A speaker line-up for the Likelihood Ratio

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

[1]Radboud University Nijmegen, The Netherlands
[2]Agnitio Research, South Africa

Abstract

We propose an analogy to eye witness line-up in order to compute calibrated likelihood ratios for speaker recognition, by including the target model in an identification trial with a cohort of foils. Expressions for the likelihood ratio as a function of cohort size, identification rank and system ROC performance are derived, and some properties of the likelihood ratio function are discussed. The line-up procedure is used as a method to calibrate recognition scores. Using NIST SRE 2010, we find calibration loss comparable to linear calibration (FoCal), while the proposed method gives improved discrimination.

Index Terms: Likelihood Ratio, calibration, speaker recognition, forensics.

1. Introduction

In the Bayesian Paradigm the *likelihood ratio* (LR) is the proper way to present the strength of forensic evidence to court [1, 2]. The likelihood ratio is the probability that the evidence is observed given the prosecutor's hypothesis divided by the probability that the evidence is observed given the defense hypothesis. A traditional method of producing a likelihood ratio with eye witnesses is the line-up. The eye witness is requested to identify the perpetrator of a crime in a line-up where the suspect is amidst a number of other subjects (foils) that are not, as a group, markedly different from the suspect in any particular way.[1] If the eye witness identifies the suspect as the perpetrator, this evidence has a weight that can be expressed as having a likelihood ratio close to N. This is quite an intuitive interpretation of the likelihood ratio.

In this paper we want to extend the notion of an eye witness in a visual line-up setting to a biometric system with a reference database, and we will later apply this to automatic speaker recognition. One of the criticisms made to using biometric technology in a court setting is that the systems are known to make errors. We will use this, in fact, to our advantage and include the probability of making errors in the computation of the likelihood ratio. One of the advantages of the study of biometric technologies is that the error rates are one of the best studied properties of a biometric system. For instance, in text independent automatic speaker recognition the proper determination of the performance of a system has almost reached the level of an art, with regular system evaluations organized by an independent authority (the National Institute of Standards and Technology, NIST), and well studied evaluation methodology and evaluation measures. Note that thus far, calibration in automatic speaker recognition has always been carried out in a data-driven, evaluative way, which is in sharp contrast to the calculation of likelihood ratios in forensic DNA, where the LR calculation is motivated purely by constructing transparent and well-motivated probabilistic models [1].

The performance of a speaker recognition system is, for the purposes in this paper, completely determined by its detection error trade-off (DET) performance, more generally known as the Receiver Operating Characteristic (ROC). From this, the identification error rates for a line-up with N speakers can be computed, and hence the likelihood ratios for the cases when the suspect is the author of the trace, and when it is not.

This paper is organized as follows. In Section 2 the expressions for the likelihood ratio are derived, and in Section 3 the behaviour of the LR is studied varying performance and size parameters. Then, in Section 4 experiments of using this method for calibrating an automatic speaker recognition system are present, before the discussion in Section 5.

2. Theory

In this section we will derive expressions for the likelihood ratio for a system that makes recognition errors. The expressions equally apply to other line-up settings, such as the eye witness, but in these cases it may be difficult to determine the error rates involved.

2.1. The likelihood ratio for positive and negative identification

Consider a line-up setting with the suspect (target) speaker and a cohort $\mathcal{F}$ of $N-1$ foils. The numerator of the likelihood ratio $P(E|H_p)$, where E is the evidence and H_p the prosecutor's hypothesis that the suspect is the author of the trace, can be expressed as the probability of finding

[1]The subject of choosing foils is a study by itself [3] and is not the study of this paper.

the suspect in a closed set identification experiment

$$P(r=1|H_p,\mathcal{F}) = 1-e_N, \quad P(r>1|H_p,\mathcal{F}) = e_N \tag{1}$$

where r is the rank of the suspect model score among the $N = \|\mathcal{F}\| + 1$ speaker models in the line-up, and e_N is the system's identification error rate for a closed set of N speakers. Similarly, the denominator of the likelihood ratio in a line-up is

$$P(r=1|H_d,\mathcal{F}) = \frac{1}{N}, \quad P(r>1|H_d,\mathcal{F}) = \frac{N-1}{N}. \tag{2}$$

It is argued that if the defense hypothesis H_d is true, the suspect has just random chance to have top ranking amidst $N-1$ other, equally unrelated, speakers. This expression assumes that the perpetrator is not (accidentally) among the foils, and that the recording of the foils are not in any particular way different from that of the suspect. Note that $P(r|H_d)$ is independent of the error rate e_N as any positive identification is erroneous under H_d.

With the expressions (1) and (2) the likelihood ratio ℓ can be computed

$$\ell(r=1) = N(1-e_N), \quad \ell(r>1) = \frac{Ne_N}{N-1} \tag{3}$$

It may be interesting to observe this expression in some detail. The logarithm of the likelihood ratio is bounded above mostly by the size of the line-up, N, in case of "positive identification," while in the case of a "negative identification" it is bounded only by the error rate of identification.

2.2. Identification Error Rate

We are still left with determining e_N, the expected error rate for an identification set up with N competing speaker models. Consider the detection error trade-off performance of a biometric detection system. In speaker recognition this is graphically shown in a DET plot [4], more traditionally this is carried out in a Receiving Operating Characteristic (ROC). In a detection system the task is to determine whether or not trace and reference are from the same source. A biometric systems gives a score s for a trial consisting of trace and reference that is the result of a comparison of trace and reference sample. We will take the 'direction sense' of comparison score to grow when the likelihood is higher for same-source comparisons than for different source comparisons (i.e., the opposite sense of a distance). Let $p_t(s)$ and $p_n(s)$ be the continuous probability densities of the same-source (or *target*) comparisons and different source (or *non-target*) comparisons, respectively. The ROC is the parametric relation between false negatives (*misses*) and false positives (*false alarms*) sweeping a threshold t for the scores.[2]

[2]For comparison with the normal DET analysis we choose our y-axis as an error rate. Traditionally, a ROC curve is drawn with the hit rate, or true positive rate, as y-axis. This is just the complement of the miss rate.

Writing

$$M(t) = \int_{-\infty}^{t} p_t(s)ds, \quad F(t) = \int_{t}^{\infty} p_n(s)ds \tag{4}$$

the ROC curve can be described as $M(F^{-1}(p))$, or its inverse $F(M^{-1}(p))$, where $F^{-1}(p) = t$ is the inverse[3] of function $F(t) = p$.

Now consider a line-up of size N, under the prosecution hypothesis, when there is one target score and $N-1$ non-target scores. Then correct identification results when all non-target scores are smaller than the target score. Under the assumption of independence of the non-target scores, we can express the probability for correct identification as:

$$\begin{aligned} 1-e_N &= \int_{-\infty}^{\infty} p_t(s)\left(1-\int_{s}^{\infty} p_n(s')ds'\right)^{N-1} ds \\ &= \int_{-\infty}^{\infty} p_t(s)\big(1-F(s)\big)^{N-1} ds \qquad (5) \\ &= \int_{0}^{1} \big(1-R(p)\big)^{N-1} dp, \qquad (6) \end{aligned}$$

where we have made the transformation $p = M(s)$, giving $dp = p_t(s)ds$ and $s = M^{-1}(p)$, and used $R = F \circ M^{-1}$, the inverse ROC function.

A special case of the complement of (6) for $N = 2$ is the area under the ROC curve[4] $p_{\mathrm{AUC}} = \int_0^1 R(p)dp$.

2.3. Likelihood for rank

Having carried out the exercise of determining the identification error rate e_N given the ROC curve of a recognition system, the likelihood ratio can be computed using (3). Here, the possible outcome of an identification has been quantized in two bins $r = 1$ and $r > 1$. However, we can determine the likelihood of finding the suspect in any rank r in the sorted list of scores found in the identification process. Realizing that $N-r$ non-target scores should be lower than the target score, and $r-1$ higher, this gives[5]

$$P(r|H_p) = \binom{N-1}{r-1} \int_0^1 \big(1-R(p)\big)^{N-r} \big(R(p)\big)^{r-1} dp, \tag{7}$$

the combinatorial factor stemming from the fact that there are multiple combinations of non-target scores that give rank r. Using the same arguments given after (2) the likelihood for the defense hypothesis for any rank is constant $P(r|H_d) = \frac{1}{N}$ so that the likelihood ratio becomes

$$\ell(r) = NP(r|H_p). \tag{8}$$

[3]Since M and F are definite integrals of continuous positive functions, their derivatives exist and they are continuous, strictly monotonic and invertible.

[4]Again, the complement of what is traditionally known as 'the area under the curve' AUC.

[5]Note that this integral also plays a role in the Cumulative Match Curve (CMC) in Biometric Identification [5].

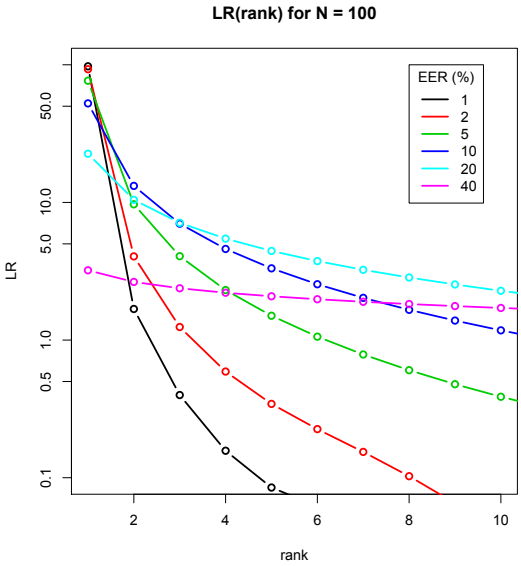

Figure 1: The log likelihood ratio as a function of identification rank r, for a line-up of 100 speakers, for various system performances. The integration of (7) has been carried out numerically.

3. Properties of the likelihood ratio

Although the ROC function $R(p)$ must be determined empirically, we can model this function in order to study the behaviour for different r, N and discrimination powers $R(p)$. In order to do so we model the distributions $p_t(s)$ and $p_n(s)$ as Gaussians with equal variance σ^2 and difference in mean $d'\sigma$, analogous to psycho-physical detection theory. In Fig. 1 we show the likelihood ratio for the first few ranks in a 100-speaker line-up of a '45° DET' recognition system of different performance, specified by the Equal Error Rate $e_=$, which defines d' via

$$d' = -2\Phi^{-1}(e_=) = -2\sqrt{2}\,\mathrm{erf}^{-1}(2e_= - 1) \qquad (9)$$

where Φ^{-1} is the inverse of the cumulative normal distribution, the warping function of the axes of the much beloved DET plot. With these score distributions we cannot solve the integral (7), but we may observe the following characteristics:

- $\ell(r, N)$ is a decreasing function of r for $e_= < 50\,\%$,
- $\ell(r, N) = 1$ for a system with $e_= = 50\,\%$, or $p_{AUC} = 0.5$,
- $\ell(r, N) - \ell(r+1, N)$ is larger for lower $e_=$ (the dependence on r is stronger),
- $\ell(r, N_1) > \ell(r, N_2)$ for $N_1 > N_2$ and $r \leq N_2$.

4. Experiments using speaker line-up

In this section we will present some experiments with the speaker line-up approach to compute calibrated likelihood ratios. We take one of the RU sub-systems used in SRE-2010 as the speaker recognition system performing speaker comparisons.

4.1. Speaker Recognition System

The speaker recognition system is a modified version of a so-called 'dotscoring' system, submitted to the Evalita 2009 evaluation [6] which is based on SDV's submission to SRE 2008. The system uses feature-warped MFCC features, UBMs of size 512, channel compensation of rank 50, and linear scoring, making the whole system very fast. Scores are submitted to zt-normalization.

4.2. Calibration and evaluation data

We use the trials of the NIST Speaker Recognition Evaluation (SRE) 2008 that consisted of 1 conversation side for training and test, English telephone trials, for calibration. We test using the normal vocal effort telephone trials from NIST SRE 2010 condition, a.k.a. 'condition 5,' which are also all in English. We use the 'extended trial list' distributed by NIST during the evaluation, consisting of 7169 target and 408950 non-target trials.

4.3. Evaluation measures

Since we are interested in the calibration of the likelihood ratio we use C_{llr} as evaluation measure [7]. As a measure for calibration we can specifically look at $C_{\mathrm{llr}} - C_{\mathrm{llr}}^{\mathrm{min}}$, the extra costs incurred by suboptimal calibration. C_{llr} evaluates the quality of the likelihood ratio over all likelihood ratios, not just the likelihood corresponding to the effective prior given by the cost function parameters $\mathcal{O}_{\mathrm{eff}} = (P_{\mathrm{tar}}/(1-P_{\mathrm{tar}}))(C_{\mathrm{miss}}/C_{\mathrm{FA}})$ as is the case in NIST evaluations. We will further present the Equal Error Rate $e_=$ as a measure for overall discrimination ability.

4.4. Experiments

We will calibrate the evaluation scores in two ways: using linear calibration with FoCal [8] and using the method proposed above. In linear calibration we use all the supervised trials of the calibration set in order to train an affine transform of the recognizer scores to likelihood ratios, which is a logistic regression optimization. The transformation is applied to the evaluation trials, and the corresponding likelihood ratios are evaluated using C_{llr}.

In the line-up calibration method we determine for each test segment the rank of the target model amidst all speaker models from the calibration data. The ranks are then converted to likelihood ratios using (8), where the integrals are computed by summation over the empirical ROC curve of the calibration trials. All experiments are carried out conditioned on gender.

system	C_{llr}	$C_{\text{llr}}^{\text{min}}$	$e_=$	$\min \log \ell$
male linear	0.173	0.158	4.13	$-$ 15.5
male line-up	0.290	0.129	3.09	-695
male bounded	0.160	0.130	3.09	$-$ 5.91
fem. linear	0.246	0.224	5.87	$-$ 16.8
fem. line-up	0.211	0.189	4.81	-107
fem. bounded	0.209	0.190	4.81	$-$ 6.38

Table 1: Results of the calibration experiment

5. Results and Discussion

In Table 1 we have tabulated the results of the experiments, separated for male and female, linear calibration and line-up calibration.

We see that the line-up calibration works, because the values for C_{llr} are well below unity. Still, for the male line-up, the difference between C_{llr} and $C_{\text{llr}}^{\text{min}}$ is quite remarkable. This is due to a few trials with very negative $\log \ell$. As was remarked earlier, there is not really a lower bound to $\log \ell$, while there is a clear upper bound determined by the size of the line-up. We have added a row 'male bounded' where we have set a lower bound to $\log \ell$, equal to the negative of the upper bound given by (3). This makes the $\log \ell$ more symmetric, which seems a reasonable thing to do. The result is that the small calibration problem is fixed, and the performance is actually better than the state-of-the-art linear calibration.

The female trials behave more normally, here the very negative $\log \ell$ does not harm, apparently because there are no target trials with low likelihood ratio. Still, putting the same lower bound on the calibration does not harm the performance either.

A remarkable side-effect of the line-up calibration is that the discrimination improves as shown by $e_=$ and $C_{\text{llr}}^{\text{min}}$, and the DET curves in Fig. 2. Although welcome, this is surprising, because the scores had already been zt-normed. The line-up calibration does not only have parallels with t-norm, but also to short time Gaussianization at the feature level, where the rank in a cohort of values is mapped to a theoretical distribution.

The procedure put forward in this paper can be used in forensic evidence reporting: the forensic specialist can supply the cohort of foils according to proper criteria. The method naturally has an upper bound to the likelihood ratio which is more satisfactory than using parametric calibration, which is known to sometimes produce unrealistically large LR values.

However, for such forensic use, further study is required, which should include: (i) examining the score independence assumption. (ii) discussing cohort selection. (iii) quantifying uncertainty in the recognizer error-rates and how to integrate out this uncertainty to produce a more honest LR that takes this uncertainty into account.

If this method is to be used for calibration of the new NIST operating point at $\mathcal{O}_{\text{eff}} = 10^{-3}$, the cohort should

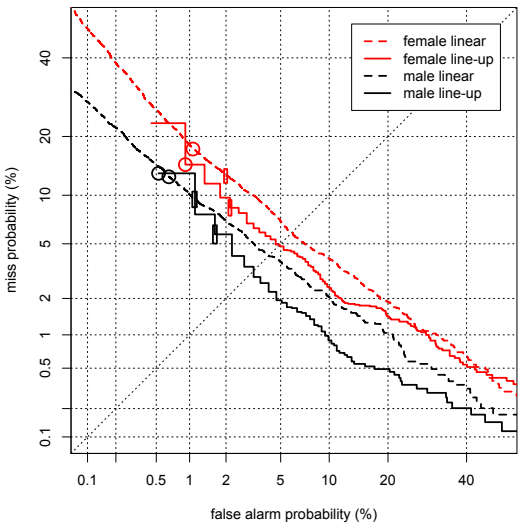

Figure 2: DET plot of the calibration experiments. Note how the line-up improves the discrimination.

contain thousands of speakers, in order to reach values of $\ell = 1000$, with very low error rates. This shows once more, that the new NIST operating point is a real challenge to calibration.

In future research we intend to investigate the possibility to subject 'the other side of the trials,' the trace, to a similar line-up normalization procedure, comparable to z-norm in score normalization.

6. References

[1] D. J. Balding, *Weight-of-evidence for Forensic DNA Profiles*. John Wiley & Sons, 2005.

[2] J. González-Rodríguez, A. Drygajlo, D. Ramos-Castro, M. Garcia-Gomar, and J. Ortega-Garcia, "Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition," *Computer Speech and Language*, vol. 20, no. 2–3, pp. 331–355, 2006.

[3] G. L. Wells, M. Small, S. Penrod, M. Roy S, S. M. Fulero, and C. A. E. Brimacombe, "Eyewitness identification procedures: Recommendations for lineups and photospreads," *Law and Human Behavior*, vol. 22, no. 6, pp. 1–39, 1998.

[4] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech 1997*, Rhodes, Greece, 1997, pp. 1895–1898.

[5] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior, *Guide to Biometrics*, ser. Springer Professional Computing. Springer, 2004.

[6] M. Huijbregts and D. A. van Leeuwen, "The RU submission to the EVALITA "application track" speaker recognition evaluation," in *Evalita Workshop*, 2009.

[7] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, pp. 230–275, 2006.

[8] N. Brümmer, *FoCal-II: Toolkit for calibration of multiclass recognition scores*, August 2006, software available at http://www.dsp.sun.ac.za/~nbrummer/focal/index.htm.