is the *only consistent* way, in qualitative correspondence with common sense, of doing quantitative inductive reasoning [2]. The concept of "random variables" is not used here, but rather "unknown quantities". A probability distribution based on a well-defined, given state of knowledge is not *estimated* – it is *assigned*. (Unknown quantities are estimated.) If there are conceptual problems with the interpretation of knowledge, or practical calculation problems, we shall talk of *approximating* probability distributions. The approximation is to the ideal distribution that would be assigned (based on given knowledge) if we had enough skill and resources.

We shall use the Bayesian notation for probability distributions: $p(. \mid state\text{-}of\text{-}knowledge)$.

3. Definition of speaker detection

It will pay to define this well-known problem in very general terms:

[1] $P(A,B)=P(A|B)P(B)$ and $P(A+B)=P(A)+P(B)-P(A,B)$, where $A+B$ denotes logical disjunction.

An agent (human or machine) is faced with a situation in which a *course of action,* $a \in A \equiv \{a_1, a_2, \ldots, a_N\}$ must be chosen. This choice may be facilitated by the use of some data $x \equiv (d_1, d_2)$, where d_1 and d_2 are two segments of speech that were produced either by the same speaker (hypothesis H_1), or by two different speakers (hypothesis H_2). Our problem is to build a machine, the speaker detector, that calculates a function $w = w(x)$, on which the agent can base the choice of action. The agent (henceforth called the *user*) chooses action $a = a(w)$. The detector *summarizes* the *information* available in the speech.

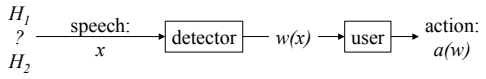

Figure 1: Detector use

The problem that developers of detector machines face is how to optimize the machine as to be maximally useful to any given user.

4. Evaluation

The question addressed here is how to evaluate speaker detection systems in order to encourage their design to be maximally useful in the above context. First we note *what* we can measure with an evaluation. Then we ask what is the *best* we can do? Next, we analyze current evaluation practice and motivate how this may be improved.

4.1. What does an evaluation measure?

An *evaluation* may be viewed as an *estimate* of how well a given speaker detection system will perform in actual future usage. "How well" is measured with a *utility function*. We start by considering a theoretical estimate of the utility of a detection system and in the next section show how this could theoretically be optimized.

In the evaluation context, we shall refer to input data $x = (d_1, d_2)$ as a *trial*. Let a *supervised trial* consist of a pair (h, x), where $h \in \{H_1, H_2\}$ is the hypothesis that is true for trial x. We consider first evaluation of action a, then find what form w should take for optimum action choice. Finally evaluation of w is considered.

Most detectors consist of a number of consecutive stages, but to facilitate analysis, we lump the stages together into 2 generic stages: the *extraction* stage $\sigma(x)$ and the *presentation* stage $v(\sigma)$:

$$w(x) \equiv v(\sigma(x)) \qquad (1)$$

(Examples of these stages, for different forms of detector, are considered below in section 6.) We further define function $\rho(.)$ to be the combination of the presentation and decision stages:

$$\rho(\sigma) = \rho(\sigma(x)) \equiv a(v(\sigma)) = a(w(x)) \qquad (2)$$

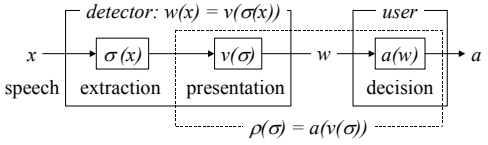

Figure 2: Detection stages

The *utility* for a single supervised trial is a real-valued function of the chosen action and of the *true* hypothesis: $u = u(a,h)$. We shall take as an estimate for utility, the *prior*[1] *expected utility*, conditioned on knowledge K : [2]

$$\begin{aligned} \hat{u} &\equiv E\{u \mid K\} \\ &= \int p(u \mid K) u \, du \\ &= \int p(\sigma \mid K) E\{u(\rho(\sigma), h) \mid \sigma, K\} d\sigma \end{aligned} \qquad (3)$$

where

$$E\{u(\rho(\sigma), h) \mid \sigma, K\} = \sum_{h \in \{H_1, H_2\}} P(h \mid \sigma, K) u(\rho(\sigma), h) \qquad (4)$$

is the *posterior* expected utility, given the output $\sigma = \sigma(x)$ of the extraction stage. The expectation $\hat{u}$ is dependent on: the user decision function $a(.)$, the detector $w(.)$, the utility $u(.)$ and on the knowledge K on which the probability distributions are conditioned.

Note further that the significance of using the expectation as an estimate is that it minimizes mean squared error: If K asserts that $p(u|K)$ is the distribution for a *single* unseen trial with utility u, then $\hat{u}$ is the estimate that minimizes $E\{(\hat{u}-u)^2 | K\}$. If furthermore K asserts[3] independence between trials, such that:

$$p(u_1, u_2, \cdots, u_N \mid K) = \prod_{i=1}^{N} p(u_i \mid K) \qquad (5)$$

for N unseen trials, then $\hat{u}$ also minimizes $E\{(\hat{u}-\bar{u})^2 | K\}$, where $\bar{u}$ is the true average over those N trials. In what follows we shall take independence between trials to hold.

In summary: The expectation $\hat{u}$ is a theoretical minimum-mean-squared-error estimate of future (or unseen) performance.

4.2. What is the best we can do?

We shall not fantasize about zero error probabilities here. The *best* any speaker detection system can do is limited by:

- The information in the input data. This is particularly relevant for short speech segments over poor and/or variable channels.
- The knowledge that can effectively be embedded in the machine, which includes knowledge induced from some quantity of development data.
- Conceptual and practical problems[4] in calculating probability distributions based on the above two information sources.

Often nothing can be done about the quality or quantity of input data. The knowledge embedded in the machine can be

[1] We call this expectation *prior* because it is the expectation before the data is given. This is to differentiate it from the *posterior* expectation introduced below.

[2] Note: (i) In the case of discrete u: $p(u|K) \equiv \Sigma_i P(u_i|K)\delta(u-u_i)$. The same applies to discrete σ. (ii) For multidimensional σ, $\int d\sigma$ is understood to denote a multidimensional integral.

[3] With probability as logic, independence need not be considered an "assumption". If our knowledge about the problem is unchanged if trials are arbitrarily *exchanged*, then this state of knowledge gives independent distributions [1][2].

[4] Conceptual problems are mostly encountered in converting knowledge into (prior) probabilities. As noted, this is the source of most of the problems in statistics. Practical problems include performing calculations such as integration over multidimensional spaces, when calculating expectations and marginal distributions.

improved, at a cost, e.g. by obtaining and using more development data. Thirdly, conceptual and practical solutions may be found to overcome some of the last class of limitation.

The evaluation methods discussed here effectively take all three of these issues into consideration, but we start by considering the last point: With a given quality of input data, and with a given state of knowledge *K* that can be built into the detector, what is the best we can do? A theoretical answer can be deduced trivially from equations 3 and 4:

The prior expected utility $\hat{u}$ can be maximized, under the *constraint* of a fixed extraction stage (i.e. the function $\sigma(x)$ is given), and with respect to varying the function $\rho(\sigma)$, by maximizing the posterior expectation for every value of σ. That is, the maximizing function $\rho(.)$ is[1]:

$$a = \rho(\sigma) = \arg\max_{a' \in A} \sum_{h \in \{H_1, H_2\}} P(h \mid \sigma, K) u(a', h) \qquad (6)$$

In the special case where we let $\sigma(x) \equiv x$, there is no constraint, and the optimization is global.

This establishes that if we could calculate the *posterior* $P(h|\sigma,K)$, then eq.6 forms an optimal decision function, without further difficulty: We have reduced this optimization problem to that of calculating the posterior. This result holds very generally, for different decision sets *A* and for different forms of utility $u(.)$.

What is the significance of the posterior? The input *x* or the output of the extraction stage σ carries *information* about the true hypothesis *h*. The posterior *presents* this information in a form that is most useful for subsequent decision making. What role does the extraction stage play? If we could directly calculate the *x* - posterior it would yield a greater $\hat{u}$, (no constraint) than the σ - posterior. However, since the dimensionality of σ is typically much smaller than that of *x,* the presentation of the information in σ is easier. The greater amount of information[2] in *x* does not help us unless we can present it in a useful way.

It is further important to note that this optimality is only achieved when the probability distributions that form the expectations are based on the *same* state of knowledge on which the optimizing posterior is based. The best any developer of a detector can do is to optimize the detector based on the knowledge *K* at his disposal. Knowledge *K* will be partly based on a quantity of development data. If a detector thus optimized is evaluated by expectations based on a different state of knowledge, which is partly based on some new data, it will no longer be optimal. The amount of utility lost in this way depends on how much the new data changes the probability distributions.

The final step in answering the question of what form *w* should take, is to note that the posterior $P(h|\sigma,K)$ can be formed (by Bayes' rule) from a *likelihood-ratio* and the *prior distribution* $P(h|K)$. In order for the detector to be as application-independent as possible, we take it as the responsibility of the user to supply this prior[3] and we shall take state-of-knowledge *K* to subsume this prior as given:

$$(P_1, P_2) \equiv \big(P(H_1 \mid K), P(H_2 \mid K)\big) \qquad (7)$$

An ideal detector output form is therefore the likelihood-ratio:

$$w(x) = R_\sigma\big(\sigma(x)\big) \equiv \frac{p(\sigma(x)|H_1,K)}{p(\sigma(x)|H_2,K)} \qquad (8)$$

which via Bayes' rule would give the posterior:

$$r_1 \equiv P(H_1 \mid \sigma(x), K) = \frac{w(x)}{w(x)+\frac{P_2}{P_1}} \qquad (9)$$

where we use the short-hand $r_i \equiv P(H_i|\sigma,K)$. It is up to the designer of the detector to define $\sigma(.)$. The user is unaware of this detail and also of *K*: The user sees the detector simply as a function $w(x)$, from which, ideally, the posterior r_1 can be obtained. The posterior empowers the user to choose his own output set *A,* and his own utility $u(.)$ and then to apply eq.6, to decide on action *a*. We rewrite eq.6 from the user's view: The ideal decision would be $a = B\{r_1, u(.)\}$, where we define:

$$B\{q_1; u(\cdot)\} \equiv \arg\max_{a' \in A}\big[q_1 u(a', H_1) + q_2 u(a', H_2)\big] \qquad (10)$$
$$q_2 \equiv 1 - q_1$$

where (q_1,q_2) is a probability distribution for *h*. We group equations 8-10 as follows:

- Equations 8 and 9 form the *inference* stage: This is the act of summarizing the total information about *h* that is obtained from *K* and *x* (without making any decisions). This summary is in the form of a posterior distribution.
- The *decision* stage (eq.10) is known as the *Bayes criterion*.

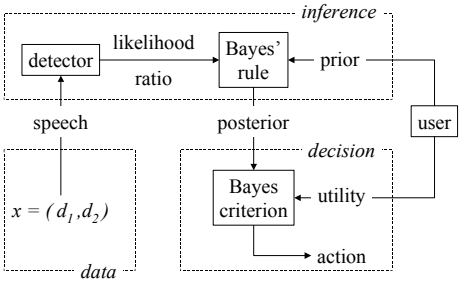

Figure 3: Application-independent detector

Note that this result is not new or indeed exclusively Bayesian. The *Bayes criterion* is accepted by all three of the probability schools mentioned above, [7][4] [1][5] [2] and is well-known in the speech engineering literature, see e.g. [21]. But why is the desirability of using a posterior in a decision problem not more widely recognized in practice? Probably because:

A. of the orthodox legacy which has conceptual difficulty with interpretation of the prior and posterior as a relative frequency.

B. it *is* practically difficult to calculate a posterior.

---

[1] To strictly make this a function, a disambiguation rule is needed in cases where there is more than one maximizing *a'*.

[2] There is more information about *h* in *x* than in $\sigma(x)$. This is formally stated by the *data processing inequality,* in terms of mutual information: $I(h;x) \geq I(h;\sigma)$ [19].

[3] If the user has no relevant knowledge, $(P_1,P_2) = (1/2,1/2)$ is *assigned*. This is an objective statement of the ignorance of the user. The fact that this may be far from the relative frequency of occurrence in future use presents no problem with probability as logic [2][22].

[4] According to [2], decision theory finally forced orthodox statistics to admit the use of some kind of "prior weighting", without the use of which *admissible* decisions could not be made [7]. This led to wider acceptance of Bayesian methods.

[5] In [1], the basic probability rules are derived axiomatically via the use of utility and the Bayes criterion..

C. of the question of how to evaluate the "quality" of a posterior.

Problem *A* is addressed by adopting *probability theory as logic*. Below, we show that the difficulty *B* is equivalent to that of setting decision thresholds. Finally we address *C* in the hope that it will help to stimulate solutions to *B*. (All problems in speaker recognition are difficult, but the more research they attract, the better they are solved.)

As an example of presentation of information to unskilled users (lacking knowledge *K*), consider how some weather forecasts are given: *Probabilities* for certain events are given, not *decisions*. A weather forecaster cannot make decisions for every member of the public. Each user of the forecast has his own (implicit or explicit) utility $u(.)$, which can be used together with the forecast probabilities to make decisions.

4.3. Evaluation practice

The problem of optimizing expected utility is different to the problem of estimating utility for a given system. To gain insight into the latter, we expand $\hat{u}$ in a different way:

$$\begin{aligned}\hat{u} &= E\{u \mid K\} \\ &= \sum_{i\in\{1,2\}} P_i E\{u \mid H_i, K\} \\ &\equiv P_1\hat{u}_1 + P_2\hat{u}_2\end{aligned} \qquad (11)$$

Where $\hat{u}_i$ is defined to be the expectation conditioned on H_i. If the $\hat{u}_i$ are obtained separately in this way, then $\hat{u}$ can be calculated for any given prior (P_1,P_2). This decomposition can be generalized to:

$$\tilde{u} \equiv P_1\tilde{u}_1 + P_2\tilde{u}_2 \approx \hat{u} \qquad (12)$$

where $\tilde{u}$ and $\tilde{u}_i$ are practical estimates for utility. An example would be the average utility over a database of supervised trials:

$$\tilde{u}_i = \frac{1}{\|D_i\|}\sum_{x\in D_i} u\big(a\big(w(x)\big), H_i\big) \approx \hat{u}_i \qquad (13)$$

where D_i is a set of trials for which H_i is true. The expected value $\hat{u}$ is a theoretical estimate for the average over unseen data. The average over given data is an empirical estimate for the average over unseen data. The average over given data is also an approximation to the expected value. The quality of the approximation depends on the relationship between *K* and the data. If *K* contains little knowledge other than this data, then the approximation will be good.

This approach is followed in the NIST evaluations [8][9], where equations 12 and 13 are applied: (D_1,D_2) is the evaluation database and (P_1,P_2) is a synthetic prior (different to the frequency of occurrence of H_1 vs H_2 in the set of evaluation trials). Estimate $\tilde{u}$ is based on two distinct sources of knowledge: the evaluation data and an independently specified prior, which is representative of an envisaged application.

The evaluator can be viewed as a *user* of the detection system and just as in the case of other users, the evaluator supplies the prior.

5. Evaluating quality of decision

Since the object of the whole process is to make decisions, it is natural to evaluate quality of decisions via a *detection cost* function. (The utility is then the negative of this cost: To optimize utility, cost must be minimized.)

For the NIST evaluations, it is required that systems output a real *detection score* as well as a *decision* for each test. The score is not formally used for evaluation, but DET [10] curves are plotted for inspection and discussion.

The formal evaluation is made by applying equations 12 and 13 with the utility defined below: Usually only two courses of action are considered: $a \in \{accept, reject\}$. The utility function is:

$$u_D(a,h) \equiv \begin{bmatrix} 0, & (a,h) = (accept, H_1) \\ 0, & (a,h) = (reject, H_2) \\ -c_{miss}, & (a,h) = (reject, H_1) \\ -c_{fa}, & (a,h) = (accept, H_2) \end{bmatrix} \qquad (14)$$

6. Detection practice

In practice, detectors cannot calculate the ideal likelihood-ratio of eq.8. We consider two practical detector types:

TYPE I:

- *Extraction stage:* outputs an amorphous real score $\sigma = s = s(x)$ of which it can only be said that larger scores favour H_1 and smaller scores H_2.
- *Presentation stage:* outputs a decision $w = \Delta(s)$, where $w \in \{accept, reject\}$. The decision is most often made by comparison of the score with a single pre-set threshold. The threshold is chosen to optimize a specific utility in a specific envisaged application.
- *User:* takes the detector output as is: $a = w$.

TYPE II:

- *Extraction stage:* This could be any $\sigma = \sigma(x)$, but a real score $\sigma = s(x)$ as in *type I,* is easiest to work with.
- *Presentation stage:* outputs an approximation to the ideal likelihood-ratio:

  $$w = \tilde{R}_\sigma(\sigma) \equiv \frac{p(\sigma|H_1,K')}{p(\sigma|H_2,K')} \approx R_\sigma(\sigma) \equiv \frac{p(\sigma|H_1,K)}{p(\sigma|H_2,K)} \qquad (15)$$

  where K' is defined to be the implicit state of knowledge on which the practically calculated likelihood-ratio is effectively conditioned.
- *User:* supplies a prior (P_1,P_2) and applies Bayes' rule to calculate the approximate posterior:

  $$\begin{aligned} q_1 &\equiv P(H_1 \mid \sigma, K') = \frac{\tilde{R}_\sigma(\sigma)}{\tilde{R}_\sigma(\sigma)+\frac{P_2}{P_1}} \\ &\approx r_1 \equiv P(H_1 \mid \sigma, K) = \frac{R_\sigma(\sigma)}{R_\sigma(\sigma)+\frac{P_2}{P_1}} \end{aligned} \qquad (16)$$

  where we have taken both states of knowledge, *K* and *K′,* to agree on the prior. Then the user applies the Bayes criterion (eq.10) to choose a course of action: $a = B\{q_1, u(.)\}$, for any utility $u(.)$.

Detectors entered into the NIST evaluations are usually of *type I.* Below we show that *type I* can be transformed to the more generally useful *type II*:

6.1. Transformation from *type I* to *type II*

Most often the decision function $\Delta(s)$ is implemented by comparing score *s* to a single pre-set threshold *t*. The two detection error probabilities as functions of this threshold are:

$$P_{fa}(t) \equiv P(accept \mid t, H_2, K) = \int_t^{\infty} p_s(s \mid H_2)\, ds$$
$$P_{miss}(t) \equiv P(reject \mid t, H_1, K) = \int_{-\infty}^{t} p_s(s \mid H_1)\, ds \qquad (17)$$

where we use, for convenience, the definition: $p_s(s|h) \equiv p(s|h,K)$. Given the prior (P_1,P_2), the error probabilities can be used to express the expected decision cost:

$$\hat{c}(t) \equiv -E\{u_D \mid t, K\} = c_{fa} P_2 P_{fa}(t) + c_{miss} P_1 P_{miss}(t) \qquad (18)$$

To find the minimizing threshold, we differentiate:

$$\frac{d}{dt}\hat{c}(t) = -c_{fa} P_2 p_s(t \mid H_2) + c_{miss} P_1 p_s(t \mid H_1) \qquad (19)$$

Setting the derivative to zero gives the well-known solution:

$$R_s(t) \equiv \frac{p_s(t \mid H_1)}{p_s(t \mid H_2)} = T \equiv \frac{P_2}{P_1}\frac{c_{fa}}{c_{miss}} \qquad (20)$$

where $R_s(.)$ is defined to be the score likelihood-ratio and where T is defined to be a new threshold in the likelihood-ratio domain. Now if $R_s(.)$ is everywhere strictly monotonically increasing, $R_s(t) = T$ has a single solution for t, which is the minimizing score threshold. (This is the assumption that effectively justifies the use of a single score threshold.)

If monotonicity is taken to hold: The optimal score threshold t is then formally obtained by inversion of $R_s(t) = T$. But if we don't have the function $R_s(.)$, how would we set a practical threshold t, given T? An "application-ready" detection system must come equipped with a threshold t, that is suitable for the application conditions represented[1] by a given T. The developer of the system will in general obtain t by application of a procedure $\varphi(.)$, involving a quantity of calibration data D, so that $t = \varphi(T, D)$. Now if the goal of this procedure is to minimize expected cost $\hat{c}(t)$, this procedure is also an approximation to the formal minimizing solution[2]:

$$\tilde{R}_s^{-1}(T) \equiv \varphi(T, D) \approx \arg\min_t \hat{c}(t) = R_s^{-1}(T) \qquad (21)$$

Now if $\varphi(T,D)$ can be evaluated at one value of T, it can also be evaluated for a range of values. In this way the calibration data can be used to map out an approximation to the likelihood-ratio: $\tilde{R}_s(.)$. If this mapping is done so that exactly $\tilde{R}_s(t) = T$, then thresholding $\tilde{R}_s(s)$ at T, will produce *identical* decisions to thresholding score s at t. But now the user can set his own T-threshold at will, instead of relying on a pre-set, built-in t-threshold, which is only optimal for one value of T. Note also that as long as the score transformation $w = \tilde{R}_s(s)$ is strictly monotonically increasing, the DET curves of s and w will be identical. In summary: the range of application of the detector has been improved, without sacrificing performance. See figure 4.

If monotonicity of $R_s(.)$ is not taken to hold: Then the decision function $\Delta(s)$ will have a more general form, which effectively has *multiple* thresholds. A similar but more tedious analysis shows that the set of optimizing thresholds is $\{t|R_s(t)=T\}$. Note: for different values of T, the cardinality of this set may be different. The set can even be empty: This happens if $R_s(.)$ is bounded. For values of T outside the range of $R_s(.)$, the requirements are effectively too strict for this quality of detector and the optimal strategy is then to always make the same decision, independently of the score.

The point here is that, also for this more complicated case: If you have a procedure for optimizing a decision function from score to decision, then you effectively have the means to map out an approximation to $R_s(.)$, which can be used to effect a more generally applicable detector of identical decision performance. (Note that in the non-monotonic case, the DET curve of $\tilde{R}_s(s)$ will be different from that of s in places, but if the approximation to the likelihood-ratio is good, this change will be an improvement.)

In summary: We have shown that we can transform a detector of *type I*, having extraction stage $s(x)$ and an optimized decisional presentation stage, to one of *type II*, that presents $\tilde{R}_s(s)$.

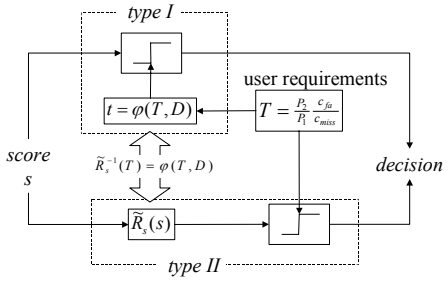

Figure 4: Relationship between types *I* and *II*

6.2. Note on score vs likelihood-ratio

We highlight the relationship between the *score* and the *score likelihood-ratio*: In most current speaker detection systems, the score s is a scaled and shifted[3] approximation to calculating the *feature log-likelihood-ratio:* $\log R_\phi(\phi) \equiv \log \frac{p(\phi \mid H_1, K)}{p(\phi \mid H_2, K)}$, where $\phi = \phi(x)$ is the total result of the front-end processing of both speech segments $x = (d_1, d_2)$. Note that if this approximation were exact, that is if $s(x) = \log R_\phi(\phi(x))$, then also[4] $s(x) = \log R_s(s(x))$. Since this is not the case in practice, we have $s(x) \neq \log R_\phi(\phi(x)) \neq \log R_s(s(x))$. Then the additional step of calculating $\tilde{R}_s(s)$ is needed to improve the *quality of presentation*. Good approximation of R_s is easier than approximation of R_ϕ, since the former is a function of the one-

[1] Note T represents the requirements, because if $\hat{c}(t)$ is minimized at $t = t^*$, then $\hat{c}'(t) \equiv c'_{fa}P'_2 p_{fa}(t) + c'_{miss}P'_1 p_{miss}(t)$ is also minimized at t^*, as long as this ratio is the same: $\frac{P'_2}{P'_1}\frac{c'_{fa}}{c'_{miss}} = T \equiv \frac{P_2}{P_1}\frac{c_{fa}}{c_{miss}}$.

[2] This relationship between $R_s(.)$ and D exists because the former is by definition based on knowledge K, which is partly based on data D. If K contains little knowledge other than D, this approximation is good.

[3] Scaling and shifting is a side-effect of some score normalization schemes such as T-norm [16].

[4] This can be shown by noting that $R_\phi(.)$ and $R_s(.)$ are both *sufficient statistics* for h [17].

dimensional score, while the latter is a function of the multi-dimensional features.

7. Evaluating quality of inference

By restricting attention to detectors of *type II*, we can shift the focus of evaluation away from *decisions,* to evaluation of the *likelihood-ratio* or of the *posterior*. This can also be done via a utility function. Since the evaluator supplies the prior (see section 4.3), there is a given one-to-one mapping between the likelihood-ratio and the posterior via Bayes' rule (eq.16). This means there is also a one-to-one mapping between utilities for likelihood-ratio w and for posterior q_1 :

$$\begin{aligned} u(q_1,h) &= u(\tfrac{w}{w+\frac{P_2}{P_1}},h) \\ &= u'(w,h) = u'(\tfrac{P_2}{P_1}\tfrac{q_1}{(1-q_1)},h) \end{aligned} \qquad (22)$$

This means that for comparison of different systems, at a given prior, the two utilities are equivalent. Here we present a way of evaluating the *posterior*. By definition (section 4.2 and figure 3), we are therefore evaluating *quality of inference*. Below we examine ways of evaluating the quality of approximation to the ideal posterior:

7.1. Scoring rules

How can the quality of a probability distribution be evaluated? Meteorologists have long used the following solution [11][12] (presented here in simplified form):

A weather forecaster uses all available data and to the best of his knowledge calculates (and therefore believes) the probability for rain tomorrow is r_1. How does one motivate him to actually present r_1 and not some other probability q_1, which he may think would be better received? One structures his reward, based on his presentation q_1 and on whether it actually rains (H_1), or does not rain (H_2), such that his own expectation of his reward is maximized if he reports what he believes: $q_1 = r_1$. This kind of reward is just a utility function (that evaluates probability distributions rather than decisions). The weatherman has the expectation of reward of :

$$\begin{aligned} &E\{u(q_1,h)\,|\,r_1\} \\ &= r_1\,u(q_1,H_1) + (1-r_1)u(1-q_1,H_2) \end{aligned} \qquad (23)$$

A utility for a probability distribution is called a *proper scoring rule* if this expectation is maximized at $q_1 = r_1$. It is *strictly* proper if it is maximized *only* at $q_1 = r_1$. There is an infinity of scoring rules that are proper or strictly proper [1][13][14]. We shall consider two families of scoring rules: *decisional scoring* and *logarithmic scoring*. The former is *proper* and develops naturally out of the NIST-type of detection cost. The latter is *strictly proper* and has many desirable properties [1][13].

7.2. Decisional scoring of posterior

Let the posterior obtained from a detector of *type II* be q_1, which is an approximation to the ideal posterior r_1 . We can effect evaluation of the quality of this approximation, via decision cost, by incorporating the Bayes criterion into a new utility function[1]:

$$\begin{aligned} u_B(q_1,h) &\equiv u_D\big(B\{q_1;u_D(.)\},h\big) \\ &= \begin{bmatrix} 0, & h=H_1, q_1 \ge C \equiv \frac{c_{fa}}{c_{fa}+c_{miss}} \\ 0, & h=H_2, q_1 < C \\ -c_{fa}, & h=H_2, q_1 \ge C \\ -c_{miss}, & h=H_1, q_1 < C \end{bmatrix} \end{aligned} \qquad (24)$$

where we have defined a *cost coefficient, C*. We have effectively removed the decision stage from the detector and built it into the new utility. By the analysis in section 4.2, the expectation of any utility $u(.)$ is maximized at $B\{r_1, u(.)\}$, therefore the expectation of u_B is maximized if $q_1 = r_1$, which shows that eq.24 is a *proper scoring rule*. This rule is a member of the family of *decisional scoring rules* [13].

By showing the equivalence of detector types *I* and *II,* and by showing that the NIST-type detection score leads to a proper scoring rule, we have therefore established that : *NIST-type evaluation does indeed implicitly evaluate quality of inference*.

But is this the best way of measuring quality of inference? The disadvantage of decisional scoring is that it is not *strictly* proper: It is possible to maximize expected score in ways other than by $q_1 = r_1$: It is maximized as long as q_1-C has the same sign as r_1-C. Suppose now that we manage to actually construct a detector that gives optimum q_1 , for every x, for a given C, then it may no longer be optimal for a different C. By evaluating with a specific cost coefficient C, we may be encouraging detection system design practices that are not optimal for other applications with very different cost coefficients.

This dependence on cost may be avoided by using a *strictly* proper scoring rule, because by definition, when this is optimized so that $q_1 = r_1$, for every x, then any other proper scoring rule will *also* have been optimized. Indeed when this is the case, any utility can be optimized via the Bayes criterion. Do note:

- For brevity of notation, we used $q_1 = P(H_1|\sigma(x),K')$ which hides the dependence on $\sigma(.)$. The optimization in the above analysis must be understood to be subject to the constraint that q_1 is a function of x only through $\sigma(x)$.
- In practice the optimum at $q_1 = r_1$ will be difficult to reach, even in constrained optimization. The above optimality is therefore only a theoretical limit.
- As pointed out before, the optimality is with respect to knowledge K. A detection system can only be optimized relative to given knowledge. Even if an optimum with respect to K is reached, this may not be optimal in a different situation where new data and knowledge is available.

7.3. Logarithmic scoring

We shall use the following logarithmic, strictly proper scoring rule:

$$u_{\log}(q_1,h) \equiv \begin{bmatrix} \log_2 \frac{q_1}{\gamma_1}, & h=H_1 \\ \log_2 \frac{q_2}{\gamma_2}, & h=H_2 \end{bmatrix} \qquad (25)$$

[1] We have included here an arbitrary disambiguation rule such as mentioned in the footnote to equation 6.

where $(\gamma_1,\gamma_2) \equiv (\gamma(H_1),\gamma(H_2))$ is a reference distribution for *h,* to be specified below; and where *(q_1, q_2)* is the approximate posterior distribution obtained from the detector. The logarithm base is a scaling factor, which is chosen here to give units of information-theoretic *bits*. Note when H_i is true and:

- if $q_i = \gamma_i$, then $u_{log} = 0$. (The posterior gives no information relative to γ_i.)
- if $q_i > \gamma_i$, then $-log_2\gamma_i > u_{log} > 0$. (The relative posterior gives information in the correct sense.)
- if $q_i < \gamma_i$, then $-\infty < u_{log} < 0$. (The relative posterior gives information in the wrong sense.)

A feature of this utility is that gambling is very strongly discouraged. By presenting a log-likelihood-ratio of large magnitude, the system can earn a positive utility of at most $-log_2\gamma_i$, if the sense is correct. But a wrong sense will effectively result in disqualification because of a very large negative utility. It is in the interest of the evaluated system to not profess to have more certainty than it really has. In the light of this *disqualification effect,* it may be wise to not adopt a monotonic transformation from score to likelihood-ratio. To be safe, a bounded, non-monotonic transformation should be used, where extreme scores (at both ends of the scale), outside the range seen in the development data, should transform to a neutral likelihood-ratio of one.

We have used the words *information* and *certainty* here: Indeed, it is known [1][2][14] that information-theoretic quantities, based on *Shannon entropy* [15], develop naturally from expectations of logarithmic utility:

7.4. Information theoretic interpretation

We do the following analysis with respect to two different states of knowledge:

- *K,* as before, is the ideal state of knowledge with respect to which the theoretical expectation of utility (eq.3) is taken. We use a shorthand notation where *r(.)* is used for distributions conditioned on *K*: $r(X \mid Y) \equiv p(X \mid Y, K)$, and with the usual convention that different distribution functions are differentiated by their arguments.
- *K′* is an implicit state of knowledge, defined to be that on which the distributions that are actually calculated by the detector are based. For distributions conditioned on *K′* we use: $q(X \mid Y) \equiv p(X \mid Y, K')$.
- The prior agrees for both states of knowledge: $q(h) \equiv r(h)$. (The prior is taken to be given by the user, which is the evaluator here. See section 4.3.)

With some manipulation we can express the expected logarithmic utility as follows:

$$\begin{aligned} \hat{u}_{\log} &\equiv E\{u_{\log}(q_1,h) \mid K\} \\ &= U_{ref} + U_{data} - U_{extr} - U_{pres} \end{aligned} \qquad (26)$$

where these terms can be specified in terms of the well-known information-theoretic quantities *divergence* and *mutual information* [19] :

$$U_{ref} \equiv D\{r(h);\gamma(h)\} \qquad (27)$$

$$\begin{aligned} U_{data} &\equiv I\{h;x \mid K\} \\ &\equiv E\{D\{r(h \mid x);r(h)\} \mid K\} \end{aligned} \qquad (28)$$

$$\begin{aligned} U_{extr} &\equiv I\{h;x \mid s,K\} \\ &\equiv E\{D\{r(h \mid x);r(h \mid s(x))\} \mid K\} \end{aligned} \qquad (29)$$

$$U_{pres} \equiv E\{D\{r(h \mid s);q(h \mid s)\} \mid K\} \qquad (30)$$

where we took the extraction stage to calculate a real score $s = s(x)$. The *Kullback-Leibler divergence* $D\{.;.\}$ between two distributions for *h* is defined as:

$$D\{\alpha(.);\beta(.)\} \equiv \sum_{h\in\{H_1,H_2\}} \alpha(h)\log_2\frac{\alpha(h)}{\beta(h)} \qquad (31)$$

The divergence is non-negative, and becomes zero if and only if $\alpha(.) \equiv \beta(.)$. We also need the *entropy* of the prior distribution *r(h)* which can be defined as[1] :

$$\begin{aligned} H\{h \mid K\} &\equiv 1 - D\{r(h);\gamma_0(h)\} \\ \gamma_0(h) &\equiv \tfrac{1}{2} \end{aligned} \qquad (32)$$

This entropy, or the *uncertainty* about *h* given by the prior*,* ranges from a minimum of zero when either hypothesis is certain, to a maximum of one when $r(.) \equiv \gamma_0(.)$.

Note:

U_{ref}: is an offset independent of *q(.)*. It is therefore unimportant for comparisons (at a fixed prior) between different systems. In practice, this term is under the control of the evaluator who specifies both *γ(h)* and *r(h)*. The choices for *γ(.)* determine the interpretation of the utility. Two possible choices are:

- If $\gamma(h) \equiv r(h)$, then $U_{ref} = 0$. This choice effects evaluation of the posterior *q(h|s)* relative to the prior *r(h)*.
- If $\gamma(.) \equiv \gamma_0(.)$, then $U_{ref} = 1 - H\{h|K\}$ which is the change in uncertainty given by having the prior *r(h)* rather than the maximally uncertain state of knowledge given by $\gamma_0(h)$. Here U_{ref} is the amount of information supplied by the user.

U_{data} $= I\{h;x|K\}$*:* is the *mutual information* between *x* and *h,* where $0 \le U_{data} \le H\{h|K\}$. It is the expected decrease in uncertainty about *h* that can be given by the data. The maximum would be reached only if the data could always give certainty about *h.* Note this term is also independent of *q(.)* .

U_{extr} $= I\{h;x|s,K\}$*:* is the *conditional* mutual information that *x* gives about *h*, in addition to that already given by *s*, where $0 \le U_{extr} \le U_{data}$. This is the amount of information lost in the extraction stage. This term vanishes for *perfect extraction*, when *s(x)* is a *sufficient statistic* [17] of *x* for *h*.

U_{pres}: is the expected *divergence* between the ideal and the calculated posterior distributions, where $0 \le U_{pres} \le \infty$. It vanishes only for *perfect presentation*, when $r(h|s) = q(h|s)$, for every *s.* All of the information in the score can be used by the user only if this penalty is zero. Note further, if the detection system makes the original score available to the evaluator and if it is within the means of the evaluator to calculate the ideal posterior *r(h|s)*, then the evaluator can force this term to zero. This is analogous to what is done in the NIST evaluations with "minimum C_{det}", where an optimal threshold is set by the evaluator to minimize detection cost over the evaluation data. In both cases this is a measure of the utility of the extraction stage alone, where the presentation has been optimized.

7.5. Note on prior dependence

The transformation between detector types *I* and *II* opens further possibilities for evaluation: In the current NIST evaluations, since detector thresholds are pre-set for a specific prior, the evaluation can only be done for this prior. But if

[1] Do not confuse the symbols $H\{.\}$ for entropy and H_i for hypothesis.

systems present a likelihood-ratio, the prior may be changed at will after system results have been submitted. Plots of utility estimates against prior (see eq.12) may be presented, instead of single values at a chosen prior. (This applies for any utility.)

All the terms of the expected logarithmic utility (eq.26) are dependent on the prior. (This is the case also for the decisional scoring.) This dependence on the prior is unavoidable, because in the limit as the prior uncertainty becomes zero, the detector would be *unnecessary*. (Mathematically the utility of the posterior would be maximized independently of the likelihood-ratio supplied by the detector.)

One solution to obtaining a prior-independent evaluation could be as follows: By using $\gamma(.) \equiv \gamma_0(.)$, the expected logarithmic utility (as a function of the prior) would approach a maximum of *one* at either extreme as the prior entropy approaches *zero*; and it would have a minimum at some intermediate prior. This minimum expected utility is a prior-independent measure. (Different systems would have minima at different priors.)

7.6. Conclusion

By combining (via eq.12) conditional averages of utility (using equation 16 in 25), over a supervised database, the quality of likelihood-ratio presentation may be evaluated, relative to the state of knowledge defined by the evaluation data. As in the case of evaluation by detection cost, *extraction* and *presentation* are evaluated simultaneously.

8. Summary

We have shown that speaker detection systems should, in order to be maximally useful for different applications, output likelihood-ratios rather than decisions. This is in agreement with what has been proposed for forensic speaker recognition. One of the obstacles to development of such systems has been the lack of the means to evaluate the quality of this form of output. We have shown both how to transform existing systems to output likelihood-ratios and how to transform the existing NIST evaluation to evaluate such outputs. In particular, we have proposed the use of a logarithmic utility, which has an attractive information-theoretic interpretation.

It is hoped that the proposed evaluation mechanism will stimulate more research into the difficult problem of explicitly calculating likelihood-ratios. By pointing out the transformation between detector types *I* and *II*, we have shown that this problem has always implicitly existed: The problem of optimizing the decision stage is equivalent to calculation of the likelihood-ratio.

9. Acknowledgement

The author whishes to thank Doug Reynolds and Joe Campbell for some stimulating input.

10. References

[1] J.M. Bernardo and A.F.M. Smith, *Bayesian Theory,* John Wiley & Sons, 1994.

[2] E.T. Jaynes, *Probability Theory: The Logic of Science,* Cambridge University Press, 2003.

[3] J. Gonzalez-Rodriguez et al. "Robust Likelihood Ratio Estimation in Bayesian Forensic Speaker Recognition", *Proc. Eurospeech 2003*, Geneva, 2003.

[4] A. Drygajlo, D. Meuwly, A. Alexander "Statistical Methods and Bayesian Interpretation of Evidence in Forensic Automatic Speaker Recognition" *Proc. Eurospeech 2003*, Geneva, 2003.

[5] C. Fredouille, J.F. Bonastre, T. Merlin, "Bayesian Approach based-Decision in Speaker Verification", *Proc. 2001: A Speaker Odyssey The Speaker Recognition Workshop*, Crete, 2001, pp. 77-81.

[6] R.T. Cox, "Probability, Frequency and Reasonable Expectation", *Am. J. Phys.,* 14: pp 1-13, 1946.

[7] A. Wald, *Statistical Decision Functions*, Wiley, New York, 1950.

[8] See the NIST Speaker Recognition Evaluations at http://www.nist.gov/speech/tests/spk/index.htm

[9] A. Martin and M. Przybocki, "The NIST 1999 Speaker Recognition Evaluation – An Overview", *Digital Signal Processing*, vol. 10, nos 1-3: pp.1-18, 2000.

[10] A. Martin et al., "The DET curve assessment of detection task performance", *Proc. EuroSpeech*, vol.4: pp.1895-1898, 1997.

[11] R. L. Winkler and A. H. Murphy "Good probability assessors". *J. Applied Meteorology*, 7: 751-758, 1968.

[12] M.S. Roulston and L.A. Smith, "Evaluating Probabilistic Forecasts Using Information Theory", *Monthly Weather Review* 130: 1653-1660, 2002.

[13] N.C. Dalkey, "Inductive Inference and the Maximum Entropy Principle", *Maximum-Entropy and Bayesian Methods in Inverse Problems,* eds.: C.R. Smith and W.T. Grandy, D. Reidel Publishing Company, Dordrecht, 1985, pp.351-364.

[14] P. Sebastiani and H.P. Wynn, "Experimental Design to Maximize Information". *MaxEnt 2000: Twentieth International Workshop on Bayesian Inference and Maximum Entropy in Science and Engineering. AIP Conference Proceedings, 2000,* pp. 192-203.

[15] C.E. Shannon, "A Mathematical Theory of Communication", *Bell Syst. Tech. J.*, vol.27, pp.379-423, 623-625, July and Oct. 1948.

[16] R. Auckenthaler et al. "Score Normalization for Text-Independent Speaker Verification Systems" *Digital Signal Processing,* vol. 10, nos 1-3, 2000, pp.42-54.

[17] G. Casella, R.L. Berger, *Statistical Inference* 2nd Edition, Duxbury 2002, Chapter 6.

[18] H. Jiang and L. Deng, "A Bayesian approach to the verification problem: -- applications to speaker verification", *IEEE Trans. on Speech and Audio Processing*, Vol. 9, No.8: pp.874-884, November 2001.

[19] R.E. Blahut, *Principles and Practice of Information Theory,* Addison-Wesley, 1987.

[20] B. Pfister and R. Beutler, "Estimating the Weight of Evidence in Forensic Speaker Verification", *Proc. Eurospeech 2003*, Geneva, 2003.

[21] Juang, B.-H.; Katagiri, S , "Discriminative learning for minimum error classification", IEEE Trans on Signal Processing , Volume: 40 , Issue: 12 , Dec. 1992, pp.3043 – 3054.

[22] Loredo, T. J., "From Laplace To SN 1987A: Bayesian Inference In Astrophysics", in *Maximum Entropy and Bayesian Methods*, P. F. Fougere (ed), Kluwer Academic Publishers Dordrecht, The Netherlands, 1990, pp. 81-142.