

# Proceedings of Biometric Technologies in Forensic Science

Editor: David van Leeuwen

Nijmegen, 14–15 October 2013



## Introduction

In criminal investigations, legal authorities can find traces that can help to identify people involved in criminal activity. Examples of traces are video footage from a security camera, a recording of a threatening telephone call, or fingermarks left on the crime scene. In the investigative part of the judicial process, the traces can be used to find suspects of the crime, while later in court, the trace can be put forward as evidence. For both these forensic activities, biometric technologies can be used to automate and objectivize the process.

Contrary to typical access control and authorization applications of biometric technologies, the trace in the forensic situation is characterized by an uncontrolled and low quality. This typically leads to a lower recognition performance of the biometric system and hence a lower evidential value. Ample well-controlled data is often available for the development and performance validation of the basic biometric classifier. However, the case-specific and privacy-sensitive nature of criminal investigation puts limitations on the forensic validation of these biometric technologies. Many applications in biometric technology are designed to perform in the low-false-accept range of the recognition system, but for presenting the weight of evidence of a biometric comparison in court, a calibrated likelihood ratio is required which is accurate over a wide range of priors.

This conference aimed at presenting and discussing scientific research undertaken to address any of the challenges set by forensic science for aspects of biometric technology such as sample quality, available data or calibration. We have therefore enjoyed the participation of researchers from both the forensic science and the biometric technology communities that showed interest to try and understand each other's motives and challenges, and willingness to bridge the potential gap between these communities due to different standpoints, terminology and experience.

David van Leeuwen, Didier Meuwly and Raymond Veldhuis, programme chairs for BTFS 2013.



## Contents

1	Calibration based on duration quality measures function in noise robust speaker recognition for NIST SRE'12 . . . . .	1
2	Estimation of cylinder quality measures from quality maps for Minutia-Cylinder Code based latent fingerprint matching . . . . .	6
3	Can Facial Uniqueness be Inferred from Impostor Scores? . . . . .	11
4	On the Improvements of Uni-modal and Bi-modal Fusions of Speaker and Face Recognition for Mobile Biometrics . . . . .	15
5	Bridging the gap between the forensic handwriting and pattern recognition communities . . . . .	23
6	The distribution of calibrated likelihood-ratios in speaker recognition . . . . .	24
7	Estimated Intra-Speaker Variability Boundaries in Forensic Speaker Recognition Casework . . . . .	30
8	Influence of the datasets size on the stability of the LR in the lower region of the within source distribution . . . . .	34
9	A human benchmark for automatic speaker recognition . . . . .	39
10	Assessing latent fingerprint distortion using forensic databases and minutiae paring by human experts . . . . .	46
11	Normalized Ordinal Distance; A Performance Metric for Ordinal, Probabilistic-ordinal or Partial-ordinal Classification Problems . . . . .	47
12	Towards Quantification of the Weight of Evidence with Partial Fingermarks on Real Forensic Casework . . . . .	51
13	Speaker recognition by means of Deep Belief Networks . . . . .	52
14	Frequency and ridge estimation by structure tensor . . . . .	58



## CALIBRATION BASED ON DURATION QUALITY MEASURES FUNCTION IN NOISE ROBUST SPEAKER RECOGNITION FOR NIST SRE'12

*Miranti Indar Mandasari, Rahim Saeidi, David A. van Leeuwen*

Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands

### ABSTRACT

This paper presents the performance evaluation of i-vector and PLDA based speaker recognition system which incorporate *quality measures function* (QMF) in linear calibration. Evaluated on the recent NIST SRE'12 corpus, the linear calibration with QMF as the additional term shows a positive gain in the system performance compared to the conventional linear calibration with only two terms. Based on the equal error rate values measured from our I4U evaluation trial set, the QMF calibration approach shows 10–37% relative improvement compared to the conventional linear calibration. It is shown that by adding 1–2 extra parameters in the linear calibration through QMF approach, there is a potential to improve the calibration and discrimination performances of a speaker recognition system.

**Index Terms**— Calibration, duration, quality measures, QMF, speaker recognition.

### 1. INTRODUCTION

In a specific field such as forensic speaker identification [1], calibration is very important in order to make the scores produced by an automatic speaker recognition system more reliable. By presenting calibrated scores in the likelihood ratio form, results can be used as legal evidence in court [2]. In the more general field of speaker recognition, the significance of calibration is becoming more recognized by the community. Especially in the 2012 edition of the speaker recognition evaluation (SRE) from the National Institute of Standard and Technology (NIST), calibration is an interesting topic to be discussed amongst the researchers in the field [3]. This is because of the requirement by NIST to participants to express their system output in the log-likelihood-ratio form.

The NIST SRE'12 provides plenty of challenges to its participants. In terms of quality measures of speech samples, there are two problems addressed in this year evaluation, duration variation (20–160 s) and noisy speech conditions. This year's evaluation also includes new performance measure called *primary cost* that is defined as the average of Bayes error rates from two detection cost functions. To address some of the interesting problems offered by the SRE'12 evaluation, this paper presents the performance analysis from calibration that is based on the duration quality measures function (QMF) approach.

The duration QMF for calibration in speaker recognition system is proposed in [4]. In this approach, we add the QMF as an extra term in the linear calibration. Evaluated on the NIST SRE'08 and SRE'10 corpora with truncations to shorter duration, it has been observed that QMF calibration is robust in the conditions where speech

duration varied. In this paper, motivated by the challenging NIST SRE'12 protocol, we use the QMF approach as a simple yet robust technique to deal with the duration variability effect on the discrimination and calibration performance of speaker recognition system.

This paper presents the speaker recognition system and corpora descriptions in Section 2 and Section 3, respectively. Calibration methods are detailed in Section 4 and performance measures are listed in Section 5. Experiment results are discussed in Section 6. This paper is concluded in Section 7.

### 2. SPEAKER RECOGNITION SYSTEM

System configuration of speaker recognition in this paper is fairly similar with the configuration in the latest papers of the authors [4,5]. The system is based on i-vector [6] framework and probabilistic linear discriminant analysis (PLDA) modeling [7,8]. The main difference is that there is an inclusion of speech enhancement algorithm with a *dynamic noise suppression* rule [9] in the system used for this paper. For this noise suppression purpose, we did noise estimation through *improved minima controlled recursive averaging* (IM-CRA) [10], and Wiener filter is applied on the amplitude spectrum as a soft mask.

Spectral features used in this system is 60 dimensional MFCCs which consist of 19 base MFCCs and log energy, augmented with deltas and double deltas. The features are extracted every 10 ms using 30 ms window. To enhance the features quality, a feature-warping was applied [11]. A speech activity detection (SAD) is implemented to extract the active speech frames from the features [12]. Gender-dependent with 2048 components UBM (universal background model) was trained using NIST SRE 2004–2006, Switchboard Cellular phase 1 and 2, and Fisher English corpora.

The i-vectors were trained using a low dimensional (400 dimensions) matrix that defines both the speaker and channel subspaces. Linear discriminant analysis (LDA) projection was applied in order to reduce the i-vectors dimension to 200. Prior to PLDA modeling, the i-vectors were processed by i-vector centering, within-class covariance normalization (WCCN) and length-normalization.

### 3. CORPORA

The speaker recognition system calibration performance is evaluated using NIST SRE'12 corpus. There are three different datasets used in the experiments. The first two is the *Dev-I4U* (development) and *Eval-I4U* (evaluation) sets from I4U<sup>1</sup> trials list [13]. The third dataset used in the experiments is the evaluation set from NIST SRE'12 trials list which refers to *Eval-SRE'12* set in this paper. For this NIST SRE 2012 evaluation, we made three sub-selections of the core-test core-training condition for different noise levels, based on the 5 common conditions (cc's) defined in the evaluation plan, using

The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement number 238803.

<sup>1</sup>I4U is a joint effort from 9 research Institutes and Universities across 4 continents in joining the NIST SRE'12 evaluation. The lists are available via <http://lands.let.ru.nl/~saeidi/I4U.tgz>



**Table 1.** Number of trials in the Dev-I4U, Eval-I4U and Eval-SRE’12 sets for female gender and “unknown” non-target trials.

Set	Noise condition	Number of Trials	
		Target	Non-target
Dev-I4U	Clean	6621	2118521
	15 dB noisy	6621	2118521
	6 dB noisy	6621	2118521
Eval-I4U	Clean	6921	2997225
	15 dB noisy	6921	2997225
	6 dB noisy	6921	2997225
Eval-SRE’12	Clean	4353	120223
	15 dB noisy	2913	7908
	6 dB noisy	2912	7908

version 1 of the trial key. The first level, “noise”, consisted of all telephone and microphone speech, without noise addition, that were not recorded in a noisy environment (the intersection of cc1 and cc2 with trials from cc5 removed). For the two noisy conditions, “15 dB” and “6 dB,” we selected trials from cc3 and cc4 with added noise of types “babble” and “HVAC” at 15 dB and 6 dB, respectively.

We divided each dataset into 3 different subsets based on the noise conditions in the *test segments* of the trials listed as *clean* (no-alteration), *15 dB* and *6 dB* noise-levels subsets. The number of trials for target and non-target scores are presented in Table 1. Partitioning the results with respect to SNR is intended for analysis of the calibration sensitivity with duration function to SNR-levels in test segments. In the training of calibration parameters, the scores were pooled without noting the SNR-levels. Only the *unknown* non-target<sup>2</sup> trials are included in the experiments, and we focus our experiments on female speakers. By looking at the durations of utterances in the NIST SRE’12 database (a histogram is provided in Figure 2), we see there is a high variability in duration, therefore performing consistent calibration is a very challenging task.

#### 4. CALIBRATION

All calibrations performed for the experiments in this paper are based on the linear transformation of scores into calibrated log likelihood ratio scores. There are two calibration approaches used. The first approach is conventional linear calibration with two parameters, and the second approach is linear calibration that applies *quality measures function* (QMF) as an extra linear term in calibration.

##### 4.1. Linear calibration

In the linear calibration, we transform a set of *raw scores*  $s$  which produced from the speaker recognition system to a set of *calibrated scores*  $\ell$  using a linear transformation

$$\ell = w_0 + w_1 s, \quad (1)$$

where  $w_0$  is the offset/gain parameter and  $w_1$  is the scaling parameter of calibration. In this paper, this two parameterized linear calibration is referred to as *conventional* calibration approach.

In the experiment, calibration parameters are trained in a set of scores, and then applied to another set of scores to be evaluated. In this paper, the calibration parameters are trained in the Dev-I4U set (including all noise condition subsets) and applied to all sets which are Dev-I4U, Eval-I4u and Eval-SRE’12. The parameters for both conventional and QMF calibration approaches were trained via *logistic regression* [14] using FoCal toolkit<sup>3</sup>.

<sup>2</sup>This is done for compatibility results with earlier SRE protocols. [3]

<sup>3</sup>Software is available at <https://sites.google.com/site/nikobrummer/focal>

**Table 2.** Duration quality measure functions (QMFs) proposed for calibration on various duration conditions.

$n$	QMF: $Q_n(d_m, d_t, \dots)$	Additional parameters
1	$Q_1 = w_2 \left  \log \frac{d_m}{d_t} \right $	$w_2$
2	$Q_2 = w_2 \log^2 \frac{d_m}{d_t}$	$w_2$
3	$Q_3 = w_2 \log \frac{d_m}{d_c} \log \frac{d_t}{d_c}$	$w_2, d_c$
4	$Q_4 = w_2 \log \frac{d_m}{d_c} \log \frac{d_t}{d_c} + w_3 \left( \log^2 \frac{d_m}{d_c} + \log^2 \frac{d_t}{d_c} \right)$	$w_2, w_3, d_c$

##### 4.2. Quality Measure Function

Quality measures function or QMF calibration is proposed by the authors in [4] and it was analyzed for the SRE’08 and SRE’10 corpora. This calibration approach basically is a linear calibration technique with several extra parameters in the linear transformation. It also includes the quality measures of speech utterance in the calibration, in this case, the duration of active speech. The QMF calibration approach is applied via linear scores transformation that can be formulated as:

$$\ell = w_0 + w_1 s + Q(d_m, d_t, w_2, \dots) \quad (2)$$

with  $Q(d_m, d_t, w_2, \dots)$  as the function that defines quality measures we use for calibration, and  $d_m$  and  $d_t$  as the duration of active speech (after SAD) in the model and test segments, respectively. There were multi-sessions enrollment in the NIST SRE’12, thus we use sum of the duration of utterances in model segments as  $d_m$ .

There are four QMFs proposed in [4] and all of this QMFs are analyzed in this paper as well. The four QMFs are presented in Table 2. All QMFs are modeled from the behavior of the calibration parameters of linear score transformation (scaling parameters) in various duration conditions of the model and test segments. Figure 1 depicts the behavior of the scaling parameters across duration of model and test segments. The first two QMFs,  $Q_1$  and  $Q_2$  are formed in order to model the large deviation of the scaling parameters when there is a large difference (mismatched) between the model and test segments. The last two QMFs,  $Q_3$  and  $Q_4$  are modeled from the saddle-plane like of the scaling calibration parameters in calibration which is presented in Figure 1 with  $d_c = 20$  s.

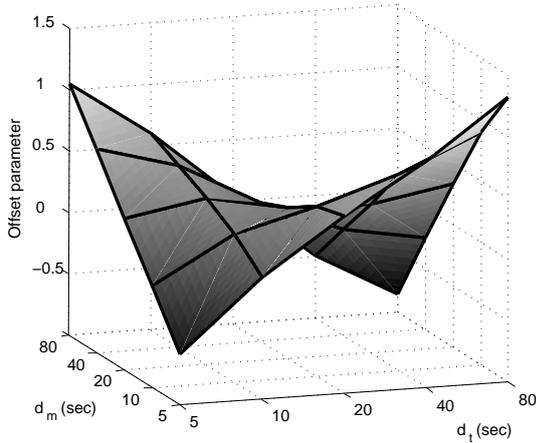
#### 5. PERFORMANCE MEASURES

There are five performance measures used to characterized the speaker recognition system performance of discrimination and classification, namely equal error rate ( $E_+$ ), primary cost from NIST SRE’12 ( $C_{\text{primary}}$ ), cost of log likelihood ratio calibration ( $C_{\text{llr}}$ ), minimum  $C_{\text{llr}}$  ( $C_{\text{llr}}^{\text{min}}$ ), and the miscalibration cost ( $C_{\text{mc}}$ ).

##### 5.1. Equal Error Rate

Equal error rate or  $E_+$  is the error rate of a binary-classifier when the probability of the false-acceptance rate and false-rejection rate is equal at a certain point in the detection error trade-off (DET) curve. The  $E_+$  was computed using `sretools` analysis package<sup>4</sup> in R using relative operating point convex hull (ROC-CH) approach.

<sup>4</sup>Software is available at <https://sites.google.com/site/sretools/>



**Fig. 1.** The saddle-plane shape of calibration (*offset*) parameters in various duration conditions on the model and test segments [4].

## 5.2. Primary cost for the NIST SRE'12

For this year's speaker recognition evaluation, NIST announced new detection cost function or  $C_{\text{det}}$  that is referred to  $C_{\text{primary}}$ . Unlike the previous evaluations, the SRE'12 cost function is a combination of two costs, the cost of NIST SRE'10 ( $P_{\text{tar}} = 1/1000$ ) and another cost with greater prior than SRE'10 ( $P_{\text{tar}} = 1/100$ ). Each of these  $C_{\text{det}}$  is computed using

$$C_{\text{det}} = C_{\text{miss}} \times P_{\text{tar}} \times P_{\text{miss|tar}} + C_{\text{FA}} \times (1 - P_{\text{tar}}) \times (P_{\text{FA|non,known}} \cdot P_{\text{known}} + P_{\text{FA|non,unknown}} \cdot (1 - P_{\text{known}})) \quad (3)$$

with  $C_{\text{miss}} = C_{\text{FA}} = 1$ . Because in the experiments, we used only unknown non-target trials<sup>5</sup>, Equation (3) becomes

$$C_{\text{det}} = C_{\text{miss}} \times P_{\text{tar}} \times P_{\text{miss|tar}} + C_{\text{FA}} \times (1 - P_{\text{tar}}) \times P_{\text{FA|non}}. \quad (4)$$

The  $C_{\text{det}}$  values are computed using BOSARIS<sup>6</sup> toolkit via Bayes error rate computation.

## 5.3. Cost of Log Likelihood Ratio Calibration ( $C_{\text{llr}}$ )

As the calibration performance measures, we use cost of likelihood ratio calibration or  $C_{\text{llr}}$  [15]. The metric  $C_{\text{llr}}$  can be empirically computed by

$$C_{\text{llr}} = \frac{1}{N_{\text{tar}}} \sum_{i \in \text{tar}} \log_2(1 + e^{-l_i}) + \frac{1}{N_{\text{non}}} \sum_{j \in \text{non}} \log_2(1 + e^{l_j}) \quad (5)$$

with  $l_i$  and  $l_j$  as the calibrated log likelihood ratio scores for the target and non-target trials, respectively. Besides  $C_{\text{llr}}$ , we also use two other measures for calibration namely  $C_{\text{llr}}^{\text{min}}$  or the minimum value of  $C_{\text{llr}}$  and  $C_{\text{mc}}$  or mis-calibration cost which is defined as the difference between  $C_{\text{llr}}$  and  $C_{\text{llr}}^{\text{min}}$ . The metric  $C_{\text{llr}}^{\text{min}}$  was computed using *isotonic regression* through pool adjacent violators (PAV) algorithm [16].

## 6. RESULTS

Results of the calibration experiments conducted in this paper are presented in Table 3. Generally in all evaluated datasets, the system tends to perform slightly better in 15 dB noise condition or 6 dB

<sup>5</sup>This corresponds to  $P_{\text{known}} = 0$  for our experiments.

<sup>6</sup>Software is available at <https://sites.google.com/site/bosaristoolkit/>

noise condition (in  $C_{\text{mc}}$  measure). This is because each of the original NIST segments which are included in the I4U lists has two noise variants included in the training of PLDA and enrollment data. The system is biased to perform better in the slightly noisy conditions compared to the unaltered (clean) condition.

The details of results analysis in Table 3 are divided into two parts: the Dev- and Eval-I4U sets, and Eval-SRE'12 set. Those analysis are discussed in the following.

### 6.1. Results on I4U Trials List

In this subsection, we present the analysis of calibration experiment results on the Dev-I4U and Eval-I4U sets. In the Dev-I4U set results, one can observe from Table 3 that all QMF calibrations give lower values across all performance measures than the linear calibration on all noise subsets. This is expected because when we applied the calibration parameters trained on the Dev-I4U set to the Dev-I4U set itself (*self-calibration*).

In the Eval-I4U set results, the QMF calibrations perform well across all performance measures when we compared it to the linear calibration results. Table 3 shows that all the QMF calibrations outperform the linear calibration in terms of  $C_{\text{llr}}^{\text{min}}$  and  $E_{=}$  performance measures. Based on these two performance metrics,  $Q_1$  appears to be the best QMF that provides the best discrimination performance compared to the linear calibration and all other QMF calibrations. The  $Q_1$  calibration results in *absolute* reduction of 0.28 %, 0.47 %, and 0.82 % in the  $E_{=}$  compared to the conventional linear calibration on the clean, 15 dB and 6 dB conditions, respectively. This equals to 10–37 % *relative* improvement in performance.

In the mis-calibration cost or  $C_{\text{mc}}$  metric, the QMFs calibration only perform better than the linear calibration in the clean condition. Even though the  $C_{\text{mc}}$  values for the 15 dB and 6 dB noise conditions for the QMF calibrations are not lower than the linear calibration, still the  $C_{\text{llr}}$  and  $C_{\text{llr}}^{\text{min}}$  values of QMF calibrations are already better than the conventional linear calibration. Using the  $C_{\text{llr}}$  and  $C_{\text{primary}}$  measures, the QMF calibrations perform better than the linear calibration in general, with the  $Q_1$  and  $Q_3$  performances slightly surpass the  $Q_2$  and  $Q_4$  calibrations. Evaluated on the Eval-I4U set, the QMF calibrations offer better performance than the conventional linear calibration based on the observations from all five performance measures.

Comparing all four QMFs for calibration, Table 3 shows that the  $Q_4$  performs the best when calibrations applied in the Dev-I4U set while  $Q_1$  performs the best in the Eval-I4U set. This results indicate that the more complex  $Q_4$  function that model the saddle-plane of calibration parameters distribution does not necessary generalize better than the more simple function such as  $Q_1$ . The  $Q_4$  training has clearly over-fitted to the calibration development set (Dev-I4U). On the other hand, the simple  $Q_1$  function can be easily and effectively implemented in the *cross-calibration*<sup>7</sup>. Regardless of which QMF is the best for calibration, all QMF calibrations indicate better performances in terms of discrimination and calibration when it is compared to the case where duration information are dismissed.

### 6.2. NIST SRE'12 Evaluation Results

The experimental results on the evaluation set from NIST SRE'12 (Eval-SRE'12) are slightly different from results on the Dev-I4U and Eval-I4U sets. As presented in Table 3, the QMF calibrations only surpass the linear calibration performance in the 6 dB noise condition subset. In other noise subsets, applying QMF calibrations does

<sup>7</sup>Applying the calibration parameters which were trained on one set to another set, in this case, from the Dev-I4U set do the Eval-I4U set.


**Table 3.** System performance in terms of  $C_{llr}$ ,  $C_{llr}^{\min}$ ,  $C_{mc}$ ,  $E_{=}$  and  $C_{primary}$  on the Dev-I4U, Eval-I4U and Eval-SRE'12 sets.

Set	Noise cond.	Calibration Method					
		N.A.*	O**	Q1	Q2	Q3	Q4
<i>Cost of log-likelihood ratio calibration (<math>C_{llr}</math>)</i>							
Dev I4U	Clean	4.373	0.195	0.183	0.183	0.192	<b>0.178</b>
	15 dB	2.918	0.078	0.070	0.070	0.071	<b>0.069</b>
	6 dB	6.100	0.115	0.100	0.099	0.103	<b>0.098</b>
Eval I4U	Clean	3.045	0.170	<b>0.148</b>	0.157	0.161	0.172
	15 dB	1.713	0.082	0.078	0.087	<b>0.072</b>	0.110
	6 dB	4.338	0.104	0.089	0.098	<b>0.088</b>	0.117
Eval SRE'12	Clean	11.099	<b>0.194</b>	0.300	0.601	0.306	0.505
	15 dB	5.199	<b>0.133</b>	0.183	0.199	0.145	0.256
	6 dB	8.310	<b>0.179</b>	0.212	0.232	0.180	0.279

<i>Minimum value of <math>C_{llr}</math> (<math>C_{llr}^{\min}</math>)</i>							
Dev I4U	Clean	0.134	0.134	0.130	<b>0.129</b>	0.130	<b>0.129</b>
	15 dB	0.066	0.066	<b>0.057</b>	<b>0.057</b>	0.058	0.058
	6 dB	0.102	0.102	0.089	0.088	0.092	<b>0.087</b>
Eval I4U	Clean	0.113	0.113	<b>0.102</b>	0.105	0.106	0.104
	15 dB	0.052	0.052	<b>0.034</b>	0.036	0.037	0.039
	6 dB	0.086	0.086	<b>0.057</b>	0.061	0.065	0.064
Eval SRE'12	Clean	<b>0.163</b>	<b>0.163</b>	<b>0.163</b>	0.266	0.188	0.244
	15 dB	<b>0.119</b>	<b>0.119</b>	0.129	0.135	0.122	0.145
	6 dB	<b>0.163</b>	<b>0.163</b>	0.173	0.180	0.165	0.194

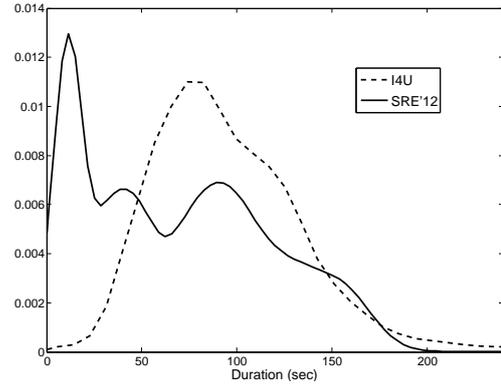
<i>Mis-calibration cost (<math>C_{mc}</math>)</i>							
Dev I4U	Clean	4.239	0.061	0.054	0.054	0.062	<b>0.050</b>
	15 dB	2.852	0.013	0.013	0.013	0.013	<b>0.011</b>
	6 dB	5.998	0.014	<b>0.011</b>	<b>0.011</b>	<b>0.011</b>	<b>0.011</b>
Eval I4U	Clean	2.932	0.057	<b>0.046</b>	0.051	0.055	0.068
	15 dB	1.661	<b>0.029</b>	0.044	0.051	0.036	0.071
	6 dB	4.251	<b>0.018</b>	0.032	0.037	0.023	0.052
Eval SRE'12	Clean	10.936	<b>0.031</b>	0.137	0.335	0.118	0.262
	15 dB	5.080	<b>0.014</b>	0.054	0.064	0.023	0.111
	6 dB	8.147	<b>0.016</b>	0.039	0.052	<b>0.016</b>	0.085

<i>Equal error rate (<math>E_{=}</math>) in %</i>							
Dev I4U	Clean	3.43	3.43	3.34	3.33	3.33	<b>3.32</b>
	15 dB	1.65	1.65	<b>1.35</b>	<b>1.35</b>	1.40	1.36
	6 dB	2.53	2.53	2.25	2.24	2.33	<b>2.17</b>
Eval I4U	Clean	2.78	2.78	<b>2.50</b>	2.53	2.55	2.56
	15 dB	1.27	1.27	<b>0.80</b>	0.81	0.86	0.93
	6 dB	2.21	2.21	<b>1.39</b>	1.48	1.57	1.64
Eval SRE'12	Clean	<b>4.22</b>	<b>4.22</b>	4.36	7.19	5.01	6.43
	15 dB	<b>2.85</b>	<b>2.85</b>	3.15	3.31	2.88	3.67
	6 dB	4.14	4.14	4.40	4.47	<b>4.12</b>	5.11

<i>Primary cost for NIST SRE'12 (<math>C_{primary}</math>)</i>							
Dev I4U	Clean	0.219	0.219	0.173	0.177	0.205	<b>0.171</b>
	15 dB	0.155	0.155	0.153	0.154	<b>0.152</b>	0.163
	6 dB	0.249	0.249	<b>0.235</b>	0.240	0.236	0.252
Eval I4U	Clean	<b>0.174</b>	<b>0.174</b>	0.204	0.381	0.236	0.382
	15 dB	0.148	0.148	0.135	0.137	<b>0.133</b>	0.160
	6 dB	0.254	0.254	<b>0.205</b>	0.215	<b>0.205</b>	0.258
Eval SRE'12	Clean	<b>0.393</b>	<b>0.393</b>	0.485	1.000	0.693	1.000
	15 dB	<b>0.340</b>	<b>0.340</b>	0.371	0.377	0.343	0.411
	6 dB	0.456	0.456	0.484	0.494	<b>0.451</b>	0.533

\* N.A. : Not applicable or no-calibration performed

\*\* O : Conventional linear calibration using  $w_0$  and  $w_1$


**Fig. 2.** Distributions of active speech duration from utterances in the I4U and NIST SRE'12 trials.

not seem to give better performance than the linear calibration. To better understand why this is happening, we had a look into the duration distributions of I4U and NIST SRE'12 segments in more details.

In Figure 2, the duration distributions of utterances included in the I4U and NIST SRE'12 lists are depicted. The durations in the plot is the durations of active speech samples for each utterances after the SAD applied. As can be seen from Figure 2, there is quite a difference between the duration distribution of utterance in the I4U and NIST SRE'12 trials lists. The duration distribution of I4U trials list is more concentrated with mean around 90 s of active speech, while the distribution of NIST SRE'12 trials list is more distributed across all duration with a lot of weight in the short duration region.

The difference in range and distribution of duration between the development and Eval-SRE'12 set may be a cause for the QMFs not working very well, but more likely, the 'data set shift' that occurs with every NIST evaluation may be the most important reason. Indeed, the absolute error rates have gone up strongly from Dev-I4U and Eval-I4U to Eval-SRE'12. The subtle changes to the calibration that the QMFs try to make may be lost in the dramatic changes that take place when the data set changes as drastically as it did in going from SRE'10 to SRE'12—despite the fact that all speakers were known in advance. We have some hope, however, that we will be able to get calibration more in line with the SRE'12 material by looking at other quality factors as well.

## 7. CONCLUSION

Using our development set Dev-I4U to calibrate both the development (self-calibration) and our evaluation set Eval-I4U, the QMF calibration approaches provide a significant performance improvement in both discrimination and calibration. This is observed in all performance metrics used to measure the performance. By adding one or two extra parameters in calibration via the QMF approaches, the system performance based on  $E_{=}$  improves by 37% relative to linear calibration without QMF. However, from the calibration results on the Eval-SRE'12 set using  $P_{known} = 0$ , this does not hold. We surmised that in applying a QMF, it is important that the development set matches the evaluation set in terms of duration range and distribution, so that it can give a positive improvement in the system performance. From the problem revealed by the different duration conditions in the Eval-SRE'12 set from the Dev-I4U set for training calibration, our future works include the *truncation* utterances from the development set or simulate the duration effect such that it can be used to better model the duration distribution in the SRE'12 evaluation set.



## 8. REFERENCES

- [1] A. Neustein and H. A. Patil, *Forensic speaker recognition*, Springer, 2011.
- [2] J. Gonzalez-Rodriguez and D. Ramos, “Forensic automatic speaker classification in the coming paradigm shift,” *Speaker Classification I*, pp. 205–217, 2007.
- [3] National Institute of Standards and Technology, *The NIST Year 2012 Speaker Recognition Evaluation Plan*, Available at <http://www.nist.gov/itl/iad/mig/sre12.cfm>.
- [4] M. I. Mandasari, R. Saeidi, M. McLaren, and D. A. van Leeuwen, “Quality measure functions for calibration of speaker recognition system in various duration conditions,” *Accepted to IEEE Trans. on Audio Speech and Language Processing*, August 2013.
- [5] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, “The effect of noise on modern automatic speaker recognition systems,” in *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4249–4252.
- [6] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, “Cosine similarity scoring without score normalization techniques,” in *Proc. of Odyssey, Speaker and Language Recognition Workshop*. IEEE, 2010, pp. 71–75.
- [7] S. J. D. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *Proc. of Int. Conf. on Computer Vision (ICCV)*. IEEE, 2007, pp. 1–8.
- [8] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brümmer, “Discriminatively trained probabilistic linear discriminant analysis for speaker verification,” in *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4832–4835.
- [9] R. Saeidi and D. A. van Leeuwen, “The Radboud University Nijmegen submission to NIST SRE-2012,” in *Proc. of the NIST Speaker Recognition Evaluation Workshop*, 2012.
- [10] I. Cohen, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.
- [11] J. Pelecanos and S. Sridharan, “Feature warping for robust speaker verification,” in *Proc. of Odyssey, Speaker and Language Recognition Workshop*. IEEE, 2001, pp. 213–218.
- [12] M. McLaren and D. A. van Leeuwen, “A simple and effective speech activity detection algorithm for telephone and microphone speech,” in *Proc. of the NIST Speaker Recognition Evaluation Workshop*, 2011.
- [13] R. Saeidi and et. al., “I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification,” in *Proc. of Interspeech*. IEEE, 2013.
- [14] S. Pigeon, P. Druyts, and P. Verlinde, “Applying logistic regression to the fusion of the NIST’99 1-speaker submissions,” *Digital Signal Processing*, vol. 10, no. 1-3, pp. 237–248, 2000.
- [15] D. A. van Leeuwen and N. Brümmer, “An introduction to application-independent evaluation of speaker recognition systems,” *Speaker Classification I*, pp. 330–353, 2007.
- [16] N. Brümmer and J. du Preez, “Application-independent evaluation of speaker detection,” *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.



# Estimation of cylinder quality measures from quality maps for Minutia-Cylinder Code based latent fingerprint matching

*M. Hamed Izadi, Andrzej Drygajlo*

Swiss Federal Institute of Technology (EPFL)  
CH-1015 Lausanne, Switzerland

{hamed.izadi, andrzej.drygajlo}@epfl.ch

## Abstract

Poor quality of fingerprint data is one of the major problems concerning latent fingerprint matching in forensic applications. Local quality of fingerprint plays a very important role in this application field to ensure high recognition performance. Although big progress has been made in matching of fingerprints using local minutiae descriptors, in particular Minutia Cylinder-Code (MCC), automatic latent fingerprint matching continues to be a challenge. Previously we proposed a matching algorithm that uses minutiae information encoded by MCC with integrated local quality measures associated to each MCC called cylinder quality measures. In our previous work, cylinder quality measures for latent case have been proposed by combining the subjective qualities of individual minutiae involved. In this paper, we propose an alternative method to estimate the cylinder quality measures directly from fingerprint quality maps, in particular ridge clarity maps, by taking into account the number of involving minutiae as well. Integration of MCC with the proposed cylinder quality measures was evaluated through experiments on the latent fingerprint database NIST SD27. These experiments show clear improvements in the identification performance of latent fingerprints of ugly quality.

## 1. Introduction

Fingerprint is one of the most widely used biometric traits for personal identification. For over one hundred years, it has been accepted as an important source of evidence for forensic human identification in the law enforcement agencies worldwide [1].

There exist three main types of fingerprint in forensics applications: 1) rolled, which is collected by rolling a finger from nail to nail on the capturing surface; 2) plain, which is collected by pressing a finger down on the capturing surface; 3) latent, which is collected from surfaces at crime scenes. Rolled and plain fingerprints have usually large amount of ridge details, which are normally believed to be sufficient information for identification. Latent fingerprints have usually the least amount of details and information available for identification.

In latent fingerprint identification, a latent fingerprint is usually compared with the rolled/plain fingerprints already registered in a database. Despite the fact that the recognition performance of Automated Fingerprint Identification Systems (AFISs) has been improved a lot for rolled/plain fingerprints, human intervention is still necessary for latent fingerprint identification, especially in feature extraction. Therefore one of the main challenges is a large number of cases, in particular high-profile cases, where forensic experts are usually under time pressure when identifying fingerprints. Therefore, it is very important that the fingerprints sent to a final visual comparison be

carefully selected so that forensic expert can spend an adequate amount of time for their final examination. One way to achieve this goal is to design an efficient automatic latent to rolled fingerprint matching system that is able to provide a quantitative estimate of the similarity between the latent and rolled prints taking into account a quality of latent fingerprint. In this case, a “Semi-Lights-Out System” [2] can be developed, where some human intervention is allowed during minutiae extraction from a latent fingerprint. The system then outputs a short list of candidates by ranking that need to be examined by the forensic expert to determine fingerprints of the highest visual similarity.

Highly discriminative minutiae-based matching is the most widely adopted approach in forensic fingerprint identification [3, 4]. However latent fingerprint matching is extremely challenging mainly due to: 1) poor quality of latent prints for which the clarity of ridge impressions is low, 2) small finger area in latent prints to be compared to rolled prints of big finger area, 3) large nonlinear distortion because of pressure variations [2].

The local minutia matching techniques have been proposed in order to address some weaknesses of global minutiae matching [3] like non-robustness to nonlinear distortions, need for accurate global alignments, and high computational effort. These techniques are based on local minutiae structures which are descriptors encoding the relationships between each minutia and its neighboring minutiae in terms of some measures invariant to rotation and translation. The local minutiae structures define a region around each minutia by considering, e.g., all minutiae closer than a given radius (fixed-radius based techniques). In particular, Minutiae Cylinder-Code (MCC) representation, recently proposed in [5], obtained remarkable performance with respect to state-of-the-art local minutiae descriptors [6, 7]. One way to improve the accuracy of latent fingerprint identification is to combine modern local minutia matching techniques with subjective or objective local quality measures [8, 9].

Latent fingerprint quality has a significant impact on matching accuracy of fingerprint identification systems [2]. Local quality measures, assigned to the blocks within the image, are mainly combined into a global quality measure for the whole fingerprint image [10, 11]. The local quality measures can be also employed directly in the local matching, for example, to weight the local scores for their contribution to the global matching score. Incorporating local quality measures in minutiae-based matching has been an interesting problem studied in this context [12, 13, 14, 9]. In our previous work [15], we proposed that cylinder quality measures for latent fingerprints can be computed by combining the subjective qualities of the involving minutiae. Recently Yoon et al. in [11] proposed an approach to generate ridge clarity maps for latent fingerprints. Based on their empirical results, they concluded that the (aver-



age) ridge clarity together with the number of minutiae are the two significant features representing the latent fingerprints. In this paper, we propose a new method to estimate the cylinder quality measures directly from fingerprint quality maps, in particular the ridge clarity maps proposed in [11]. We have also taken into account the number of contributing minutiae to the newly proposed measure, which is showed to be an important quality factor.

The rest of this paper is organized as follows. Section 2 provides a brief introduction to fingerprint quality maps and ridge clarity maps for latent fingerprints. Minutiae Cylinder Code (MCC) and its integration with the cylinder quality measures are introduced in Section 3. Then in Section 4, we propose new methods for obtaining cylinder quality measures from fingerprint quality maps. The experiments and results are presented in Section 5. Finally, conclusions are drawn in Section 6.

## 2. Fingerprint quality maps

Fingerprint quality estimation methods are generally categorized into three different groups [16]: 1) those based on local features; 2) those based on global features; 3) those based on a classification approach. The main focus of this paper is on the local approaches, which rely on local features of the fingerprint image. In such methods, the image is usually divided into nonoverlapping square blocks and the quality features are extracted from each block, resulting in a quality map where a local measure of quality is assigned to each block. This measure usually represents the clarity of the ridges and valleys, which can also estimate the extractability of the fingerprint features such as minutiae. Local orientation, Gabor filters, pixel intensity statistics, power spectrum and their combination have been the main local features already used for local quality assessment of the fingerprint images [16].

### 2.1. Ridge clarity maps

Although a variety of methods have been introduced for measuring the local quality of rolled/plain fingerprints, it is rather challenging to objectively assess the local quality of latent fingerprints due to missing ridge structures, mixture of ridge patterns, severe background noise, etc [11]. Recently in [11], Yoon et al. proposed an average ridge clarity measure for latent fingerprints, which is computed from a ridge clarity map containing the local clarity estimates in  $16 \times 16$  pixel blocks within the latent fingerprint image. The ridge clarity in each block is estimated based on the 2-dimensional sine wave representations of the ridges in the form of  $\omega(x, y) = a * \sin(2\pi f(x \cos \theta + y \sin \theta) + \phi)$  [17]. The procedure to obtain the ridge clarity map briefly consists of the four following steps:

1. Contrast enhancement of the original fingerprint image.
2. Estimating the parameters of the 2-D sine wave representations ( $a, f, \theta, \phi$ ) for each block corresponding to the top two local amplitude maxima of its Fourier spectrum within the frequency range of  $[\frac{1}{16}, \frac{1}{5}]^1$ .
3. Obtaining a ridge continuity map by evaluating continuity conditions of the 2-D sine waves in adjacent blocks.
4. Computing the ridge clarity map via multiplying the ridge continuity of each block by the highest amplitude of its 2-D sine waves.

<sup>1</sup>The frequency range is selected based on the fact that the dominant ridge frequencies in a fingerprint image are generally in this range.

In Figure 1, the ridge clarity maps are shown for a latent fingerprint and a rolled tenprint from NIST SD27.

## 3. MCC based latent fingerprint matching

### 3.1. MCC

Local minutiae descriptors became very popular in modern fingerprint matching techniques. Among recently proposed local minutiae descriptors, Minutia Cylinder-Code (MCC) [5] has shown a relatively high performance for local matching [6]. MCC is a fixed-length minutiae descriptor which encodes the relationships between a central minutia and its neighboring minutiae within a fixed-radius circular area. It is to some extent invariant to rotation and translation, robust to skin distortions, and can be computed rather fast.

Given a set of minutiae (a minutiae template), the distance of one minutia to all its neighboring minutiae is considered as basis for creating the MCC. In addition to distance, the angular difference is taken into account using an additional dimension. Finally for each minutia a discretized 3D cylinder-shaped structure is created as MCC descriptor, whose base and height are related to the spatial and directional information, respectively.

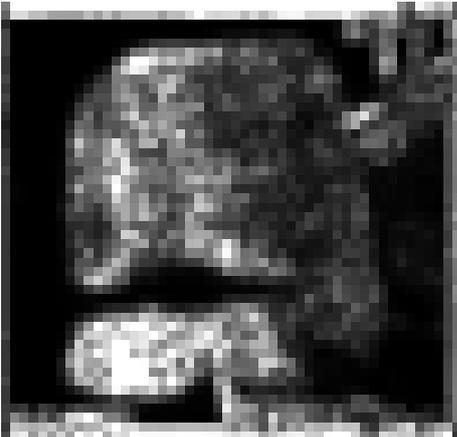
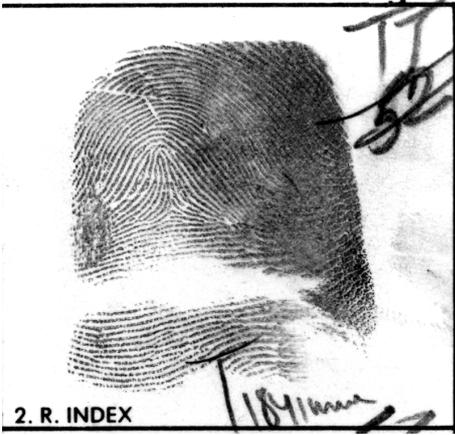
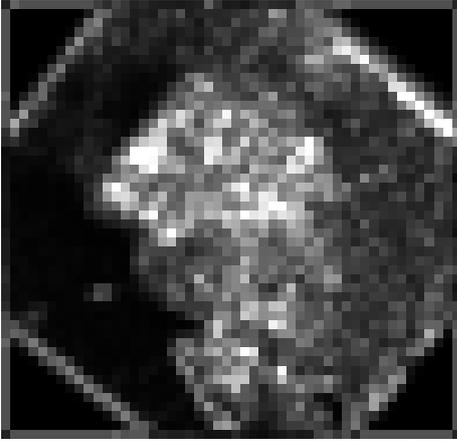
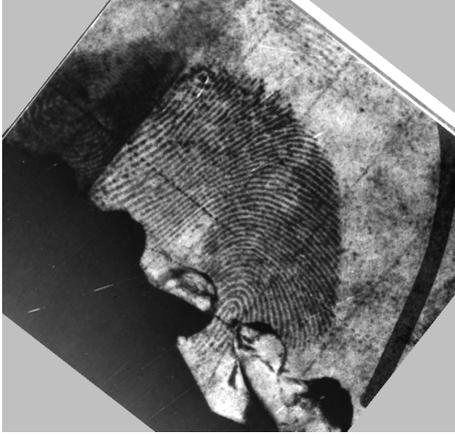
### 3.2. Cylinder quality measures for MCC based matching

In this section, we describe our proposed method to incorporate the cylinder quality measures into MCC based matching process. The MCC based matching normally begins with computing all local similarities for every possible pair of cylinders from the two templates to be compared. Then these local similarities are combined into an overall similarity score between two minutiae templates (e.g., one corresponding to the latent and the other one corresponding to the rolled fingerprint). In [5, 7] several methods have been introduced to obtain a global score from the local similarities, almost all of them involve a careful selection of candidates among all possible pairs and then averaging over their (relaxed) similarities. The cylinder quality measures can be employed here in several ways: they can be used for example to reject the very low-quality cylinders from the list of potential candidates, or they can be applied as weights to the local similarities for candidate selection or final averaging. One approach can be to weight the local similarity between two cylinders based on their pairwise quality. Therefore, if the quality is high for cylinders in a pair, this pair gains more chance to be selected as a candidate and will contribute more to the global score than a pair of low quality cylinders. However this simple weighting strategy might not be so helpful mainly because the incompatible pairs that initially obtained a high similarity by chance might obtain even higher contribution if they are of high quality as well. What we propose here, to reduce the effects of this deficiency, is to utilize the cylinder qualities after a pre-selection phase through a relaxation approach like in [18, 5].

More precisely, given two templates of MCC descriptors, say  $L = \{l_1, l_2, \dots, l_{n_L}\}$  corresponding to a latent fingerprint and  $T = \{t_1, t_2, \dots, t_{n_T}\}$  corresponding to a rolled tenprint, the global matching *Score* between the two templates is computed through the following steps<sup>2</sup>:

1. The local similarities between all possible pairs of descriptors respectively from the two templates ( $n_L \times n_T$  pairs in total) are computed as in [5].

<sup>2</sup>All MCC related parameters used in the formulations, have been named same as in [5], except otherwise stated.



- The  $n_R$  pairs having normally the top local similarities are pre-selected, e.g., using a Local Greedy Similarity (LGS) algorithm [7]. Note that  $n_R$  is usually greater than the number of final pairs ( $n_P$ ) that contribute to the global score. Let  $P$  be the set of all selected pairs:

$$P = \{(l_{r_j}, t_{c_j})\}, j = 1, \dots, n_R,$$

$$1 \leq r_j \leq n_L, 1 \leq c_j \leq n_T, n_R = \min \{n_L, n_T\}.$$

- Through the relaxation phase, the local similarity of each pair is iteratively being modified based on its global relationship with the other pairs as follows: assuming  $\lambda_j^{(0)}$  to be the initial similarity of pair  $j$  (i.e.,  $(l_{r_j}, t_{c_j})$ ), the modified local similarity at iteration  $i$  of the relaxation procedure is:

$$\lambda_j^{(i)} = \omega_R \cdot \lambda_j^{(i-1)} + \left( \frac{1 - \omega_R}{n_R - 1} \right) \cdot \sum_{\substack{k=1 \\ k \neq j}}^{n_R} (\rho(j, k) \cdot \lambda_k^{(i-1)}), \quad (1)$$

where  $\omega_R$  is a weighting parameter and  $\rho(j, k)$  is the measure of compatibility between two pairs:  $(l_{r_j}, l_{r_k})$  and  $(t_{c_j}, t_{c_k})$ , and can be computed as explained in [5] considering also its distortion-tolerant version in [7]. After executing  $n_{rel}$  iterations on all  $n_R$  pairs existing in  $P$ , the quality-based efficiency of pair  $j$  is calculated as:

$$qe_j = \frac{\lambda_j^{(n_{rel})}}{\lambda_j^{(0)}} \cdot Q_j, \quad (2)$$

where  $Q_j$  is a pairwise quality measure for pair  $j$ , and thus depends on both  $Q_{l_{r_j}}$  and  $Q_{t_{c_j}}$ , that are quality measures corresponding to the MCC descriptors  $l_{r_j}$  and  $t_{c_j}$  respectively.

- The  $n_P$  pairs with the highest quality-based efficiency  $qe_j$  are selected, and the final score is computed using a weighted average with pairwise qualities  $Q_j$  as weights:

$$Score = \frac{\sum_{j=1}^{n_P} (Q_j \cdot \lambda_j^{(n_{rel})})}{\sum_{j=1}^{n_P} Q_j}. \quad (3)$$

With the definition given for quality-based efficiency, the final pairs are selected taking into account both factors of quality and compatibility with other pairs. It can also address the previously discussed problem of the pairs that obtained randomly an initial high similarity, by penalizing them in the relaxation process.

#### 4. Cylinder quality measures from fingerprint quality maps

In our previous work [15], it was proposed that cylinder quality measures for latent fingerprints can be computed by combining the subjective qualities of the involving minutiae. In this section, we propose a new method to estimate the cylinder quality measures directly from fingerprint quality maps, in particular the ridge clarity maps described in Section 2.1. In the proposed measure, we also take into account the number of contributing minutiae to each cylinder. This is inspired by the empirical results in [11], based on which Yoon et al. concluded that the number of minutiae and the (average) ridge clarity are the

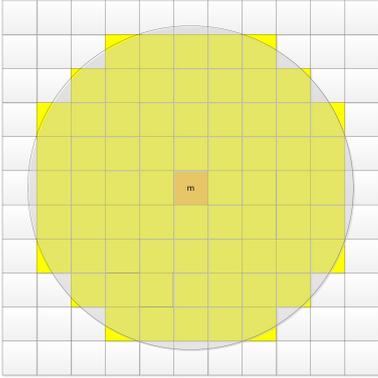


Figure 2: An example showing the  $16 \times 16$  pixel blocks considered for computing the average ridge clarity within a cylinder of radius 75 pixels.

two significant features representing the quality of latent fingerprints. Our method is generally based on averaging the entries of the quality map within the area of a given cylinder. More specifically, given a cylinder  $C_m$  with the radius  $R$  pixels centered at minutia  $m$ , the cylinder quality measure,  $Q_{C_m}$ , is defined as:

$$Q_{C_m} = N_{C_m} \cdot \tilde{Q}_{C_m}, \quad (4)$$

where  $N_{C_m}$  is the number of minutiae contributing to the cylinder and  $\tilde{Q}_{C_m}$  is an average quality contribution estimated from the fingerprint quality map.

Considering  $\mathbf{QM}[xb, yb]$  to be quality of the block residing in the row  $xb$  and the column  $yb$  of the quality map, we propose two approaches for estimating  $\tilde{Q}_{C_m}$ . In the first approach,  $\tilde{Q}_{C_m}$  is estimated based on the average quality in the blocks containing the minutiae involved in the cylinder, as follows:

$$\tilde{Q}_{C_m} = \frac{1}{N_{C_m}} \sum_{i=1}^{N_{C_m}} \mathbf{QM}[xb_i, yb_i], \quad (5)$$

where  $[xb_i, yb_i]$  is the block containing the  $i$ -th minutia contributing to the cylinder  $C_m$ .

For example, if we consider  $16 \times 16$  pixel blocks for the quality map, and a cylinder with radius 75 pixels fully inside the convex hull of all minutiae, the final blocks are those containing minutiae among the 69 blocks around the central block (including itself) as depicted in Figure 2.

With the cylinder quality measures defined in Eq. (4), we have taken into account both the average ridge clarity and the number of minutiae within the cylinder area.

## 5. Experiments and results

### 5.1. Database

For our evaluations, we have considered the only publicly available database of latent fingerprints, NIST SD27 [19], which contains 258 latent fingerprint cases and their corresponding rolled tenprints. The cases are divided into three general categories: Good (88 cases), Bad (85 cases), and Ugly (85 cases) based on the overall quality of latent fingerprint images evaluated by examiners. The minutiae information (position and direction) is provided for all fingerprints in the database. The tenprints minutiae are extracted by an automatic AFIS system

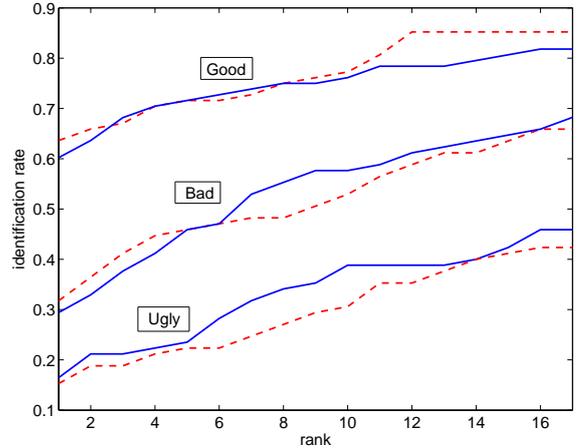


Figure 3: CMC curves showing the identification performance on Good, Bad and Ugly parts of NIST SD27 in presence (blue solid curves) and absence (red dashed curves) of cylinder quality measures.

and the latent minutiae are marked manually by the professional latent examiners.

### 5.2. Evaluations

The bit-based MCC descriptors (MCC16b) were firstly created using the parameters given in [7]. We also implemented the MCC based matching algorithm as proposed in [7], which involves a distortion-tolerant relaxation part. For our comparative evaluations, we consider two different scenarios with and without incorporating cylinder quality measures in the matching. The case without quality is equivalent to consider equal qualities for all cylinders. To incorporate cylinder quality, we firstly generate ridge clarity maps for all latent and rolled fingerprint images as explained in Section 2.1. Then for each MCC descriptor, a cylinder quality measure is computed according to Eqs (4) and (5) given in Section 4. The cylinder quality measures are used to obtain the pairwise quality ( $Q_j$ ) needed for matching (see Eqs (2) and (3)). Here we also consider a common form of  $Q_j = \sqrt{Q_{l_{r_j}} \times Q_{t_{c_j}}}$  as pairwise quality of the latent-rolled cylinder pair  $j$ . This means that the cylinder qualities of both latent and rolled cases have been taken into account in our experiments.

The Cumulative Match Characteristic (CMC) curves (identification rate vs. rank) are shown in Figure 3 separately for Good, Bad, and Ugly images within the NIST SD27 database both in presence and absence of cylinder quality measures. Incorporating the proposed cylinder quality measures shows an improvement in the identification performance for the Ugly category and also for the Bad category above some given rank. For example, the rank-10 identification rate is increased from 30.59% to 38.82% for the Ugly category and from 52.94% to 57.65% for the Bad category. On the other hand, there are some failures specially for the Good category which has also the highest average number of minutiae per image. This could be due to the fact that the estimated cylinder qualities from ridge clarity maps are not discriminating enough among different parts of the images in this category, thus being not sufficiently informative



regarding the automatic identification task.

## 6. Conclusions

Latent fingerprint matching is a biometric technique widely used in forensic applications, while it is still far from being reliable using fully automatic systems. In this paper, we combine the modern MCC based matching technique with objective local quality measures for the latent fingerprint images. We proposed a method to estimate the cylinder quality measures directly from fingerprint quality maps, in particular ridge clarity maps, by averaging the ridge clarity within the cylinder area and also taking into account the number of minutiae involved. The experiments on the NIST SD27 database showed that incorporating the estimated cylinder quality measures through the quality-based relaxation approach proposed by the authors in [9] can improve the identification performance for latent fingerprints in the Ugly category, while being not so effective for the Good category. It is however to be expected since the cylinder qualities in the Good category vary less than in the other categories, and thus there is not much to be gained there by using the proposed weighting method.

## 7. Acknowledgments

The authors would like to thank He Xu for her contribution in the code to generate ridge clarity maps, and also Leila Mirmohamadsadeghi for sharing her MCC codes. This work was partly supported by the Swiss National Science Foundation (SNSF) through the grants 200020-127321 and 200020-146826.

## 8. References

- [1] C. Champod, C. J. Lennard, P. A. Margot, and M. Stoilovic, *Fingerprints and Other Ridge Skin Impressions*. Boca Raton: CRC Press, 2004.
- [2] A. Jain and J. Feng, “Latent Fingerprint Matching,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 88–100, January 2011.
- [3] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, *Handbook of Fingerprint Recognition*, 2nd ed. Springer, 2009.
- [4] R. Cappelli, M. Ferrara, and D. Maltoni, “Minutiae-Based Fingerprint Matching,” in *Cross Disciplinary Biometric Systems*, C. Liu and V. Mago, Eds. Springer Berlin Heidelberg, 2012, vol. 37, pp. 117–150.
- [5] —, “Minutia Cylinder-Code: A New Representation and Matching Technique for Fingerprint Recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2128–2141, December 2010.
- [6] J. Feng and J. Zhou, “A Performance Evaluation of Fingerprint Minutia Descriptors,” in *International Conference on Hand-Based Biometrics (ICHB)*, November 2011.
- [7] R. Cappelli, M. Ferrara, D. Maltoni, and M. Tistarelli, “MCC: a Baseline Algorithm for Fingerprint Verification in FVC-onGoing,” in *International Conference on Control Automation Robotics Vision (ICARCV)*, December 2010, pp. 19–23.
- [8] E. Tabassi, C. L. Wilson, and C. I. Watson, “Fingerprint Image Quality,” *NISTIR 7151*, 2004.
- [9] M. H. Izadi, L. Mirmohamadsadeghi, and A. Drygajlo, “Introduction of Cylinder Quality Measure into Minutia Cylinder-Code based Fingerprint Matching,” in *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, September 2012.
- [10] N. Ratha and R. Bolle, “Fingerprint Image Quality Estimation,” *IBM Computer Science Research Report, RC21622*, 1999.
- [11] S. Yoon, E. Liu, and A. K. Jain, “On Latent Fingerprint Image Quality,” in *ICPR International Workshop on Computational Forensics (IWCF)*, November 2012.
- [12] Y. Chen, S. C. Dass, and A. K. Jain, “Fingerprint quality indices for predicting authentication performance,” in *International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, July 2005, pp. 160–170.
- [13] J. Chen, F. Chan, and Y.-S. Moon, “Fingerprint Matching with Minutiae Quality Score,” in *Advances in Biometrics*, S.-W. Lee and S. Li, Eds. Springer Berlin Heidelberg, 2007, vol. 4642, pp. 663–672.
- [14] K. Cao, E. Liu, L. Pang, J. Liang, and J. Tian, “Fingerprint matching by incorporating minutiae discriminability,” in *International Joint Conference on Biometrics (IJCB)*, October 2011.
- [15] M. H. Izadi and A. Drygajlo, “Embedding cylinder quality measures into minutia cylinder-code based latent fingerprint matching,” in *Proceedings of the ACM Workshop on Multimedia and Security (MMSEC)*. ACM, 2012, pp. 33–38.
- [16] F. Alonso-Fernandez, J. Fierrez, J. Ortega-Garcia, J. Gonzalez-Rodriguez, H. Fronthaler, K. Kollreider, and J. Bigun, “A Comparative Study of Fingerprint Image-Quality Estimation Methods,” *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 4, pp. 734–743, December 2007.
- [17] A. Jain and J. Feng, “Latent Palmprint Matching,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 1032–1047, 2009.
- [18] Y. Feng, J. Feng, X. Chen, and Z. Song, “A Novel Fingerprint Matching Scheme Based on Local Structure Compatibility,” in *International Conference on Pattern Recognition (ICPR)*, 2006, pp. 374–377.
- [19] NIST Special Database 27, “Fingerprint Minutiae from Latent and Matching Tenprint Images,” <http://www.nist.gov/itl/iad/ig/sd27a.cfm>.



# Can Facial Uniqueness be Inferred from Impostor Scores?

Abhishek Dutta, Raymond Veldhuis, Luuk Spreeuwers

University of Twente, Netherlands

{a.dutta,r.n.j.veldhuis,l.j.spreeuwers}@utwente.nl

## Abstract

In Biometrics, facial uniqueness is commonly inferred from impostor similarity scores. In this paper, we show that such uniqueness measures are highly unstable in the presence of image quality variations like pose, noise and blur. We also experimentally demonstrate the instability of a recently introduced impostor-based uniqueness measure of [Klare and Jain 2013] when subject to poor quality facial images.

## 1. Introduction

The appearances of some human faces are more similar to facial appearances of other subjects in a population. Those faces whose appearance is very different from the population are often called a unique face. Facial uniqueness is a measure of distinctness of a face with respect to the appearance of other faces in a population. Non-unique faces are known to be more difficult to recognize by the human visual system [1] and automatic face recognition systems [2, Fig. 6]. Therefore, in Biometrics, researchers have been actively involved in measuring uniqueness from facial photographs [2, 3, 4, 5]. Such facial uniqueness measurements are useful to build an adaptive face recognition system that can apply stricter decision thresholds for fairly non-unique facial images which are much harder to recognize.

Most facial uniqueness measurement algorithms quantify the uniqueness of a face by analyzing its similarity score (i.e. impostor score) with the facial image of other subjects in a population. For example, [2] argue that a non-unique facial image (i.e. lamb<sup>1</sup> as defined in [6]) “will generally exhibit high level of similarity to many other subjects in a large population (by definition)”. Therefore, they claim that facial uniqueness of a subject can be inferred from its impostor similarity score distribution.

In this paper, we show that impostor scores are not only influenced by facial identity (which in turn defines facial uniqueness) but also by quality aspects of facial images like pose, noise and blur. Therefore, we argue that a facial uniqueness measure based solely on impostor scores may give misleading results for facial images degraded by quality variations.

The organization of this paper is as follows: in Section 2, we review some existing methods that use impostor scores to measure facial uniqueness, next in Section 3 we describe the experimental setup that we use to study the influence of facial identity and image quality on impostor scores, in Section 4 we investigate the stability of one recently introduced impostor-based uniqueness measure (i.e. [2]). Finally, in Section 5, we

This work was supported by the BBfor2 project which is funded by the EC as a Marie-Curie ITN-project (FP7-PEOPLE-ITN-2008) under Grant Agreement number 238803.

<sup>1</sup>sheep: easy to distinguish given a good quality sample, goats: have traits difficult to match, lambs: exhibit high levels of similarity to other subjects, wolves: can best mimic other subject’s traits

discuss the experimental results and present the conclusions of this study in Section 6.

## 2. Related Work

Impostor score distribution has been widely used to identify the subjects that exhibit high level of similarity to other subjects in a population (i.e. lamb). The authors of [6] investigated the existence of “lamb” in speech data by analyzing the relative difference between maximum impostor score and genuine score of a subject. They expected the “lambs” to have very high maximum impostor score. A similar strategy was applied by [5] to locate non-unique faces in a facial image dataset. The authors of [3] tag a subject as “lamb” if its mean impostor score lies above a certain threshold. Based on this knowledge of a subject’s location in the “Doddington zoo” [6], they propose an adaptive fusion scheme for a multi-modal biometric system. Recently, [2] have proposed an Impostor-based Uniqueness Measure (IUM) which is based on the location of mean impostor score relative to the maximum and minimum of the impostor score distribution. Using both genuine and impostor scores, [4] investigated the existence of biometric menagerie in a broad range of biometric modalities like 2D and 3D faces, fingerprint, iris, speech, etc.

All of these methods that aim to measure facial uniqueness from impostor scores assume that impostor score is only influenced by facial identity. In this paper, we show that impostor scores are also influenced by image quality (like pose, noise, blur, etc).

The authors of [7] have also concluded that facial uniqueness (i.e. location in the biometric zoo) changes easily when imaging conditions (like illumination) change. Their conclusion was based on results from a single face recognition system (i.e. FaceVACS [8]). In this paper, we also investigate the characteristics of facial uniqueness using four face recognition systems (two commercial and two open-source systems) operating on facial images containing the following three types of quality variations: pose, blur and noise.

## 3. Influence of Image Quality on Impostor Score Distribution

In this section, we describe an experimental setup to study the influence of image quality on impostor scores. We fix the identity of query image to an average face image synthesized<sup>2</sup> by setting the shape ( $\alpha$ ) and texture ( $\beta$ ) coefficients to zero ( $\alpha, \beta = 0$ ) as shown in Figure 1. We obtain a baseline impostor score distribution by comparing the similarity between the average face and a gallery set (or, impostor population) containing 250 subjects. Now, we vary the quality (pose, noise and blur) of

<sup>2</sup>using the code and model provided with [9]



this gallery set (identity remains fixed) and study the variation of impostor score distribution with respect to the baseline. Such a study will clearly show the influence of image quality on impostor score distribution as only image quality varies while the facial identity remains constant in all the experiments.



Figure 1: Average face image

We use the MultiPIE neutral expression dataset of [10] to create our gallery set. Out of the 337 subjects in MultiPIE, we select 250 subjects that are common in session (01,03) and session (02,04). In other words, our impostor set contains subjects from  $(S_1 \cup S_3) \cap (S_2 \cup S_4)$ , where  $S_i$  denotes the set of subjects in MultiPIE session  $i \in \{1, 2, 3, 4\}$  recording 1. From the group  $(S_1 \cup S_3)$ , we have 407 images of 250 subject and from the group  $(S_2 \cup S_4)$ , we have 413 images of the same 250 subjects. Therefore, for each experiment instance, we have 820 images of 250 subjects with at least two image per subject taken from different sessions.

We compute the impostor score distribution using the following four face recognition systems: FaceVACS [8], Verilook [11], Local Region PCA and Cohort LDA [12]. The first two are commercial while the latter two are open source face recognition systems. We supply the same manually labeled eye coordinates to all the four face recognition systems in order to avoid the performance variation caused by automatic eye detection error.

In this experiment, we consider impostor population images with frontal view (cam 05\_1) and frontal illumination (flash 07) images as the baseline quality. We consider the following three types of image quality variations of the impostor population: pose, blur, and noise as shown in Figure 2. For pose, we vary the camera-id (with flash that is frontal with respect to the camera) of the impostor population. For noise and blur, we add artificial noise and blur to frontal view images (cam 05\_1) of the impostor population. We simulate imaging noise by adding zero mean Gaussian noise with the following variances:  $\{0.007, 0.03, 0.07, 0.1, 0.3\}$  (where pixel value is in the range  $[0, 1.0]$ ). To simulate  $N$  pixel horizontal linear motion of subject, we convolve frontal view images with a  $1 \times N$  averaging filter, where  $N \in \{3, 5, 7, 13, 17, 29, 31\}$  (using Matlab's `fspecial('motion', N, 0)` function). For pose variation, camera-id 19\_1 and 08\_1 refer to right and left surveillance view images respectively.

In Figure 4, we report the variation of impostor score distribution of the average face image as box plots [13]. In these box plot, the upper and lower hinges correspond to the first and third quantiles. The upper (and lower) whisker extends from the hinge to the highest (lowest) value that is within  $1.5 \times \text{IQR}$  where IQR is the distance between the first and third quartiles. The outliers are plotted as points.

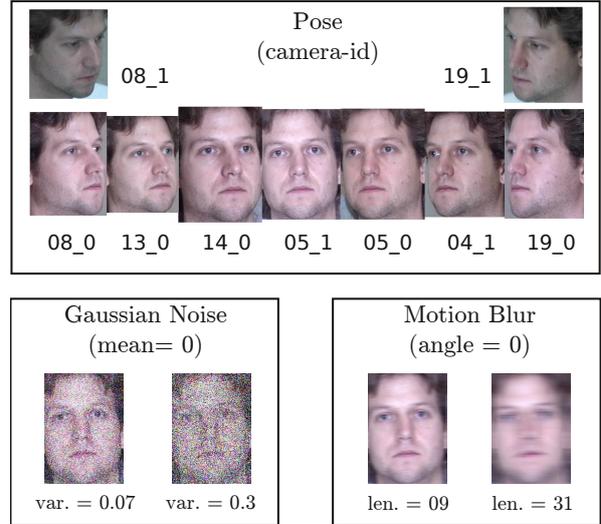


Figure 2: Facial image quality variations included in this study.

#### 4. Stability of Impostor-based Uniqueness Measure Under Quality Variation

In this section, we investigate the stability of a recently proposed impostor-based facial uniqueness measure [2] under image quality variations. The key idea underpinning this method is that a fairly unique facial appearance will result in low similarity score with a majority of facial images in the population. This definition of facial uniqueness is based on the assumption that similarity score is influenced only by facial identity.

This facial uniqueness measure is computed as follows: Let  $i$  be a probe (or query) image and  $J = \{j_1, \dots, j_n\}$  be a set of facial images of  $n$  different subjects such that  $J$  does not contain an image of the subject present in image  $i$ . In other words,  $J$  is the set of impostor subjects with respect to the subject in image  $i$ . If  $S = \{s(i, j_1), \dots, s(i, j_n)\}$  is the set of similarity score between image  $i$  and the set of images in  $J$ , then the Impostor-based Uniqueness Measure (IUM) is defined as:

$$u(i, J) = \frac{S_{max} - \mu_S}{S_{max} - S_{min}} \quad (1)$$

where,  $S_{min}, S_{max}, \mu_S$  denote minimum, maximum and average value of impostor scores in  $S$  respectively. A facial image  $i$  which has high similarity with a large number of subjects in the population will have a small IUM value  $u$  while an image containing highly unique facial appearance will take a higher IUM value  $u$ .

For this experiment, we compute the IUM score of 198 subjects common in session 3 and 4 (i.e.  $S_3 \cap S_4$ ) of the MultiPIE dataset. The IUM score corresponding to same identity but computed from two different sessions (the frontal view images without any artificial noise or blur) must be highly correlated. We denote this set of IUM scores as the baseline uniqueness scores. To study the influence of image quality on the IUM scores, we only vary the quality (pose, noise, blur as shown in Figure 2) of the session 4 images and we compute the IUM scores under quality variation. If the IUM scores are stable with image quality variations, the IUM scores computed from session 3 and 4 should remain highly correlated despite quality variation in session 4 images. Recall that the facial identity remains fixed to the same 198 subjects in all these experiments.

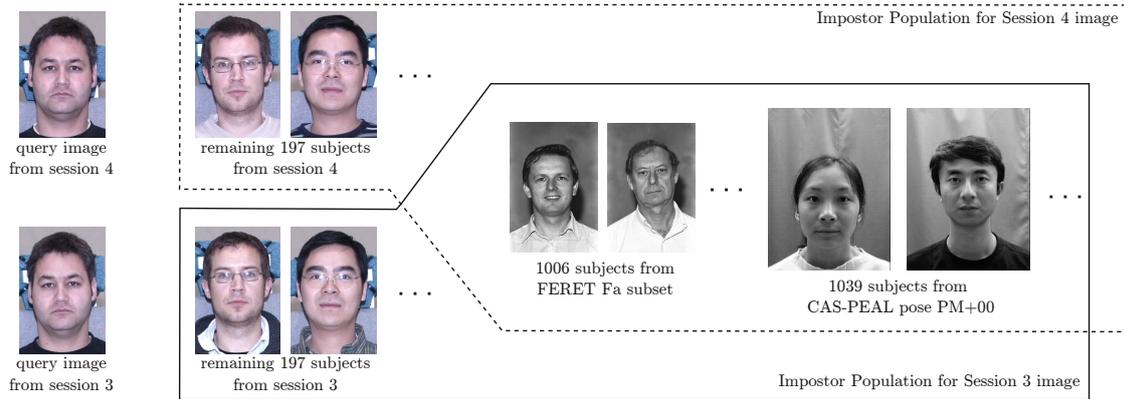


Figure 3: Selection of impostor population for IUM score computation.

In [2], the authors compute IUM scores from an impostor population of  $(16000 - 1)$  subjects taken from a private dataset. We do not have access to such a large dataset. Therefore, we import additional impostors from CAS-PEAL dataset ( $1039^3$  subjects from pose subset PM+00) [14] and FERET (1006 subjects from Fa subset) [15]. So, for computing the IUM score for subject  $i$  in session 3, we have a impostor population containing the remaining 197 subjects from session 3, 1039 subjects from CAS-PEAL and 1006 subjects from FERET. Therefore, each of the IUM score is computed from an impostor set  $J$  containing a single frontal view images of  $197 + 1039 + 1006 = 2242$  subjects as shown in Figure 3. In a similar way, we compute IUM scores for the same 198 subjects but with images taken from session 4. As the Cohort LDA system requires colour images, we replicate the gray scale images of FERET and CAS-PEAL in RGB channels to form a colour image. Note that we only vary the quality of a single query facial image  $i$  (from session 4) while keeping the impostor population quality  $J$  fixed to 2242 frontal view images (without any artificial noise or blur).

In Table 1, we show the variation of Pearson correlation coefficient ( $\text{cor}()$  [16]) between IUM scores of 198 subjects computed from session 3 and 4. The bold faced entries correspond to the correlation between IUM scores computed from frontal view (without any artificial noise or blur) images of the two sessions. The remaining entries denote variation in correlation coefficient when the quality of facial image in session 4 is varied without changing the quality of impostor set. In Figure 5, we show the drop-off of normalized correlation coefficient (derived from Table 1) with quality degradation where normalization is done using baseline correlation coefficient.

## 5. Discussion

### 5.1. Influence of Image Quality on Impostor Score

In Figure 4, we show the variation of impostor score distribution with image quality variations of the impostor population. We consider frontal view (cam 05\_1) image without any artificial noise or blur (i.e. the original image in the dataset) as the baseline image quality. The box plot corresponding to  $\text{cam-id}=05\_1$ ,  $\text{blur-length}=0$ ,  $\text{noise-variance}=0$  denotes mainly the impostor score variation due to facial identity.

<sup>3</sup>in our version of the CAS-PEAL dataset, PM+00 images for person-id 261 in the pose subset were missing. Therefore, we use only 1039 of the total 1040 subjects in the original dataset

In Figure 4, we observe that, the nature of impostor score distribution corresponding to all three types of quality variations is significantly different from the baseline impostor distribution. For instance, the impostor score distribution for FaceVACS and Verilook systems corresponding to a motion blur of length 31 pixels is completely different from that corresponding to no motion blur. Furthermore, the impostor score distribution also seem to be responding to quality variations. For example, the mean of impostor distribution for FaceVACS system appears to increase monotonically as the image quality moves towards the baseline image quality. We also observe that the impostor score distribution of the four face recognition systems respond in a different way to the three types of image quality variations. These observations clearly show that the impostor score distribution is not only influenced by identity (as expected) but also by the image quality like pose, blur and noise.

### 5.2. Stability of Impostor-based Uniqueness Measure Under Quality Variation

We observe a common trend in the variation of correlation coefficients with image quality degradation as shown in Table 1. The correlation coefficient is maximum for the baseline image quality (frontal, no artificial noise or blur). As we move away from the baseline image quality, the correlation between IUM scores reduces. This reduction in correlation coefficient indicates the instability of Impostor-based Uniqueness Measure (IUM) in the presence of image quality variations.

The instability of IUM is also depicted by the normalized correlation coefficient plot of Figure 5. For all the four face recognition systems, we observe fall-off of the correlation between IUM scores with variation in pose, noise and blur of facial images. For pose variation, peak correlation is observed for frontal view (camera 05\_1) facial images because, in this case, both pairs of IUM scores correspond to frontal view images taken from two session 3 and session 4.

The instability of IUM measure is also partly due to the use of minimum and maximum impostor scores in equation (1) which makes it more susceptible to outliers.

The authors of [2], who originally proposed the Impostor-based Uniqueness measure (IUM), report a correlation of  $\geq 0.92$  using FaceVACS system on a privately held mug shot database of 16000 subjects created from the operational database maintained by the Pinellas County Sheriff's Office. Further details about the quality of facial images in this dataset is



not available. From the sample images shown in [2], we can assume that this private mugshot database contains sharp frontal view facial images captured under uniform illumination. Our baseline image quality (frontal view without any artificial blur or noise) comes very close to the quality of images used in their experiment. However, we get a much lower correlation coefficient of  $\leq 0.68$  on a combination of three publicly released dataset. One reason for this drop in correlation may be due to difference in the quality (like resolution) of facial images. Our impostor population is formed using images taken from three publicly available dataset and therefore represents larger variation in image quality as shown in Figure 3. To a lesser extent, this difference in correlation could also be due to difference in the FaceVACS SDK version used in the two experiments. We use the FaceVACS SDK version 8.4.0 (2010) and they have not mentioned the SDK version used in their experiments.

## 6. Conclusion

We have shown that impostor score is influenced by both identity and quality of facial images. We have also shown that any attempt to measure characteristics of facial identity (like facial uniqueness) solely from impostor score distribution shape may give misleading results in the presence of image quality degradation in the input facial images.

This research has thrown up many questions in need of further investigation regarding the stability of existing facial uniqueness measures based solely on impostor scores. More research is needed to better understand the impact of image quality on the impostor score distribution. Such studies will help develop uniqueness measures that are robust to quality variations.

## 7. Acknowledgement

- We would like to thank Cognitec Systems GmbH. for supporting our research by providing the FaceVACS software. Results obtained for FaceVACS were produced in experiments conducted by the University of Twente, and should therefore not be construed as a vendor's maximum effort full capability result.
- We also acknowledge the anonymous reviewers of the BTFS 2013 conference for their valuable feedback.

## 8. References

- [1] Merideth Going and JD Read, "Effects of uniqueness, sex of subject, and sex of photograph on facial recognition," *Perceptual and Motor Skills*, vol. 39, no. 1, pp. 109–110, 1974.
- [2] Brendan F. Klare and Anil K. Jain, "Face recognition: Impostor-based measures of uniqueness and quality," in *Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on*, 2012, pp. 237–244.
- [3] Arun Ross, Ajita Rattani, and Massimo Tistarelli, "Exploiting the doddington zoo effect in biometric fusion," in *Biometrics: Theory, Applications, and Systems, 2009. BTAS'09. IEEE 3rd International Conference on*. IEEE, 2009, pp. 1–7.
- [4] Neil Yager and Ted Dunstone, "The biometric menagerie," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 2, pp. 220–230, 2010.
- [5] M Wittman, P Davis, and PJ Flynn, "Empirical studies of the existence of the biometric menagerie in the frgc 2.0 color image corpus," in *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*. IEEE, 2006, pp. 33–33.
- [6] George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation," in *Proceedings of International Conference on Spoken Language Processing*, 1998.
- [7] Jeffrey Paone, Soma Biswas, Gaurav Aggarwal, and Patrick Flynn, "Difficult imaging covariates or difficult subjects? - an empirical investigation," in *Biometrics (IJCB), 2011 International Joint Conference on*, 2011, pp. 1–8.
- [8] Cognitec Systems, "FaceVACS C++ SDK Version 8.4.0," 2010.
- [9] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter, "A 3D Face Model for Pose and Illumination Invariant Face Recognition," in *Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS.*, 2009, pp. 296–301.
- [10] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker, "Multi-PIE," in *Automatic Face Gesture Recognition, 2008. FG 08. 8th IEEE International Conference on*, 2008, pp. 1–8.
- [11] Neurotechnology, "VeriLook C++ SDK Version 5.1," 2011.
- [12] "CSU Baseline Algorithms - Jan. 2012 Releases," <http://www.cs.colostate.edu/facerec/algorithms/baselines2011.php>.
- [13] Hadley Wickham, *ggplot2: elegant graphics for data analysis*, Springer New York, 2009.
- [14] Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao, "The cas-peal large-scale chinese face database and baseline evaluations," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 38, no. 1, pp. 149–161, 2008.
- [15] P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 10, pp. 1090–1104, oct 2000.
- [16] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013.



# On the Improvements of Uni-modal and Bi-modal Fusions of Speaker and Face Recognition for Mobile Biometrics

*Elie Khoury, Manuel Günther, Laurent El Shafey and Sébastien Marcel*

Idiap Research Institute, Martigny, Switzerland

{elie.khoury,manuel.guenther,laurent.el-shafey,sebastien.marcel}@idiap.ch

## Abstract

The MOBIO database provides a challenging test-bed for speaker and face recognition systems because it includes voice and face samples as they would appear in forensic scenarios. In this paper, we investigate uni-modal and bi-modal multi-algorithm fusion using logistic regression. The source speaker and face recognition systems were taken from the 2013 speaker and face recognition evaluations that were held in the context of the last International Conference on Biometrics (ICB-2013). Using the unbiased MOBIO protocols, the employed evaluation measures are the equal error rate (EER), the half-total error rate (HTER) and the detection error trade-off (DET). The results show that by uni-modal algorithm fusion, the HTER's of the speaker recognition system are reduced by around 35 %, and of the face recognition system by between 15 % and 20 %. Bi-modal fusion drastically boosts recognition by a relative gain of 65 % - 70 % of performance compared to the best uni-modal system.

## 1. Introduction

During the last years, more and more surveillance devices are installed in public places or in private properties to possibly capture crime scenes. Some of them are able to record both image and voice data. Similarly, instant messaging systems such as Skype, Google talk, Yahoo messenger and Facebook messenger support both video and audio plugins. In forensic investigations, usually a human operator would compare these recordings to samples from suspects. The studies in [1, 2] have shown that automatic speaker and face recognition algorithms can outperform humans in comparing speech utterances or images from unfamiliar identities. Therefore, using automatic algorithms to verify suspects' identities based on their voice and their face are favorable in these cases.

Automatic speaker recognition is investigated since the 1970s [3] and regularly evaluated by the National Institute of Standards and Technology (NIST)<sup>1</sup> since 1996. Similarly, automatic face recognition started in the late 1980s [4], and many evaluations were conducted. Since 2000, face recognition vendor tests (FRVT) [5], which

are also executed by NIST,<sup>2</sup> evaluate capabilities of automatic face recognition applications under controlled conditions.

The appropriate framework for the admissibility of the results of automatic forensic face and speaker recognition systems in front of court is to use evidence interpretation, and to standardize the procedures and the protocols for testing [6]. This allows a scientific and logical methodology to clearly determine the capability of the systems and to be conscious of the error rates. Recent scientific efforts have converged towards exploiting the Bayesian approach for the analysis of evidence, such that opinions about the prosecution and defense hypotheses are expressed in the form of posterior probabilities [7]. In this sense, using log-likelihood ratio (LLR) as a degree of support of one hypothesis over the other has become a crucial demand that successful forensic speaker and face recognition systems should afford.

After the success of the first edition in 2010 [8], the Biometric Group at the Idiap Research Institute organized the second edition of speaker [9] and face [10] recognition evaluations in mobile environment. These evaluations are conducted on the MOBIO database, which provides the unique opportunity to analyze two mature biometrics, i.e., speaker and face recognition side by side in a challenging environment. The conditions in MOBIO are closer to forensic scenarios than during NIST evaluations and, hence, it is more suitable to show algorithm capacities for forensic investigations. Two unbiased evaluation protocols exist for MOBIO, which allow a direct comparison of results of different algorithms with figures published in literature.

In total, 12 institutions participated to the speaker recognition evaluation [9], while the face recognition evaluation [10] analyzed 9 systems. All participants of both evaluations had to strictly follow the unbiased evaluation protocols. To assure a fair comparison, during the evaluations the file names of the test data were anonymized so that the participants could not use the name of the probe file to infer identity. In [9, 10], the speaker and face recognition systems were assessed, and [9] already showed that fusing different speaker recognition systems

<sup>1</sup><http://www.nist.gov/itl/iad/mig/sre.cfm>

<sup>2</sup><http://www.nist.gov/itl/iad/ig/frvt-docs.cfm>



outperforms each single system. In this paper, we investigate whether the integration of state-of-the-art speaker and face recognition systems can further improve performance.

The remainder of the paper is as follows. Section 2 introduces the MOBIO database and the evaluation protocols. In section 3, the techniques that were used in systems submitted to the speaker and face evaluations are described briefly, and the employed multi-algorithm and bi-modal fusion technique is detailed. The experimental evaluation is provided in section 4, while section 5 concludes the paper.

## 2. MOBIO Database

The MOBIO database is a bi-modal, face and speaker, video database recorded from 152 people. MOBIO is challenging since the data is acquired on mobile devices with real noise. The extracted images contain faces with uncontrolled illumination, facial expression, near-frontal pose and occlusion, while the extracted speech segments are relatively short, sometimes less than 2 seconds. Therefore, this database is suited to evaluate algorithms under uncontrolled conditions as they would appear in surveillance scenarios. More technical details about the MOBIO database can be found in [11] and on its official web page.<sup>3</sup>

Based on the gender of the clients, two different evaluation protocols *male* and *female* exist. These protocols are identical for speaker and face recognition. Each five recordings per client are used to enroll client models, the remaining recordings serve as probes. Similarity scores are computed between all models and all probes. In order to have unbiased protocols, the clients of the database are split up into 3 different sets: training, development and evaluation, which are statistically detailed in table 1. The use of the sets is restricted to:

**Training set** The data of this set is used to learn the background parameters of the algorithm (projection matrices, background models, etc.). It can also be used as a cohort for score normalization.

**Development set** The data of this set is used to optimize meta-parameters of the algorithm. Scores produced with this set can be exploited to train calibration parameters for system fusion.

**Evaluation set** The data of this set is used for computing the final evaluation performance. No training or tuning is allowed to be performed on this set.

## 3. Multi-algorithm and Bi-modal Fusion

The MOBIO database permits to integrate information from voice and face samples to recognize clients. In [9]

and [10], several state-of-the-art speaker and face recognition systems provided score files for MOBIO. In this section, we give a short summary of the systems that participated in the evaluations.<sup>4</sup> For more detailed information, please refer to [9, 10].

### 3.1. Speaker Recognition

A text-independent speaker recognition system generally contains 3 main modules: feature extraction, modeling and scoring. The feature extraction modules employed in the speaker recognition evaluation [9] included feature computation (MFCC, LFCC, PLP, F0, etc.), voice activity detection (energy-based, phoneme-based, etc.), speech enhancement (spectral subtraction, Wiener filtering, etc.), and feature post-processing (feature warping, cepstral mean and variance normalization, etc.).

The modeling and scoring modules were often related. Different techniques were used in the evaluation. They can be divided into 4 main groups:

- Gaussian mixture modeling (GMM) [12] first estimates a universal background model (UBM), which is then adapted to each client using maximum *a posteriori* (MAP). Scores are computed by estimating the log-likelihood ratio of the probe with regards to the client model and the UBM.
- In Gaussian super-vector (GSV) modeling [13], the mean vectors of the Gaussians are concatenated. Nuisance attribute projection (NAP) [14] is generally used for session compensation. The scores are computed using support vector machines (SVM).
- Inter-session variability (ISV) modeling [15] aims to estimate and eliminate the effects of the session variability. The scoring employed a linear approximation of the log-likelihood ratio.
- Total variability modeling (i-vector) [16] extracts a low-dimensional vector from each of the speech segments. Different i-vector post-processing types like whitening [17], length normalization [18], linear discriminant analysis (LDA) and within-class covariance normalization (WCCN) [19] were used. At scoring level, between i-vectors the cosine distance or probabilistic linear discriminant analysis (PLDA) [20, 21] was computed.

The best system in the evaluation, i. e., Alpineon (cf. table 2(a)) is based on the fusion of 9 different i-vector sub-systems, each with a different set of features. More details about this and the other speaker recognition systems can be found in [9].

<sup>4</sup>Note that after the evaluations some of the participants submitted corrected score files so that the entries in table 2 partially do not correspond to the results published in [9, 10].

<sup>3</sup><http://www.idiap.ch/dataset/mobio>



Table 1: PARTITIONING OF THE MOBIO DATABASE. This table details the number of clients and recordings of the training set, as well as the number of clients and enrollment recordings, and the number of probes for the development and the evaluation set, for the male and female protocols of the MOBIO database.

	Training		Development				Evaluation			
	Clients	Files	Enrollment		Probe		Enrollment		Probe	
	Clients	Files	Clients	Files	Files	Scores	Clients	Files	Files	Scores
male	37	7104	24	120	2520	60480	38	190	3990	151620
female	13	2496	18	90	1890	34020	20	100	2100	42000
Total	50	9600	42	210	4410	94500	58	290	6090	193620

### 3.2. Face Recognition

In the algorithms submitted to the face recognition evaluation [10], the first step of all participants was to align the faces using the provided hand-labeled eye positions. Afterward, different image normalization techniques [22, 23, 24] were used to reduce illumination effects. Various kinds of features like edge information (POEM) [25], Gabor features [26, 27], local binary patterns (LBP) [28], local phase quantization (LPQ) [29] and color information were extracted. The best single-feature based system, i. e., UC-HU (see table 2(b)) learned how to extract features using a convolutional neural network [30].

On top of these features, different kinds of face recognition systems were executed. Several algorithms computed histograms of various kinds and used histogram comparisons to compute scores. Other systems used principal component analysis (PCA) or linear discriminant analysis (LDA) to reduce feature dimensionality. Furthermore, partial least squares (PLS) classifiers or support vector machines (SVM) were trained to enroll models and compare them to probe features.

Additionally, some participants fused face recognition systems of different kinds. The best performing system in [10], i. e., UNILJ-ALP (cf. table 2(b)) was the multi-representation PCA, which fused in total 30 different face recognition sub-systems. For more detailed descriptions of this and all the other submitted face recognition systems, please refer to [10].

### 3.3. Fusion

To fuse different recognition systems, we take the well-known *linear logistic regression* approach, which has successfully been employed to combine heterogeneous speaker [31, 32] and face [33] authentication systems, as well as for bi-modal authentication [34].

Linear logistic regression combines a set of  $Q$  classifiers using the sum rule. Let the probe  $\mathcal{O}_t$  be processed by  $Q$  classifiers, each of which produces an output score  $h_q(\mathcal{O}_t, \mathbf{g}_i)$  between the current probe sample  $\mathcal{O}_t$  and a given client model  $\mathbf{g}_i$ . These scores are fused using a linear combination:

$$h_\beta(\mathcal{O}_t, \mathbf{g}_i) = \beta_0 + \sum_{q=1}^Q \beta_q h_q(\mathcal{O}_t, \mathbf{g}_i) \quad (1)$$

where  $\beta = [\beta_0, \beta_1, \dots, \beta_Q]$  are the fusion weights (also known as regression coefficients).

The coefficients  $\beta$  are computed by estimating the maximum likelihood of the logistic regression model on the scores of the development set. Let  $\mathcal{X}_{\text{cli}}$  be the set of true client access trials, i. e., the set of those pairs  $\mathbf{x} = \{\mathcal{O}_t, \mathbf{g}_i\}$  where the identities of test sample  $\mathcal{O}_t$  and client  $\mathbf{g}_i$  match. Let furthermore  $\mathcal{X}_{\text{imp}}$  be the set of impostor trials, i. e., those pairs where the identities of  $\mathcal{O}_t$  and  $\mathbf{g}_i$  differ. Then, the objective function to maximize is [35]:

$$L(\beta) = - \sum_{\mathbf{x} \in \mathcal{X}_{\text{imp}}} \log(1 + \exp(h_\beta(\mathbf{x}, \beta))) - \sum_{\mathbf{x} \in \mathcal{X}_{\text{cli}}} \log(1 + \exp(-h_\beta(\mathbf{x}, \beta))) \quad (2)$$

The maximum likelihood estimation procedure converges to a global maximum. In our work, this optimization is done using the conjugate-gradient algorithm [35].

One important fact of the fusion procedure is that the fused scores are already in form of log-likelihood ratios. Hence, after fusing scores from different systems, not only the verification accuracy is increased, but the scores can directly be used to present evidence in front of court.

## 4. Experiments

For several different speaker and face verification algorithms, scores for the MOBIO database are provided in the speaker and face evaluations [9, 10]. To evaluate multi-algorithm and bi-modal system fusion, we executed several experiments using these scores. All experiments are run using the open source software library Bob [36], and we provide both scripts<sup>5</sup> and data<sup>6</sup> for the scientific community affording reproducible research. The first set of experiments assessed the uni-modal multi-algorithm fusion, while the second set of tests fused algorithms of both data types.

### 4.1. Evaluation metrics

The metrics that we use to evaluate verification performance are based on the false acceptance rate (FAR) and the false rejection rate (FRR), which are calculated for

<sup>5</sup><http://pyipi.python.org/pyipi/xbob.paper.BTFS2013>

<sup>6</sup><http://www.idiap.ch/dataset/mobio>



Table 2: RESULTS FROM ICB. *These tables repeat<sup>A</sup> the results from [9] and [10], ordered by EER on the development set of the male protocol from the MOBIO database.*

(a) Speaker recognition

Id	System	male		female	
		EER	HTER	EER	HTER
S-1	Alpineon	<b>5.04</b>	<b>7.08</b>	<b>7.98</b>	<b>10.68</b>
S-2	L2F-EHU	7.89	8.14	11.01	13.59
S-3	Phonexia	9.60	10.78	8.36	14.18
S-4	GIAPSI	9.68	8.86	11.59	12.81
S-5	IDIAP	9.96	10.03	12.01	14.27
S-6	Mines-Telecom	10.20	9.11	11.43	11.63
S-7	L2F	10.60	11.05	13.48	14.73
S-8	EHU	11.31	10.06	17.94	19.51
S-9	CPqD	11.82	10.21	14.35	15.99
S-10	CTDA	12.74	19.40	19.47	22.64
S-11	RUN	13.73	12.13	13.39	14.09
S-12	ATVS	14.88	15.43	16.84	17.86

(b) Face recognition

Id	System	male		female	
		EER	HTER	EER	HTER
F-1	UNILJ-ALP	<b>1.71</b>	7.45	<b>2.75</b>	<b>10.46</b>
F-2	GRADIANT	3.14	9.52	5.38	12.27
F-3	UC-HU	3.49	<b>6.21</b>	4.71	10.83
F-4	CPqD	5.48	7.67	6.30	11.21
F-5	TUT	5.48	10.02	7.35	12.05
F-6	UTS	6.11	11.96	7.46	13.57
F-7	Idiap	6.63	10.29	6.24	12.51
F-8	CDTA	7.65	11.93	10.74	15.90
F-9	baseline	14.80	17.11	14.71	20.94

the development and evaluation sets independently. The definition of these rates depends on a *threshold*  $\theta$ :

$$\begin{aligned} \text{FAR}(\theta) &= \frac{|\{s_{\text{imp}} \mid s_{\text{imp}} \geq \theta\}|}{|\{s_{\text{imp}}\}|} \\ \text{FRR}(\theta) &= \frac{|\{s_{\text{cli}} \mid s_{\text{cli}} < \theta\}|}{|\{s_{\text{cli}}\}|} \end{aligned} \quad (3)$$

Here,  $s_{\text{cli}}$  are client (true target) and  $s_{\text{imp}}$  impostor (non-target) scores, both of which might come from a single speaker or face recognition system, or might have been created by fusing scores of many systems using Eq. (1).

The first evaluation metric is based on the *equal error rate* (EER) on the development set and the *half total error rate* (HTER) on the evaluation set. Particularly, the optimal threshold  $\theta^*$  is based on the EER of the development set, and the HTER is computed using this threshold:

$$\begin{aligned} \theta^* &= \arg \min_{\theta} |\text{FAR}_{\text{dev}}(\theta) - \text{FRR}_{\text{dev}}(\theta)| \\ \text{EER} &= \frac{\text{FAR}_{\text{dev}}(\theta^*) + \text{FRR}_{\text{dev}}(\theta^*)}{2} \\ \text{HTER} &= \frac{\text{FAR}_{\text{eval}}(\theta^*) + \text{FRR}_{\text{eval}}(\theta^*)}{2} \end{aligned} \quad (4)$$

Table 3: UNI-MODAL FUSION. *These tables show the results of uni-modally fusing the N best speaker or face recognition systems from table 2.*

(a) Speaker recognition

Pool of classifiers	male		female	
	EER	HTER	EER	HTER
S-1	5.04	7.08	7.98	10.68
+ S-2	3.81	5.92	6.14	8.85
+ S-3	3.41	5.43	4.18	7.97
+ S-4	3.05	4.76	3.59	6.91
+ S-5	2.86	4.70	3.65	6.87
+ S-6	2.86	4.75	3.65	<b>6.73</b>
+ S-7	2.90	4.75	<b>3.54</b>	6.87
+ S-8	2.90	4.75	3.54	6.89
+ S-9	2.90	4.76	3.59	6.96
+ S-10	2.98	4.97	3.61	7.04
+ S-11	2.81	4.69	3.60	6.87
all	<b>2.78</b>	<b>4.63</b>	3.60	6.87

(b) Face recognition

Pool of classifiers	male		female	
	EER	HTER	EER	HTER
F-1	1.71	7.45	2.75	10.46
+ F-2	1.59	7.30	2.59	10.03
+ F-3	1.50	6.62	2.44	9.99
+ F-4	1.47	6.69	2.49	9.97
+ F-5	1.51	6.89	2.37	10.10
+ F-6	1.47	6.90	2.22	9.31
+ F-7	1.51	6.83	2.12	9.29
+ F-8	1.51	6.72	2.02	9.01
all	<b>1.39</b>	<b>6.27</b>	<b>1.97</b>	<b>8.47</b>

The second type of evaluation is based on the detection error trade-off (DET) [37]. In this curve, the FRR is plotted against the FAR in normal deviate scale. In opposition to receiver operating characteristics (ROC), DET curves allow easy observation of system contrasts, especially for low FAR values. The DET curve of a system is linear when client and impostor score are Gaussian distributed, and with an angle of  $45^\circ$  the variances of the Gaussians are equal.

## 4.2. Uni-modal fusion

The final verification results of the evaluations are given in table 2, where the different systems are sorted according to the EER of the male protocol. Apparently, compared to the best speaker recognition system S-1, the best face recognition system F-1 has lower error rates on the development set, but similar evaluation set errors.

For both the speaker and face recognition systems, we assessed how the fusion of the N best systems improves performance, varying N from 1 to 12 for the speaker and to 9 for the face recognition systems. The results of these experiments can be found in table 3. Clearly, fusing scores from multiple algorithms improves performance. For speaker recognition, incorporating scores from the best two

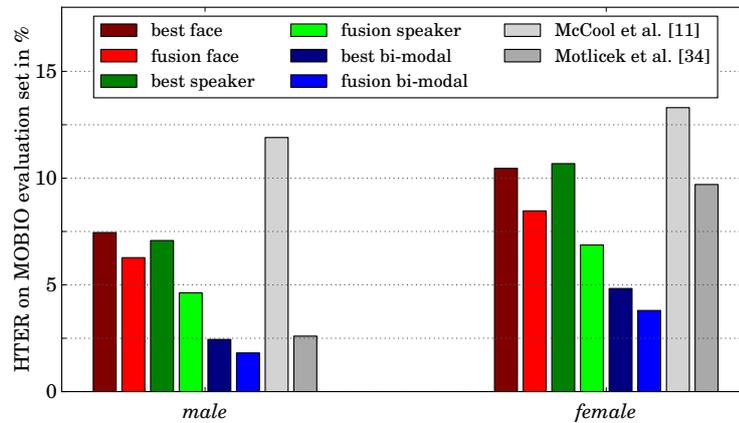


Figure 1: SYSTEM COMPARISON. This figure shows the HTER on the evaluation set of MOBIO for the best speaker and face recognition system and their fusion, as well as the fusions of all speaker, all face and all recognition systems. Additionally, results of the bi-modal systems from [11] and [34] are included.

systems already gains more than one percent of absolute error rate, for the female protocol even around two percent. Incorporating all speaker recognition systems results in a overall relative reduction of around 50% in the development set, and still 30% in the evaluation set.

For face recognition, a similar trend is obtainable, though in general there is a higher difference between development and evaluation set. The best face recognition system has already quite low error rates on the development set. Still, the fusion of all face recognition algorithms results in a relative reduction of around 20% - 30% for both protocols and both sets.

Note that the difference in error rates between the development and evaluation set is an artifact of the small data set and confirms previous findings [34].

### 4.3. Bi-modal fusion

As we showed in [34], fusion of speaker and face recognition systems can tremendously improve performance on the MOBIO database. A similar behavior is observed combining the speaker and face recognition systems provided by the participants of the evaluations. Two different configurations were evaluated:

#### 4.3.1. Fusion of all speaker and face systems

In figure 1, the HTER results on the evaluation set are displayed for the best speaker and face recognition systems, as well as their bi-modal fusion, and the uni-modal and bi-modal fusion of all speaker and face systems. Clearly, uni-modal fusion gives improvements over the best uni-modal system. Bi-modally fusing the best speaker and the best face recognition system outperforms the uni-modal systems significantly, and the fusion of all systems gives by far the best results, which is 0.16 % EER and 1.78 %

HTER for male and 0.16 % EER and 3.80 % HTER for female clients.

In figure 1, the results from [11, 34] are added. While [34] exploits bi-modal single-algorithm fusion, [11] fuses bi-modal multi-algorithm recognition systems using the sum rule. Clearly, both systems are outperformed by our bi-modal multi-algorithm fusion strategy of systems with various type of features that is based on linear logistic regression.

The DET curves, which are shown in figure 2, reveal similar trends. For different working points (different thresholds) the order of the systems is stable, ranging from the single uni-modal systems over the uni-modal fusions to the bi-modal fusions.

#### 4.3.2. Optimal bi-modal fusion

In the second experiment we assessed, which of the submitted algorithms are best suited for fusion. Starting with the best performing system on the development set, i. e., the face recognition system F-1, we tested which other algorithm gets the highest improvement in EER on the development set, i. e., is most complementary to the F-1 system. After finding the speaker recognition systems S-1 (male) and S-5 (female) to be most suitable, we added the next best algorithm and so forth. Figure 3 shows the EER and the according HTER values of fusing the optimal set of systems. From left to right, the indicated system was added to the set of fused algorithms.

Apparently, the biggest gain is in fusing one face and one speaker recognition system, and further adding more systems improves performance only moderately. After fusing approximately 10 systems, which differ between the male and female protocols, performance on the development set settles. Adding more systems does not improve the EER on the development set any more, though

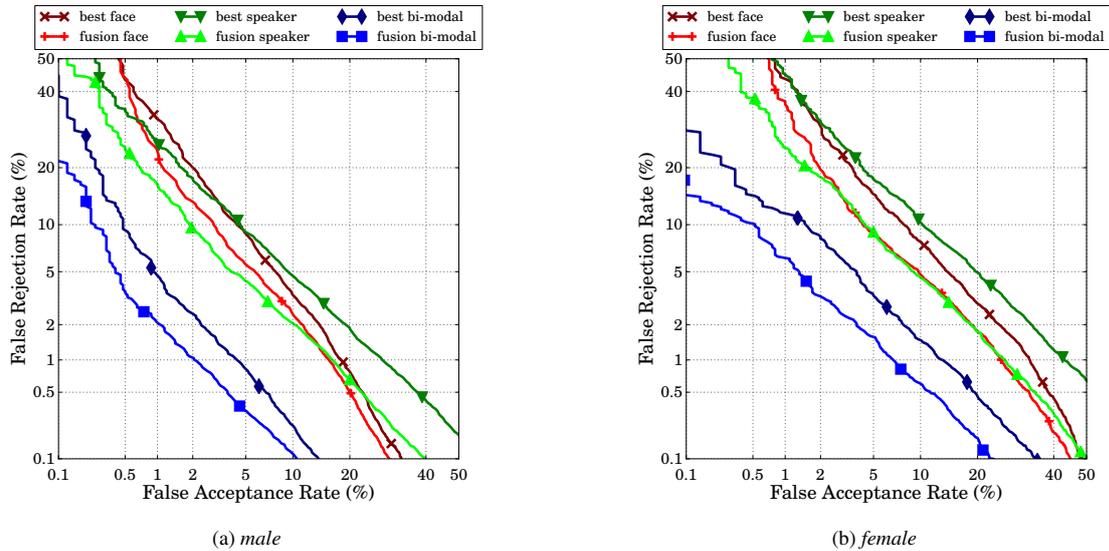


Figure 2: DET CURVES. This figure shows DET curves for uni-modal and bi-modal speaker and face recognition systems and their fusions on the evaluation set of the MOBIO database for the male and female protocols.

the evaluation set performance varies slightly.

Note that speaker and face recognition systems contribute similarly. For the *male* protocol, the top 10 systems contain 6 face and 4 speaker recognition systems, while for *female* the top 10 comprise speaker and face recognition systems in an equal number.

## 5. Conclusions

In the present paper, we tested how the decision level fusion of several state-of-the-art speaker and face recognition algorithms can help to improve recognition performance. We used the 12 speaker and 9 face recognition systems that were submitted to the speaker and face recognition evaluations in mobile environments [9, 10]. We showed that the uni-modal fusion of speaker or face recognition systems is able to improve performance moderately, i. e., by approximately 35 % or 15 % relative gain, respectively. Already fusing the best speaker and the best face recognition algorithm improved recognition performance by more than 55 % compared to the best uni-modal system. By integrating more speaker and face recognition systems into the fusion process, this gain can be increased up to 70 %. The final 1.78 % HTER for the *male* protocol and 3.80 % HTER for the *female* protocol outperform previously published bi-modal fusion algorithms that used the same database with the same evaluation protocols.

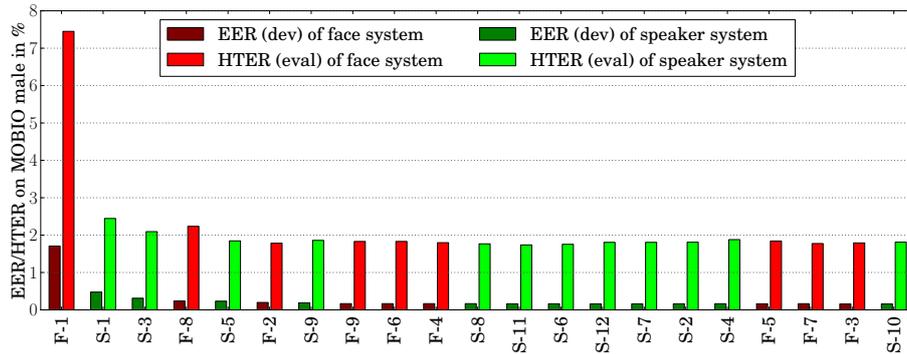
These findings lead to the conclusion that multi-modal multi-algorithm fusion should be the future trends in biometric recognition. It can also help forensic applications, especially since the resulting fused scores are already in terms of log-likelihood ratios.

## 6. Acknowledgment

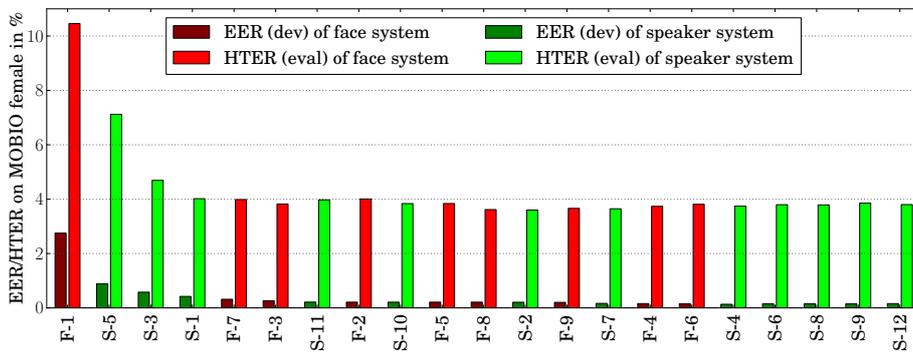
The research leading to the results presented in this paper has received funding from the European Community's Seventh Framework Program (FP7) under grant agreements 238803 (BBfor2) and from the Swiss National Science Foundation under the LOBI project.

## 7. References

- [1] V. Hautamäki, T. Kinnunen, M. Nosratighods, K.-A. Lee, B. Ma, and H. Li, "Approaching human listener accuracy with modern speaker verification," in *INTERSPEECH*, 2010, pp. 1473–1476.
- [2] A.J. O'Toole, P.J. Phillips, F. Jiang, J. Ayyad, N. Pearnard, and H. Abdi, "Face recognition algorithms surpass humans matching faces over changes in illumination," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 1642–1646, 2007.
- [3] J.P. Campbell Jr., "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.
- [4] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *Journal of the Optical Society of America A*, vol. 4, no. 3, 1987.
- [5] D.M. Blackburn, M. Bone, and P.J. Phillips, *Face Recognition Vendor Test 2000: Evaluation Report*, Storming Media, 2001.



(a) Pool of classifiers for male



(b) Pool of classifiers for female

Figure 3: OPTIMAL BI-MODAL FUSION. This figure shows the improvements in EER and HTER of optimally fusing speaker and face recognition systems. From left to right, the indicated system is added to the set of fused systems, starting with the best system F-1.

- [6] O. Ribaux, S.J. Walsh, and P. Margot, “The contribution of forensic science to crime analysis and investigation: Forensic intelligence,” *Forensic science international*, vol. 156, no. 2, pp. 171–181, 2006.
- [7] D. Ramos-Castro, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, “Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework,” in *Speaker and Language Recognition Workshop*. 2006, pp. 1–8, IEEE.
- [8] S. Marcel et al., “On the results of the first mobile biometry (MOBIO) face and speaker verification evaluation,” in *International Conference on Pattern Recognition*. 2010, pp. 210–225, Springer-Verlag.
- [9] E. Khoury et al., “The 2013 speaker recognition evaluation in mobile environment,” in *International Conference on Biometrics*, 2013.
- [10] M. Günther et al., “The 2013 face recognition evaluation in mobile environment,” in *International Conference on Biometrics*, 2013.
- [11] C. McCool et al., “Bi-modal person recognition on a mobile phone: Using mobile phone data,” in *IEEE International Conference on Multimedia and Expo, Workshop on Hot Topics in Mobile Multimedia*, 2012.
- [12] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.
- [13] W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff, “SVM based speaker verification using a GMM supervector kernel and NAP variability compensation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006.
- [14] A. Solomonoff, C. Quillen, and W.M. Campbell, “Channel compensation for SVM speaker recognition,” in *Proceedings of Odyssey, Speaker and Language Recognition Workshop*, 2004, pp. 57–62.
- [15] R. Vogt and S. Sridharan, “Explicit modelling of session variability for speaker verification,” *Computer Speech & Language*, vol. 22, no. 1, pp. 17–38, 2008.
- [16] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker



- verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 788–798, 2011.
- [17] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brümmer, “Discriminatively trained probabilistic linear discriminant analysis for speaker verification,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2011, pp. 4832–4835.
- [18] D. Garcia-Romero and C.Y. Espy-Wilson, “Analysis of i-vector length normalization in speaker recognition systems,” in *INTERSPEECH*, 2011, pp. 249–252.
- [19] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, “Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification,” in *INTERSPEECH*, 2009, pp. 1559–1562.
- [20] S.J.D. Prince and J.H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.
- [21] L. El Shafey, C. McCool, R. Wallace, and S. Marcel, “A scalable formulation of probabilistic linear discriminant analysis: Applied to face recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1788–1794, 2013.
- [22] X. Tan and B. Triggs, “Enhanced local texture feature sets for face recognition under difficult lighting conditions,” *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1635–1650, 2010.
- [23] G. Heusch, Y. Rodriguez, and S. Marcel, “Local binary patterns as an image preprocessing for face authentication,” in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2006, pp. 9–14.
- [24] N. Vu and A. Caplier, “Illumination-robust face recognition using retina modeling,” in *IEEE International Conference on Image Processing*, 2009, pp. 3289–3292.
- [25] N. Vu and A. Caplier, “Face recognition with patterns of oriented edge magnitudes,” in *European Conference on Computer Vision*. 2010, pp. 313–326, Springer.
- [26] L. Wiskott, J.-M. Fellous, N. Krüger, and C. v.d. Malsburg, “Face recognition by elastic bunch graph matching,” *IEEE Transactions on Pattern Recognition and Artificial Intelligence*, vol. 19, pp. 775–779, 1997.
- [27] M. Günther, D. Haufe, and R.P. Würtz, “Face recognition with disparity corrected Gabor phase differences,” in *Artificial Neural Networks and Machine Learning*, 2012, pp. 411–418.
- [28] T. Ahonen, A. Hadid, and M. Pietikainen, “Face recognition with local binary patterns,” in *European Conference on Computer Vision*, 2004, vol. 3021, pp. 469–481.
- [29] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkil, “Recognition of blurred faces using local phase quantization,” in *International Conference on Pattern Recognition*, 2008, pp. 1–4.
- [30] D.D. Cox and N. Pinto, “Beyond simple features: A large-scale feature search approach to unconstrained face recognition,” in *IEEE International Conference on Automatic Face Gesture Recognition*, 2011, pp. 8–15.
- [31] S. Pigeon, P. Druyts, and P. Verlinde, “Applying logistic regression to the fusion of the NIST’99 1-speaker submissions,” *Digital Signal Processing*, vol. 10, no. 1-3, pp. 237–248, 2000.
- [32] N. Brümmer et al., “Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006,” *IEEE Transactions on Speech, Audio and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.
- [33] C. McCool and S. Marcel, “Parts-based face verification using local frequency bands,” in *IEEE/IAPR Third International Conference on Advances in Biometrics*. 2009, pp. 259–268, Springer-Verlag.
- [34] P. Motlicek, L. El Shafey, R. Wallace, C. McCool, and S. Marcel, “Bi-modal authentication in mobile environments using session variability modelling,” in *International Conference on Pattern Recognition*, 2012, pp. 1100–1103.
- [35] T.P. Minka, “Algorithms for maximum-likelihood logistic regression,” Tech. Rep. 758, CMU Statistics Department, 2001.
- [36] A. Anjos, L. El Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel, “Bob: a free signal processing and machine learning toolbox for researchers,” in *ACM International Conference on Multimedia*, 2012, pp. 1449–1452.
- [37] A.F. Martin, G.R. Doddington, T. Kamm, M. Ordowski, and M.A. Przybocki, “The DET curve in assessment of detection task performance,” in *EUROSPEECH*, 1997.



# Bridging the gap between the forensic handwriting examiners and pattern recognition community

Linda Alewijnse

Netherlands Forensic Institute  
Department of Digital Technology and Biometrics  
The Hague, The Netherlands  
l.alewijnse@nfi.minvenj.nl

**There is a continuous need for new, unpublished data to train and evaluate new algorithms for signature verification systems. Handwriting samples that make up the current publicly available databases have all been collected under controlled conditions. Research databases constituted of case related forensic data in general are scarce. To suit forensic purposes, it is preferred to start building databases with forensically relevant data. When verification and identification systems are trained on this type of material, the output will be more suited for forensic examination purposes. The challenge is to bridge the gap between the forensic handwriting examiners and the pattern recognition community.**

Offline signature verification is a biometric technique with promising results for the near future to aid the forensic handwriting examiner in drawing a conclusion. The pattern recognition discipline has made rapid developments in the last ten years [1]. Implementing an analysis tool in the forensic practice is the next challenge. Before automated signature verifications can be used, the forensic community must be ascertained that the systems are trained, evaluated and validated on data that is collected under appropriate environmental conditions.

In the past years, from 2009 until 2013, several signature datasets were collected by researchers from the Netherlands Forensic Institute for the Signature Competition (SigComp) [2, 3]. This competition allows researchers and practitioners from academia and industries to compare performance on signature verification on new and unpublished datasets. Because all participating parties in the competition are provided with the same data, results are comparable. While the competition provides an overview of involved parties and shows the performance of the available systems to the forensic community, the pattern recognition researchers are more concerned about which features are most discriminative. The aim of the SigComp is to improve the interaction and bridge the gap between the two communities. Nevertheless, much work still needs to be done to bring together researchers in the field of automated handwriting analysis and signature verification and experts from the forensic handwriting examination community.

Data for signature verification research aimed at forensic implementation have specific requirements. Data must reflect the variation of handwriting in the relevant population, and should represent the whole intra-writer variation. Additional forensic requirements are: known sex, age, handedness, level of education, profession, cultural origin of the writer, a substantial amount of reference signatures, specifications of the conditions under which a simulated or disguised signature was produced, and the time span over which the data was collected.

Where biometric systems usually have access to high quality and uniform data, in forensic practice the trace under investigation is often of poor quality. This is not represented by the currently existing handwriting databases. The next step in bridging the gap between the pattern recognition community and forensic handwriting examiners should logically involve the use of real forensic samples. Simulated data can be used in the training phase of system development, because the ground truth of the origin is known. The evaluation phase should at least contain real life handwriting samples. However, the validation of the system should completely be performed with real data samples [4]. The best would be using forensic casework data to evaluate and validate automated systems, but legal aspects regarding privacy form an obstacle, as well as uncertainty about the ground truth of the writing.

The best solution would be acquiring existing handwriting samples in a similar way as a forensic handwriting examiner collects specimen writings in a case investigation. All writing conditions (intrinsic and extrinsic factors) are represented in the dataset and the ground truth of the sample is known.

- [1] M. Caligiuri and L. Mohammed, "The Neuroscience of Handwriting: Applications for Forensic Document Examination," CRC Press, 2012.
- [2] M.I. Malik, M. Liwicki, L. Alewijnse, and W. Ohyama, "ICDAR2013 Competitions on Signature Verification and Writer Identification for On- and Offline Skilled Forgeries (SigWiComp2013)," in press.
- [3] M. Liwicki et al., "Signature Verification Competition for Online and Offline Skilled Forgeries (SigComp2011)," in Document Analysis and Recognition (ICDAR), 2011 International Conference on, pp. 1480-1484, 2011.
- [4] D. Meuwly and R.N.J. Veldhuis, "Forensic biometrics: From two communities to one discipline," IEEE Conference publications BIOSIG 2012, Darmstadt Germany, pp. 1-12, Sep 2012.



# The distribution of calibrated likelihood-ratios in speaker recognition

David A. van Leeuwen<sup>1</sup> and Niko Brümmner<sup>2</sup>

<sup>1</sup>Netherlands Forensic Institute, The Hague and Radboud University Nijmegen, The Netherlands

<sup>2</sup>AGNITIO Research, Somerset West, South Africa

## Abstract

This paper studies properties of the score distributions of calibrated log-likelihood-ratios that are used in automatic speaker recognition. We derive the essential condition for calibration that the log likelihood ratio of the log-likelihood-ratio is the log-likelihood-ratio. We then investigate what the consequence of this condition is to the probability density functions (PDFs) of the log-likelihood-ratio score. We show that if the PDF of the non-target distribution is Gaussian, then the PDF of the target distribution must be Gaussian as well. The means and variances of these two PDFs are interrelated, and determined completely by the discrimination performance of the recognizer characterized by the equal error rate. These relations allow for a new way of computing the offset and scaling parameters for linear calibration, and we derive closed-form expressions for these and show that for modern i-vector systems with PLDA scoring this leads to good calibration, comparable to traditional logistic regression, over a wide range of system performance.

*This is a slightly elaborated version of a paper with the same title that has appeared at Interspeech 2013 [1].*

## 1. Introduction

In recent years, calibration in automatic speaker recognition has received more attention [2–12]. Intuitively, calibration is related to the ability to properly set a threshold in a speaker detection system so as to minimize the expected error [13]. In speaker detection, the task is to decide whether or not two speech signals originate from the same speaker. Because all speaker recognition systems internally work with some scalar *score* that expresses speaker similarity, a score threshold can control the trade-off between the two types of errors that a system can make [14, 15]. Indeed, in the series of NIST Speaker Recognition Evaluations (SRE) the primary evaluation measure has been sensitive to calibration. Until SRE 2010, calibration was assessed in a single operating point, through a single decision cost function known as  $C_{\text{det}}$ . Also other technologies in speech technology or biometrics utilize calibration-sensitive evaluation measures, such as the cost functions  $C_{\text{avg}}$  in language recognition [16] and the Half Total Error Rate in face recognition [17].

Since around 2004 [2, 3] the concept of calibration in speaker recognition has been generalized to a range of operating points by using proper scoring rules [18] to evaluate probabilistic statements about whether a trial is a same-speaker (target) or different-speaker (non-target) trial. A system that represents its score as a *likelihood-ratio* can be well-calibrated over a wide range of operating points simultaneously. This representation of the speaker recognition score has direct application in speaker detection, as the decision threshold follows directly from the cost function parameters [15], but also in evidence reporting in forensic speaker comparison cases [5, 19]. In the NIST SRE 2012, for the first time, hard decisions were no longer required, and instead the recognition score had to be submitted in the form of a likelihood-ratio. The evaluation measure effectively sampled the decision cost function at two different parameters [20, 21].

Since a calibrated likelihood-ratio is still just a score, all properties of normal scores apply to likelihood-ratios as well, and we can draw DET and ROC plots, determine EERs and inspect the score distributions. The axis warping of the DET plot [14] in combination with the observed more-or-less straight DET curves suggests that target and non-target score distributions could be accurately modelled with Gaussians. These score distributions and the relation to the DET have been studied previously [22, 23] and are very instructive to the understanding of basic detection theory and the concepts of calibration [15, 24]. In this paper we are interested in properties of the distributions of *calibrated log-likelihood-ratios*. This may help situations where we carry out a calibration transformation on raw recognition scores, because it can tell us what the calibrated distributions should look like.

The paper is organized as follows. We define the very nature of a calibrated likelihood-ratio in Section 2. In Section 3 we investigate the properties of log-likelihood-ratio distributions when they are Gaussian, and we will then apply these in Section 4 as a new method for calibration. We then present experiments and conclusions.

## 2. Likelihood-ratio idempotence

Here we carefully define the *likelihood-ratio* (LR) and show that it has the interesting property: *the LR of the LR is the LR*, which forms a definition of calibration.



The speaker recognition system has as input two speech segments, denoted  $X$  and  $Y$ , which it processes in two steps. We represent the first step as  $s = f(X, Y)$ . To keep things general,  $s$  may represent different kinds of output, e.g., a pair of acoustic feature vector sequences, a pair of i-vectors, or just a single, scalar recognition score. The second step is to compute the likelihood-ratio  $r$  as a function of  $s$ , as:

$$r = \frac{P(s | H_1, \mathcal{M})}{P(s | H_2, \mathcal{M})} \quad (1)$$

where  $H_1$  is the (target) hypothesis that  $X$  and  $Y$  originate from the same speaker,  $H_2$  the (non-target) hypothesis that they are from two different speakers, and  $\mathcal{M}$  is a generative probabilistic model for  $s$ . In current practice,  $s$  is always the recognition score, so that  $\mathcal{M}$  merely models scalar scores—not i-vectors, acoustic feature sequences or speech signals. But our theory below is sufficiently general to remain applicable in future to more ambitious models, when  $s$  might have a more complex form. We now assume there is given the *hypothesis prior*,  $\pi = P(H_1)$ , which allows us to express the *hypothesis posterior*, via Bayes' rule as:

$$P(H_1 | s, \mathcal{M}, \pi) = \frac{\pi r}{\pi r + (1 - \pi)} \quad (2)$$

This shows that  $r$  is a *sufficient statistic*: the posterior depends on  $s$  only through  $r$ . This allows rewriting the posterior as:

$$P(h | s, \mathcal{M}, \pi) = P(h | r, \mathcal{M}', \pi), \quad h \in \{H_1, H_2\} \quad (3)$$

where we have introduced  $\mathcal{M}'$  to denote  $\mathcal{M}$ , augmented by asserting (1). Although  $r$  contains all the relevant information that  $\mathcal{M}$  can extract from  $s$  to recognize the unknown hypothesis, it must be stressed that  $r$  and  $s$  do *not* necessarily contain all the relevant information that could have been extracted from the original input  $X, Y$  by some more elaborate model. Now we use the *odds form* of Bayes' rule:

$$\frac{P(H_1 | \rho, M, \pi)}{P(H_2 | \rho, M, \pi)} = \frac{\pi}{1 - \pi} \frac{P(\rho | H_1, M)}{P(\rho | H_2, M)} \quad (4)$$

where  $\rho$  is a placeholder for  $r$  or  $s$  and  $M$  for  $\mathcal{M}$  or  $\mathcal{M}'$ . Combining this with (3), we find the desired relationship (the LR of the LR is the LR [25]):

$$r = \frac{P(s | H_1, \mathcal{M})}{P(s | H_2, \mathcal{M})} = \frac{P(r | H_1, \mathcal{M}')}{P(r | H_2, \mathcal{M}')} \quad (5)$$

If we define  $x$  to be the *log-likelihood-ratio* (LLR):

$$x = \log r \quad (6)$$

we also find<sup>1</sup> (the LLR of the LLR is the LLR):

$$x = \log \frac{P(x | H_1, \mathcal{M}'')}{P(x | H_2, \mathcal{M}'')} \quad (7)$$

where  $\mathcal{M}''$  augments  $\mathcal{M}'$  by addition of (6).

## 2.1. Implications

Rewriting (5) as:

$$P(r | H_1, \mathcal{M}') = r P(r | H_2, \mathcal{M}') \quad (8)$$

we see that if either of the two distributions is given, then the other distribution is completely determined—they cannot vary independently. Moreover, a further restriction is placed on these distributions: since the LHS must integrate to 1, the *expected value* of the non-target distribution (the integral of the RHS) must be:  $\langle r \rangle = 1$ . Similarly, for targets:  $\langle \frac{1}{r} \rangle = 1$ . By applying Jensen's inequality [26] we also find for targets:  $\langle x \rangle \geq 0$  and for non-targets:  $\langle x \rangle \leq 0$ .

## 2.2. Good and bad calibration

How does (5) function as a definition of calibration? Since it is an equality, won't all LRs calculated via (1) by some model  $\mathcal{M}$ , just automatically satisfy (5)? Yes they will, but only if  $\mathcal{M}$  and  $\mathcal{M}'$  are related as explained above. If we want to independently judge the goodness of the calibration of  $r$ , we do not condition the distributions for  $r$  on the recognizer's model  $\mathcal{M}$ . Instead, we could empirically observe the target and non-target values of  $r$  as calculated by the recognizer over an independent, supervised database of speaker detection trials. Letting  $\mathcal{O}$  denote the empirical observation, we could then say the model  $\mathcal{M}$  is well calibrated if:

$$r = \frac{P(s | H_1, \mathcal{M})}{P(s | H_2, \mathcal{M})} \approx \frac{P(r | H_1, \mathcal{O})}{P(r | H_2, \mathcal{O})} \quad (9)$$

Bad calibration is when the LRs given respectively by the recognizer's  $\mathcal{M}$  and empirical observation  $\mathcal{O}$ , do not agree in this way. This can and does happen, since  $\mathcal{O}$  is independent of any development data that was used to determine the form and parameters of  $\mathcal{M}$ .

It should be noted that (9) does not give a practical recipe to judge degree of goodness of calibration—it specifies neither how to assign  $P(r | h, \mathcal{O})$ , nor how to numerically evaluate the agreement between LHS and RHS. For practical solutions for calibration-sensitive objective functions, see for example [27].

## 3. Gaussian distributed log-likelihood-ratios

Inspired by the fact that DET curves in speaker recognition tend to be straight [22], we explore a Gaussian solution to the LLR distribution constraint (7). Since target

<sup>1</sup>To see this, note the log transformation is monotonic and the Jacobian of the transformation cancels in the ratio.



and non-target LLR distributions are so tightly coupled, it turns out that if the one is assumed to be Gaussian, then the other must also be. We shall use the shorthand:  $e(x) = P(x | H_1, \mathcal{M}'')$  and  $d(x) = P(x | H_2, \mathcal{M}'')$ . Arbitrarily assuming a Gaussian distribution for non-targets (*different-speaker trials*):

$$d(x) = \mathcal{N}(x | \mu_d, \sigma_d) = \frac{1}{\sqrt{2\pi}\sigma_d} e^{-(x-\mu_d)^2/2\sigma_d^2}. \quad (10)$$

We derive the functional form for targets<sup>2</sup>,  $e(x)$ , when (7) applies:

$$e(x) = e^x d(x) = \frac{1}{\sqrt{2\pi}\sigma_d} e^{x-(x-\mu_d)^2/2\sigma_d^2}. \quad (11)$$

We collect the terms in  $x$  in the exponent, which itself can be written like

$$-\frac{x^2 - 2\mu_d x + \mu_d^2}{2\sigma_d^2} + \frac{2\sigma_d^2 x}{2\sigma_d^2} \quad (12)$$

$$= -\frac{x^2 - 2(\mu_d + \sigma_d^2)x + \mu_d^2}{2\sigma_d^2} \quad (13)$$

$$= -\frac{(x - (\mu_d + \sigma_d^2))^2}{2\sigma_d^2} + \frac{2\mu_d\sigma_d^2 + \sigma_d^4}{2\sigma_d^2} \quad (14)$$

The first term is in the familiar form of a Gaussian exponent, the second will result in a constant factor. Gathering terms, and writing

$$\mu_e = \mu_d + \sigma_d^2, \quad (15)$$

the expression for the same-speaker comparison log-likelihood-ratio scores becomes

$$e(x) = \frac{1}{\sqrt{2\pi}\sigma_d} e^{\sigma_d^2/2 + \mu_d} e^{-(x-\mu_e)^2/2\sigma_d^2} \quad (16)$$

$$= e^{\sigma_d^2/2 + \mu_d} \mathcal{N}(x | \mu_e, \sigma_d). \quad (17)$$

We see that  $e(x)$  is of Gaussian shape, with

$$\sigma_e = \sigma_d \equiv \sigma. \quad (18)$$

Since  $e(x)$  must be a proper PDF, its integral over  $x$  must be unity, from which follows that

$$e^{\sigma^2/2 + \mu_d} \int_{-\infty}^{\infty} \mathcal{N}(x | \mu_e, \sigma) dx = 1 \quad (19)$$

$$-2\mu_d = \sigma^2. \quad (20)$$

Finally, with (15) we find

$$\mu_e = \mu_d + \sigma^2 = -\mu_d \equiv \mu, \quad (21)$$

This shows that  $d(x)$  and  $e(x)$  are equal variance Gaussians with means symmetric around zero at  $\pm\mu$ , and where the variance and mean are related (20)

$$\sigma^2 = 2\mu. \quad (22)$$

<sup>2</sup>trials where the speakers are equal

### 3.1. Equal Error Rate and $d'$

Using the symmetry of the solution, it is clear that the threshold for the equal error rate is at  $x = 0$ . Using the expression for the miss probability, the equal error rate  $E_=$  is

$$E_= = \int_{-\infty}^0 \mathcal{N}(x | \mu, \sigma) dx \quad (23)$$

$$= \int_{-\infty}^{-\mu/\sigma} \mathcal{N}(x | 0, 1) dx \equiv \Phi(-\mu/\sigma), \quad (24)$$

where  $\Phi(x)$  is the cumulative normal distribution.

It is sometimes useful to recognize the parameter  $d'$  from detection theory, which is the difference in means expressed in terms of the standard deviation, here  $d' = 2\mu/\sigma$ . With (24) the relation becomes

$$E_= = \Phi\left(-\frac{1}{2}d'\right). \quad (25)$$

$$d' = \sigma = -2\Phi^{-1}(E_=), \quad (26)$$

introducing  $\Phi^{-1}(y)$ , the inverse of the cumulative normal distribution. The importance of the relations above is that  $\mu$  and  $\sigma$  are determined by the discrimination performance measured by  $E_=$ , using (22) and (26)

$$\mu = \frac{\sigma^2}{2} = 2[\Phi^{-1}(E_=)]^2. \quad (27)$$

## 4. A new calibration method

In practice, automatic speaker recognition systems do not deliver scores that can directly be interpreted as a log-likelihood-ratio, even though they are computed as such, for instance in the good old UBM-GMM scoring [28] or the latest i-vector PLDA scoring [29]. A practical solution to this is to convert raw scores  $s(X, Y)$  to calibrated log-likelihood-ratios by some transformation function  $x(s)$ , usually constrained to be monotonic increasing. There are many ways of doing this. The FoCal [30] and BOSARIS [31] toolkits use logistic regression to discriminatively train linear calibration transformations. Other possibilities include isotonic regression (PAV [31]) and line-up calibration [10] that uses the rank in a line-up of foil speakers. In FoCal or BOSARIS, the score-to-LLR function is affine:

$$x(s) = as + b \quad (28)$$

and the parameters  $a$  and  $b$  are found by optimizing cross-entropy, a calibration-sensitive objective function defined on a supervised set of speaker recognition trials.

Here we contrast the popular discriminative logistic regression solution to a new generative, constrained maximum-likelihood (ML) solution. Our constraints follow from assuming (i) Gaussian LLR distributions, and (ii) an affine score-to-LLR transform (28). This implies



that (i) the LLR distributions are constrained as derived in Section 3, and (ii) the score distributions are also Gaussians, with equal variances. With no LLR distribution constraints, we would have had 6 free parameters: 2 means, 2 variances and 2 calibration parameters. But we have imposed 3 constraints, equal variances (18), symmetric means (21) and (22). We find the remaining 3 free parameters by maximizing the following weighted likelihood:

$$\frac{\alpha}{N_e} \sum_{i \in \mathcal{E}} \log \mathcal{N}(s_i | m_e, v) + \frac{1 - \alpha}{N_d} \sum_{i \in \mathcal{D}} \log \mathcal{N}(s_i | m_d, v) \quad (29)$$

where  $\mathcal{E}$  indexes  $N_e$  target scores,  $\mathcal{D}$  indexes  $N_d$  non-target scores, and where we have generalized the usual maximum-likelihood (ML) criterion by relative weighting of targets by  $0 \leq \alpha \leq 1$  and non-targets by  $1 - \alpha$ . This weighting helps to compensate for the restrictive equal-variance modeling assumption, by allowing the ML criterion to focus on an operating point with a target proportion of  $\alpha$ . This mechanism will be demonstrated in Fig. 2 below. The score distribution parameters that need to be optimized are the means  $m_e, m_d$  and common variance  $v$ . Setting derivatives to 0, we find the maximum likelihood at the sample means:

$$m_e = \frac{1}{N_e} \sum_{i \in \mathcal{E}} s_i, \quad m_d = \frac{1}{N_d} \sum_{i \in \mathcal{D}} s_i \quad (30)$$

and at a weighted combination of sample variances:

$$v = \frac{\alpha}{N_e} \sum_{i \in \mathcal{E}} (s_i - m_e)^2 + \frac{1 - \alpha}{N_d} \sum_{i \in \mathcal{D}} (s_i - m_d)^2 \quad (31)$$

By (28), the LLR distribution parameters become  $\sigma^2 = a^2 v$ ,  $\mu_e = a m_e + b$  and  $\mu_d = a m_d + b$ . Finally, applying the constraints  $\sigma^2 = \mu_e - \mu_d$  and  $\mu_e = -\mu_d$ , we can solve for the calibration parameters:

$$a = \frac{m_e - m_d}{v}, \quad b = -a \frac{m_e + m_d}{2} \quad (32)$$

We call this recipe *constrained, maximum-likelihood, Gaussian* (CMLG) calibration. An advantage of CMLG is that it has a closed form, in contrast to the iterative optimization required by logistic regression.

#### 4.1. Experiment

In order to test CMLG we apply it to a number of recognition trials sets. We use a set of trials crafted for duration-dependence experiments [9] from the NIST SRE 2008 and 2010 trial sets, the telephone-telephone “extended” trial lists. We constructed short duration segments of 5, 10, 20, and 40 seconds from both train and test segments by simply selecting the first frames after speech activity detection. All durations, including the full conversation

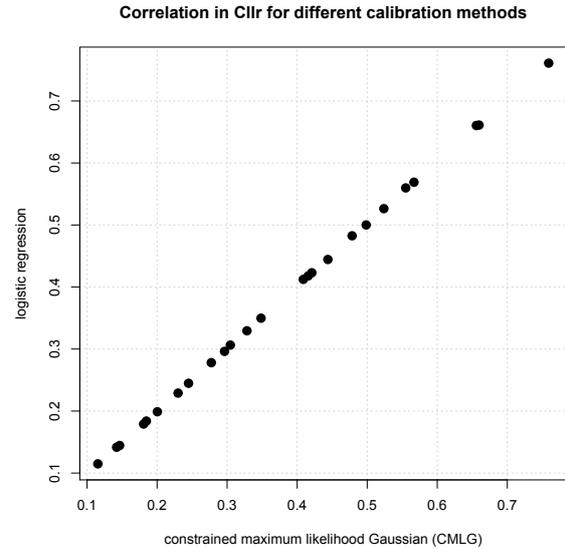


Figure 1:  $C_{llr}$  values of the 25 trial lists for the CMLG method (horizontal) versus logistic regression (vertical).

side, were tested in all combinations, leading to 25 different trial lists. The NIST SRE10 ‘det-5’ performance over these lists ranges from  $E_{\perp} = 2.9\text{--}26\%$ . The recognition system is a standard i-vector based system with PLDA scoring described elsewhere [21].

We contrast CMLG (with  $\alpha = \frac{1}{2}$ ) to the traditional logistic regression method. The calibrations are trained on NIST SRE 2008 data (427 375 trials) and applied to SRE 2010 trials for evaluation (10 007 900 trials), all gender mixed. We evaluate the 25 different trial list combinations using  $C_{llr}$ , a cost function that is sensitive to calibration over the whole DET curve [3]. We used R’s `glm` routine for logistic regression.

The results are shown in Fig. 1, where we have plotted the  $C_{llr}$  obtained using CMLG calibration versus  $C_{llr}$  obtained using logistic regression. The values are highly correlated. For CMLG, the average  $C_{llr}$  over all 25 conditions is 0.375, for logistic regression it is 0.376. These can be called good, as the mean  $C_{llr}^{\min}$  is 0.370.

We also tried CMLG on *different* data, with scores generated by a *different* speaker recognition system. The scores are from an i-vector PLDA speaker recognizer submitted by the ABC team [32] to the NIST SRE 2012 evaluation [20]. Calibration was trained on about 120 million scores, obtained by processing multiple microphone and telephone speech segments of 2019 male and female speakers from the SRE’04, ’05, ’06, ’08 and ’10 Mixer databases. Performance was evaluated on about 80 million scores, obtained by pooling all five “common evaluation conditions” of SRE 2012 [20].

Figure 2 compares the performance of the proposed CMLG (blue triangles) against the traditional prior-weighted logistic regression (red circles) on the SRE 2012 scores. On the horizontal axis we show the target proportion, which was used to weight the training criterion, in

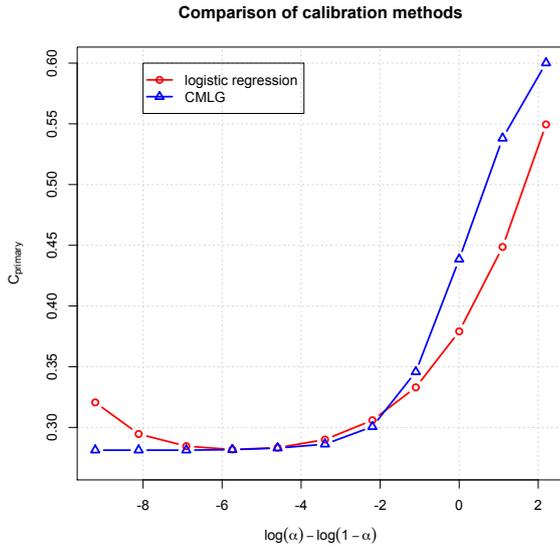


Figure 2:  $C_{\text{primary}}$  for logistic regression and CMLG calibration methods for ABC’s SRE12 submission, as a function of prior  $\alpha$  used in the objective / ML optimization.

*log odds* form:  $\log \alpha - \log(1-\alpha)$ . On the vertical axis, we show  $C_{\text{primary}}$ , the calibration-sensitive SRE 2012 evaluation criterion [20], where smaller values indicate better performance. In the case of logistic regression,  $\alpha$  was used in a similar manner to relatively weight targets and non-targets in the discriminative logistic regression training criterion, see [27] for details. Observe that CMLG and logistic regression have similar minima, but CMLG has a wider, flatter minimum. CMLG is also much faster to train, because it has a closed-form solution, while logistic regression needs iterative numerical optimization.

## 5. Discussion and Conclusions

We have shown in this paper, that if the different-speaker calibrated log-likelihood-ratio scores from a speaker recognition system follow a Gaussian distribution, then the distribution of the same-speaker scores must also be Gaussian after calibration, with the same variance but opposite mean. Because monotonically increasing score-to-likelihood-ratio functions do not change the DET plot, such equal-variance distributions in the calibrated score domain imply  $45^\circ$  DET-plots in the raw score domain as well—which is neither observed with real data<sup>3</sup> nor desired for applications operating in the low false alarm region. The logical conclusion then is that real scores, if they are well-calibrated, will not be Gaussian. However, we see that our PLDA system can be calibrated quite well under the Gaussian assumptions, and indeed we have noticed that i-vector PLDA systems tend to have score distributions that appear more Gaussian than earlier tech-

<sup>3</sup>We have measured the slope of the DET in the conventional error region 0.1–50% for the data in the experiment. The mean slope over the 25 conditions is  $-0.99$  with a standard deviation of 0.06, so in fact this data appears to honour the equal variance condition quite well.

nologies, such as i-vector LDA cosine distance scoring, support vector machines or the UBM-GMM likelihood ratio scoring.

The Gaussian solution to the LLR equation (7) is one where both distributions are shaped by the same mathematical function. In signal detection theory, where the distribution represents noise, this seems almost mandatory, but in speaker recognition this is not an obvious assumption. We have experimented with other distributions, e.g., in the likelihood-ratio domain (5) a pair of Gamma distributions is a solution to the calibration condition, and these are asymmetric in the log-likelihood-ratio domain. However, such distributions seem to be not at all representative of real score distributions. Also, an arbitrary linear combination of Gaussians with different means and corresponding variances is a solution to (7) which allows some freedom in fitting a shape of score distribution. In principle, there is no need for real score distributions to follow any mathematical description, but we have observed that many researchers like to use some form of idealized shape of the score distributions to understand the data [5, 22]. When calibration methods are designed, condition (7) should therefore be taken into account.

The relations derived in Section 3 open up more possibilities for relations between the various evaluation measures. For instance, we can compute  $C_{\text{llr}}$  by numerical integration as

$$\frac{1}{\log 2} \int_{-\infty}^{\infty} \mathcal{N}(x | \mu, \sigma) \log(1 + e^{-x}) dx \quad (33)$$

and this relates  $C_{\text{llr}}$  to  $E_{=}$  via (26) and (27) for Gaussian score distributions. E.g., for our set of 25 trial lists this expression differs from  $C_{\text{llr}}^{\text{min}}$  only 0.006 in root mean squared difference, or about 2%. Instead of for calibration, the relations can also be used for fusion of systems. For pre-calibrated systems this leads to solutions that transparently depend on the correlation between the scores.

The fact that we can obtain the linear calibration parameters under the Gaussian assumption is an interesting side-effect of this study. The calibration parameters can be expressed in closed-form, and do not explicitly consider cross entropy or  $C_{\text{llr}}$  as an optimization objective. For score distributions that do not resemble a Gaussian, this calibration method is likely to fail—we therefore do not recommend CMLG calibration as a general technique. Still, we are quite pleased that the experiments support the mostly theoretical results of this paper.

## 6. Acknowledgments

The research leading to these results has received funding from the European Community’s Seventh Frame work Programme (FP7/2007–2013) under grant agreement no. 238803.



## 7. References

- [1] D. A. van Leeuwen and N. Brümmer, “The distribution of calibrated likelihood-ratios in speaker recognition,” in *Proc. Interspeech*. ISCA, 2013, pp. 1619–1623.
- [2] N. Brümmer, “Application-independent evaluation of speaker detection,” in *Proc. Odyssey 2004 Speaker and Language recognition workshop*. ISCA, June 2004, pp. 33–40.
- [3] N. Brümmer and J. du Preez, “Application-independent evaluation of speaker detection,” *Computer Speech and Language*, vol. 20, pp. 230–275, 2006.
- [4] N. Brümmer and D. A. van Leeuwen, “On calibration of language recognition scores,” in *Proc. Odyssey 2006 Speaker and Language recognition workshop*, San Juan, June 2006.
- [5] D. Ramos-Castro, J. González-Rodríguez, and J. Ortega-García, “Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework,” in *Proc. Odyssey 2006 Speaker and Language Recognition Workshop*, 2006.
- [6] D. Ramos, “Forensic evaluation of the evidence using automatic speaker recognition systems,” Ph.D. dissertation, Universidad Autonoma de Madrid, November 2007.
- [7] N. Brümmer, L. Burget, J. Černocký, O. Glembek, F. Grezl, M. Karafiát, P. Matějka, D. A. van Leeuwen, P. Schwarz, and A. Strassheim, “Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006,” *IEEE Transactions on Speech, Audio and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.
- [8] Z. Jancik, O. Plchot, N. Brummer, L. Burget, O. Glembek, V. Hubeika, M. Karafiát, P. Matejka, T. Mikolov, A. Strassheim *et al.*, “Data selection and calibration issues in automatic language recognition—investigation with BUT-AGNITIO NIST LRE 2009 system,” in *Proc. Speaker and Language Odyssey*, 2010.
- [9] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, “Evaluation of i-vector speaker recognition systems for forensic application,” in *Proc. Interspeech*. Firenze: ISCA, August 2011.
- [10] D. A. van Leeuwen and N. Brümmer, “A speaker line-up for the likelihood ratio,” in *Proc. Interspeech*. Firenze: ISCA, August 2011.
- [11] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, “The effect of noise on modern automatic speaker recognition systems,” in *Proc. ICASSP*. Kyoto: IEEE, March 2012.
- [12] G. R. Doddington, “The role of score calibration in speaker recognition,” in *Proc. Interspeech*, 2012.
- [13] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, “The NIST speaker recognition evaluation—Overview, methodology, systems, results, perspective,” *Speech Communication*, vol. 31, pp. 225–254, 2000.
- [14] 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.
- [15] D. A. van Leeuwen and N. Brümmer, “An introduction to application-independent evaluation of speaker recognition systems,” in *Speaker Classification*, ser. Lecture Notes in Computer Science / Artificial Intelligence, C. Müller, Ed. Springer, 2007, vol. 4343.
- [16] A. F. Martin and A. N. Le, “NIST 2007 language recognition evaluation,” in *Proc. Speaker and Language Odyssey*. Stellenbosch, South Afrika: IEEE, 2008.
- [17] R. Wallace, M. McLaren, C. McCool, and S. Marcel, “Cross-pollination of normalization techniques from speaker to face authentication using gaussian mixture models,” *Information Forensics and Security, IEEE Transactions on*, vol. 7, no. 2, pp. 553–562, 2012.
- [18] M. DeGroot and S. Fienberg, “The comparison and evaluation of forecasters,” *The Statistician*, pp. 12–22, 1983.
- [19] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, and J. Ortega-Garcia, “Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2104–2115, September 2007.
- [20] C. S. Greenberg, “The NIST year 2012 speaker recognition evaluation plan,” 2012. [Online]. Available: [http://www.nist.gov/itl/iad/mig/upload/NIST\\_SRE12\\_evalplan-v17-r1.pdf](http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf)
- [21] D. A. van Leeuwen and R. Saeidi, “Knowing the non-target speakers: the effect of the i-vector population for PLDA training in speaker recognition,” in *Proc ICASSP*. Vancouver: IEEE, 2013, pp. 6778–6782.
- [22] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, “Score normalization for text-independent speaker verification systems,” *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.
- [23] J. Navrátil and G. N. Ramsawamy, “The awe and mystery of t-norm,” in *Proc. Eurospeech*, 2003, pp. 2009–2012.
- [24] N. Brümmer, “Measuring, refining and calibrating speaker and language information extracted from speech,” Ph.D. dissertation, Stellenbosch University, 2010.
- [25] K. Slooten and R. Meester, “Forensic identification: Database likelihood ratios and familial DNA searching,” *arXiv:1201.4261 [stat.AP]*, 2012.
- [26] J. L. W. V. Jensen, “Sur les fonctions convexes et les inégalités entre les valeurs moyennes,” *Acta Mathematica*, vol. 30, no. 1, pp. 175–193, 1906.
- [27] N. Brümmer and G. Doddington, “Likelihood-ratio calibration using prior-weighted proper scoring rules,” in *Proc. Interspeech*. ISCA, 2013.
- [28] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.
- [29] S. J. D. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2007, pp. 1–8.
- [30] N. Brümmer, *FoCal-II: Toolkit for calibration of multi-class recognition scores*, August 2006, software available at <http://www.dsp.sun.ac.za/~nbrummer/focal/index.htm>.
- [31] E. de Villiers and N. Brümmer, *The Bosaris Toolkit*, BOSARIS, 2010, software available at <https://sites.google.com/site/bosaristoolkit/>.
- [32] AGNITIO, BUT, and CRIM, “ABC SRE’12 presentation,” in *NIST SRE 2012 Workshop*, Orlando, 2012.



# Estimated Intra-Speaker Variability Boundaries in Forensic Speaker Recognition Casework

*Yosef A. Solewicz<sup>1</sup>, Gaëlle Jardine<sup>2</sup>, Timo Becker<sup>3</sup> and Stefan Gfroerer<sup>3</sup>*

<sup>1</sup>National Police, Israel

<sup>2</sup>Police Technique et Scientifique, France

<sup>3</sup>Federal Criminal Police Office, Germany

solewicz@police.gov.il, gaelle.jardine@interieur.gouv.fr, timo.becker@bka.bund.de,  
stefan.gfroerer@bka.bund.de

## Abstract

Current automatic forensic speaker recognition algorithms are known to be highly dependent on matched reference data for producing statistically reliable outcomes. In day-to-day forensic casework, this data is often not available. This paper addresses this problem by introducing a data-driven modular method which does not require precise reference data. Results are more general and given as p-values and likelihood ratio ranges. Experiments performed with a real forensic database support our proposal.

## 1. Introduction

Likelihood ratios (LRs) have become increasingly popular as a way of reporting the outcome in forensic speaker comparison casework [1-5]. They are seen as bringing speaker identification in line with the wider forensic community, where LRs have been used for some time now.

In speaker comparison, LRs reflect the relation between intra- and inter-speaker variabilities. It is relatively easy to model inter-speaker variability for any of the combinations of mismatched conditions one regularly comes across in forensic casework. Each combination can be established from a pool of monolingual speakers that are recorded through one channel. Plenty of data is therefore available for modeling inter-speaker variability in a large variety of settings (involving same or different channels, languages etc.).

By contrast, the estimation of intra-speaker variability is problematic. Strictly speaking, the suspect's own intra-speaker variability should be calculated using several recordings in settings that reflect those of the suspect and the offender recordings. The forensic circumstances obviously render this generally impossible.

A common workaround is to compute average parameters using a pool of reference speakers that each are recorded in different, case-specific settings. Since the settings may differ greatly from one case to another, any reference corpus, in order to be useful across many cases, would have to comprise for each speaker a great number of recordings involving different channels and possibly different languages. In practice, most forensic labs find this an impossible task.

Current speaker recognition systems usually respond to this lack of data by extrapolating from intra-speaker variability data that is not strictly comparable to the specific case. This is a fix that many forensic labs rightly refuse to use; for a large portion of casework they therefore prefer to opt for a phonetic analysis.

The driving idea behind our proposed method is that the expert should be given the means to do speaker comparisons even in conditions where relevant intra-speaker variability statistics are not available. We believe that this is possible, albeit at the expense of reporting with a lesser degree of accuracy. We continue the work in [6], where it was suggested that it is preferable, in cases where intra-speaker variability would have to be based on non-representative estimates, to concentrate on inter-speaker variability and to take only limited account of intra-speaker variability. In practice, this means either going back some way towards a pure "frequentist" approach, where the estimated inter-speaker distribution returns a p-value (reflecting the probability of obtaining identification scores higher than the ones obtained for the specific case) and the intra-speaker variability is ignored, or estimating a range of likelihood ratios based on past performance of the algorithm under specific, known conditions.

The paper is organized as follows. Section 2 describes the proposed methodology. Section 3 describes the experimental setup and section 4 the actual experiments performed, followed by analyses of the results in Section 5. Section 6 concludes the paper and discusses future plans.

## 2. Methodology

We start off by using an estimated inter-speaker variability distribution on the basis of a reference speaker population (see [6] for a brief discussion on using either the offender or suspect recording for this purpose). In its basic form, our method adopts a typical frequentist approach, simply returning p-values based on the estimated inter-speaker distribution, below which we reject the null-hypothesis of a random match for our suspect recording.

This basic module can be expanded to return an outcome in the form of a LR if we estimate boundaries for intra-speaker variability statistics. To obtain these statistics, we run our algorithm off-line under two "extreme" conditions — extremely matched and extremely mismatched — and measure system performance in both conditions in terms of the distances between the respective intra- and inter-speaker variability distributions. These two conditions must be sufficiently different, and in particular the mismatched condition sufficiently mismatched or "extreme", to be able to cover the expected range of scenarios in any given casework. We expect that using a variety of microphones in the extreme mismatched condition should essentially cover most of the forensic intra-speaker variability in terms of channel mismatch. Other sources of noise such as health condition



(e.g., a sore throat or a cold), language and type of expression (e.g., shouting) are also known to have an impact on system performance. They must be taken account of separately by the forensic expert.

In operating mode, we create two virtual intra-speaker distributions – matched and mismatched – by shifting the mean of the computed inter-speaker distribution by the two pre-calculated distribution distances. Two virtual LRs are then estimated. They consist of a common denominator (the recognition score set against the real, case-specific inter-speaker variability distribution) and a different numerator each (the recognition score set against each of the two virtual intra-speaker variability distributions). These two LRs constitute an upper and a lower boundary that are specific to our condition. For any given case, it is the forensic expert's role to appreciate whether the condition specific to the case lies within the range of scenarios covered by these two extreme conditions.

An example of such a setup is given in Figure 1, where the matching and mismatching conditions refer to the channel (telephone vs. microphone). It shows the distances between intra- and inter-speaker distribution means for our pre-determined boundaries. The entire process is summarized in Figure 2.

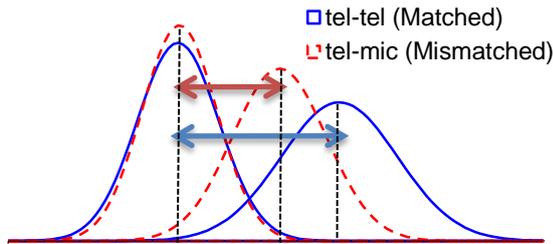


Figure 1. Target and impostor scores for matched and mismatched conditions

**In training mode:**

Calculate parametrized normal intra- and inter-distributions using reference benchmarks for the extreme conditions 1 and 2:

$$X_{intra,1,2} \sim N(\mu_{intra,1,2}, \sigma_{intra,1,2}^2)$$

$$X_{inter,1,2} \sim N(\mu_{inter,1,2}, \sigma_{inter,1,2}^2)$$

**In operating mode:**

Calculate the parametrized normal inter-speaker distribution for the questioned recording:

$$X_{inter,q} \sim N(\mu_{inter,q}, \sigma_{inter,q}^2)$$

Obtain the parametrized virtual intra-speaker distributions for the questioned recording in the extreme conditions:

$$X_{intra,q,1,2} \sim N(\mu_{intra,q,1,2}, \sigma_{intra,q,1,2}^2), \text{ where}$$

$$\mu_{intra,q,1,2} = \mu_{inter,q} + (\mu_{intra,1,2} - \mu_{inter,1,2}),$$

$$\sigma_{intra,q,1,2}^2 = \sigma_{intra,1,2}^2$$

Figure 2. Variability estimation scheme

### 3. Experimental setup

Our virtual LR methodology was validated using a real forensic benchmark and a regular i-vector speaker verification system as described below.

#### 3.1 Verification system

The verification system used in the current experiments is based on standard i-vector features extracted through maximum a posteriori adaptation followed by linear discriminant analysis (LDA) as a second processing layer and cosine scoring [7] without any further score normalization. Data from NIST 2004, 2006 and 2008 evaluations [8] had been used for system development.

#### 3.2 Corpus

The corpus used in these experiments is the ‘‘GFS1.0 Corpus (German Forensic Speech Version 1.0)’’ (referred to as GFS from now on). It is a compilation of original forensic telephone taps from the German Federal Criminal Police Office and contains spontaneous German speech of male individuals. The corpus is part of the EU project Correlation between phonetic–acoustic–auditory and automatic approaches in forensic speaker identification that aimed at making authentic forensic data available to external research institutes. The current version consists of two protocol variants, one of which is more oriented towards analyses by automatic systems. This is the one used in the present evaluations. It is organized as follows.

- 39 offender recordings with a minimum duration of 30 seconds, originating from 24 speakers
- 21 suspect recordings with a minimum duration of 60 seconds, originating from 21 speakers
- 49 reference population recordings with a minimum duration of 60 seconds, originating from 49 speakers

#### 3.3 Extreme intra-speaker distribution estimation

Scores for parametrizing the two extreme (matched and mismatched) distributions were obtained using a subset of the Nist 2008 Speaker Recognition Evaluation with a few hundred, mostly English-speaking, male-speaker conversations. Telephone–telephone trials were defined as the one extreme (matched), telephone–microphone trials as the other extreme (mismatched). In forensic casework, of course, mismatched conditions can go far beyond the simple difference between two channels, and can include combinations of unknown recording devices, different audio encoding, different languages and so on. Recall also that GFS is a German-language corpus (possibly containing some unknown channels), whereas our algorithm was trained on mostly English-language telephone recordings.

In order to simulate more realistic forensic conditions, original training conversations in the Nist evaluation were truncated to one minute length and testing conversations to 30 seconds. Equal Error Rates of 5.1% and 11.5% were obtained for the matched and mismatched conditions.

Finally, target and impostor scores for both conditions (Figure 1) were parametrized as normal distributions and their means and standard deviations used for deriving the normal virtual intra-speaker distributions, as explained in Figure 2.



## 4. Experiments

The GFS corpus is limited in the amount of possible trials; some caution must therefore be used in the interpretation of the following results. Comparing each of the 39 offender recordings in the corpus with each of the 21 suspect recordings produces 36 target trials and 783 impostor trials. We obtained an EER of 10.7% for this evaluation. This value, as expected, lies between the 5.1% and 11.5% EERs obtained for the extreme matched and mismatched NIST conditions. It suggests that the GFS trials are fairly mismatched.

### 4.1 p-values

The basic variant of casework reporting discussed in this paper is a simple p-value reporting, below which we reject the null hypothesis of a random match for a suspect recording. This hypothesis is modeled by matching the suspect (or offender) recording against the relevant reference population. It is tested by comparing it to the matching score obtained by the offender-suspect comparison. To be on the (suspect's) safe side, one can also estimate two distributions and test two null-hypotheses. This is done by matching the reference population against both the suspect and the offender recordings. If the suspect-offender comparative match lies squarely within either distribution, the null hypothesis cannot be rejected. The further in the positive distribution tail (if not to the right of the entire distribution) the comparative match lies, the more doubts are cast on the null hypothesis, but strictly speaking, nothing can be said about the alternative hypothesis. Using the p-value as a recognition score (actually,  $1-(p\text{-value})$ , so that target scores are higher than impostors'), we obtained an EER of 11.1% in the GFS evaluation.

### 4.2 Virtual LRs

However, we can define our extreme conditions and obtain estimated intra-variability distributions by modeling extreme alternative hypotheses, thereby improving on the frequentist approach. This leads us to the more elaborate proposal in this paper of a virtual LR reporting. We will focus on this and skip the results analysis of the pure p-value experiments.

In these experiments we initially used the raw target and impostor scores obtained for the GFS evaluation as described in the beginning of this section to parametrize normal inter- and intra-speaker distributions. These distributions were then used to convert the same raw evaluation scores into LRs. This is an over-optimistic, a-posteriori calibration and will be used for comparison purposes only.

We then estimated LR boundaries for each trial in the GFS benchmark as follows. The population of 49 reference speakers was matched against each offender recording to compute a normal inter-speaker variability distribution. We then derived the corresponding virtual matched and mismatched intra-speaker distributions as described in Figure 2. Finally, virtual matched and mismatched LRs were calculated for the specific trial. Since the intra-variability distributions used are virtual and not based on the GFS trials themselves, the LRs are not considered to be a-posteriori calibrated, and could be called a-priori LRs.

At this point we produced an additional a-priori baseline through a different approach. Instead of deriving extreme distributions for each trial as before and setting upper and lower LR boundaries, we evaluated “normalized” matched and mismatched extreme distributions once and kept them fixed

for all GFS trials, as follows. We re-evaluated the matched and mismatched Nist evaluations using t-normed scores obtained through the 49 reference speakers from GFS. In this way, we came up with extreme inter- and intra- speaker distributions normalized to the specific GFS benchmark, our specific forensic scenario. (The matched and mismatched evaluations attained EERs of 6.2% and 14.4%, respectively. Note that these performances are inferior to those obtained without score normalization (Section 3.3), probably due to the mismatch between the evaluation (Nist) and reference population (GFS) data.)

Similarly, we re-evaluated the GFS trials applying to its scores the same t-norm procedure used for obtaining the normalized extreme distributions. The t-normed GFS scores were then converted into LRs using the t-normed extreme distributions.

The additional a-priori baseline approach described above conceptually differs from the proposed framework. Within the variability estimation approach, we only transfer the estimated system performances in extreme conditions to our specific casework. For the baseline approach, we normalize the extreme condition distributions to the specific casework conditions. In other words, the proposed framework decouples the original system development setup from the casework in terms of data, whereas the baseline (as an example of other current forensic workarounds) confounds development and casework data in order to match the system performance in these different setups.

## 5. Results

In this section we will present two results for our proposed LR framework and the baseline approach as described in Section 4.2. The first one uses an a-priori setup which reflects a typical forensic scenario: just as one normally cannot obtain case-specific intra-variability distributions, we report a specific outcome without relying on other GFS comparisons to model intra-speaker distributions. The second one uses an a-posteriori setup which simulates a utopian situation: for a specific piece of casework we do have compatible intra-speaker comparisons and are able to express our forensic comparison in terms of a customized LR. In this case we do use intra-variability distributions obtained with other GFS comparisons.

From a forensic perspective it is interesting to compare different setups in terms of their corresponding log-likelihood-ratio costs (Cllr) [9, 10]. In addition, it is convenient to differentiate between target and impostor costs as we proposed in [11]. The costs for individual target and non-target trials are given by:

$$C_{llr}^{trial}(S) = \begin{cases} \log_2 \left( 1 + \frac{1}{S} \right), & \text{target} \\ \log_2 (1 + S), & \text{non-target} \end{cases} \quad (1)$$

where  $S$  is the trial's likelihood ratio. Our goal is to find which of the a-priori setups best matches the optimum a-posteriori baseline in terms of the trial costs. Figure 3 depicts the estimated and baseline average costs for target and impostor trials.

For each configuration, we show the matched and mismatched extreme LR boundaries (left-hand, blue and right-hand, green, columns). The middle, red column is the



average a-posteriori cost. Ideally we would like to have our a-priori boundaries tightened around the a-posteriori values. This would indicate that our LR range is properly estimated. The graph shows that in these experiments the variability estimation approach provided a better LR generalization than the baseline. In addition, note that the baseline approach shows a poorer calibration (especially in the mismatched condition) than the proposed framework. As pointed out earlier, this is probably an effect of score normalization performed with mismatched data, which does not occur within the variability estimation approach. However, in terms of performance both approaches produced 11.1% EER using either matched or mismatched intra-speaker modelling, which is very close to the a- posteriori performance (10.7%).

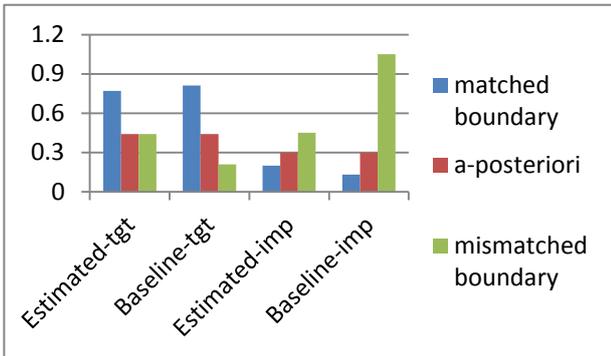


Figure 3. Estimated and baseline average  $C_{IIr}$

In these experiments we are also interested in measuring how close our estimated log likelihood ratio (LLR) ranges fit the actual a-posteriori LLRs. We therefore compute the average LLR range for all evaluation trials. In addition, we penalize target and impostor trials whose a-posteriori LLRs are lower resp. higher than their assigned ranges. For these inaccurate trials we average the distances from their a-posteriori LLRs to their closest LLR boundary. These values are given in Table 1. Note that the variability estimation approach shows relatively narrow ranges and low LLR bias compared to the baseline.

	Target Trials	Impostor Trials
Estimated range	0.6	0.4
Baseline range	1.3	0.1
Estimated bias	0.9	0.7
Baseline bias	2.5	1.1

Table 1. Average LLR ranges and bias

### 6. Discussion

We propose a data-driven framework for reporting forensic speaker comparison outcomes in terms of virtual LR ranges. The LRs are obtained by setting an inter-speaker distribution that is empirically estimated and based on a case-specific reference population against virtual intra-speaker distributions. These are derived from estimated but system-specific, pre-calculated distances between intra- and inter-speaker distributions that are assessed under controlled conditions which can be compared to our specific casework conditions. Comparability must be ascertained for each case

by the forensic expert. We argue that a large number of forensic cases that typically do not meet the data matching requisites for a LR report can be reasonably addressed by the methodology proposed.

Our framework avoids explicit normalization of scores that uses mismatched data. It was assessed with a real forensic benchmark and obtained satisfactory results compared to a similar approach which attempts to normalize reference distributions to the forensic scenario. We recommend further evaluations of this methodology using other forensic benchmarks and systems and focusing especially on the issue of calibration, which is a particularly sensitive topic in forensic situations with scarce matched data [4].

### 7. References

- [1] Champod, C., & Meuwly, D. (2000). The inference of identity in forensic speaker recognition. *Speech Communication*, 31(2), 193–203.
- [2] Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M., & Ortega-Garcia, J. (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech & Language*, 20(2), 331–355.
- [3] Rose, P. (2006). Technical forensic speaker recognition: Evaluation, types and testing of evidence. *Computer Speech & Language*, 20(2), 159–191.
- [4] Campbell, J. P., Shen, W., Campbell, W. M., Schwartz, R., Bonastre, J. F., & Matrouf, D. (2009). Forensic speaker recognition. *IEEE Signal Processing Magazine*, 26(2), 95–103.
- [5] Morrison, G. S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, 51(3), 91–98.
- [6] Solewicz, Y., Koppel, M. & S. Sofer (2004). A robust framework for forensic speaker verification. *SPECOM*, 393–397.
- [7] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4), 788–798.
- [8] Martin, A. & Przybocki, M. The NIST speaker recognition evaluation series, National Institute of Standards and Technology’s Web site [Online]. Available: <<http://www.nist.gov/speech/tests/sre>>.
- [9] Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2), 230–275.
- [10] van Leeuwen, D. A., & Brümmer, N. (2007). An introduction to application-independent evaluation of speaker recognition systems. In *Speaker Classification I* (pp. 330–353). Springer Berlin Heidelberg.
- [11] Solewicz, Y., Becker, T., Jardine, G. & Gfroerer, S. (2012). Comparison of Speaker Recognition Systems on a Real Forensic Benchmark, *Odyssey*.



## Influence of the datasets size on the stability of the LR in the lower region of the within source distribution

Rudolf Haraksim, Didier Meuwly

Netherlands Forensic Institute  
Laan van Ypenburg 6, 2497GB, The Hague, The Netherlands  
r.haraksim(d.meuwly)@nfi.minvenj.nl

### Abstract

This article focuses on the statistical evaluation of the fingerprint evidence using the likelihood ratio (LR) approach. It studies the influence of the quantity of data used to model the within (WS) and between (BS) source variability. The LR system built for the experiment uses an Automated Fingerprint Identification System (AFIS) feature extraction and comparison algorithm, fingerprint and fingerprint datasets coupled with a generative approach for modeling the WS and BS variability. This article concentrates on the computation of LRs of the same source in the lower region of the WS distribution. It analyzes the behavior of the LR with an increasing number of entries in the WS datasets while maintaining the constant proportion of the BS dataset in an attempt to estimate the amount of same source scores necessary to achieve consistent LR performance.

### 1. Introduction

While the question of the comparison of complete fingerprints seems to be an issue long solved in the biometric world with many commercial algorithms and applications available, quite some issues arise when analyzing forensic fingerprints (traces). When a fingerprint and a fingerprint are subjected to forensic evaluation, the fingerprint is almost always partial, its quality severely degraded due to uncontrolled imposition (clarity, distortion) and due to the effects of the development methods.

While the AFIS matching and comparison algorithm is able to achieve great results in terms of performance and speed while producing shortlists of candidates, it is not used in the current practice for the statistical evaluation of fingerprints and fingerprint evidence. Forensic evidence ( $E$ ) in this case is considered the similarity score resulting from the fingerprint and fingerprint comparison. In order to quantify the weight of the forensic evidence we start off with a set of mutually exclusive propositions, the one of the prosecution  $H_p$  and the one of the defense  $H_d$ :

- $H_p$  – the fingerprint originates from the individual that is also the source of the fingerprint
- $H_d$  – the fingerprint originates from an unknown individual, randomly selected

With the propositions defined we can now proceed to the LR calculation which can be derived from the odds form of the Bayes theorem in the following way:

$$LR = \frac{\Pr(E | H_p)}{\Pr(E | H_d)} \quad (1)$$

where  $\Pr$  indicates the probability of observing the evidence  $E$  given one of the two hypotheses.

The calculation of the LR implies the modeling of the WS and BS scores distributions using a discriminative, generative or hybrid approach [1]). The main objective of this article is to study the influence of the size of the datasets on the stability of the LR. The influence will be studied using a generative approach<sup>1</sup> for the modeling of the within and between source variability.

An ideal situation would be to dispose of a quantity of score observations large enough to cover the whole range of the BS and WS distributions. However in the tails of these distributions a good estimate of the LR is difficult to obtain, due to the rarity of the scores. In the regions where the number of scores is sufficient to describe reliably the WS and BS the LR value is generally low, and the stability of the LR can be considered as an indicator for the robustness<sup>2</sup> and of the reliability<sup>3</sup> of the method.

In this work we shall analyze the region of the lower tail of the WS score distribution - see figure 1 (similar issues addressed in [6]). We are interested in this region mainly due to the fact that similarity scores in this particular area can “shift” the scales in favor of either of the propositions. Ideally we would like to observe a stable LR support to either of the propositions, however with the varying number of the WS scores we observe variation in the LRs as well.

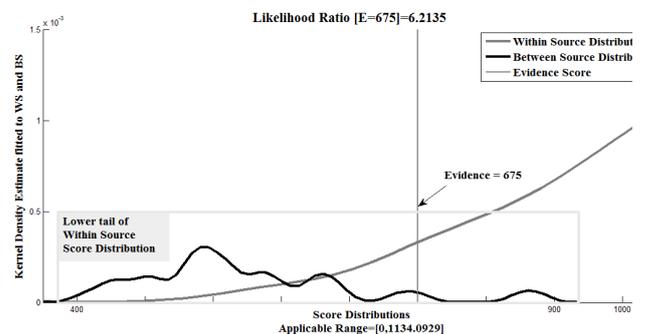


Figure 1. – Area of interest (lower tail of the WS score distribution)

<sup>1</sup> In the generative approach we “generate” the score distributions from the discrete datasets (similarity scores).

<sup>2</sup> Robustness is defined as the ability of a method to maintain the tendency of its performance when reducing the quality conditions of the data under examination

<sup>3</sup> Reliability defined as the capability of the method of not degrading the trueness of the LR when used in all the possible cases for which it has been designed



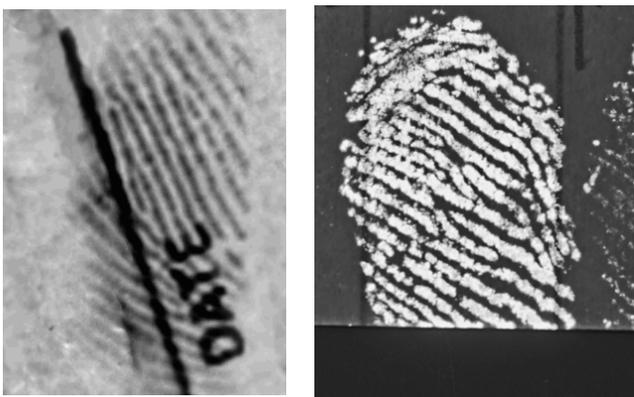
In this initial study we will model the similarity scores produced by the AFIS algorithm using the Kernel Density Function (KDF). This choice is based on the fact that we are dealing with discrete datasets and because the comparison algorithm produces multimodal score distributions. Since we are interested in observing the influence of the different sizes of datasets on the LR stability, the over-fitting, which in most of the cases is considered a drawback of the KDF seems to be a desirable side-effect for this particular application.

Before any method developed can be used in a forensic casework, a validation step needs to provide insight about its robustness and reliability ( $LR > 1$  if  $H_p$  true,  $LR < 1$  if  $H_d$  is true). The aim of this article is to study the stability of the LR produced and in particular the variations due to data on the probability estimates for both the numerator and denominator of the LR. We will show the influence of lowering the quantity of data used for modeling the WS and BS scores on the stability of the LRs. Despite the fact that relatively small number of individuals is used in this study, it provides a valuable insight on the LR stability depending on the decreasing number of WS scores.

## 2. Datasets used

For modeling the BS scores, large quantities of reference fingerprints are available, for example ten-print cards originating from a police fingerprint databases. It is not necessarily the case for WS scores, where a limited number of fingerprints and corresponding fingerprints with the ground truth known is available. Different approaches have been proposed in the literature to handle the data sparsity under  $H_p$  [3, 4].

Both methods rely on the use of simulated fingerprints from the suspected individual. In [4] these simulated fingerprints are compared with a set of corresponding fingerprints (multiple fingerprints per finger), when in [3] large quantities of simulated fingerprints are compared with a single fingerprint in order to obtain the WS score distribution.



**Figure 2.** – Simulated fingerprint on the left vs. visualized real fingerprint from a crime-scene on the right

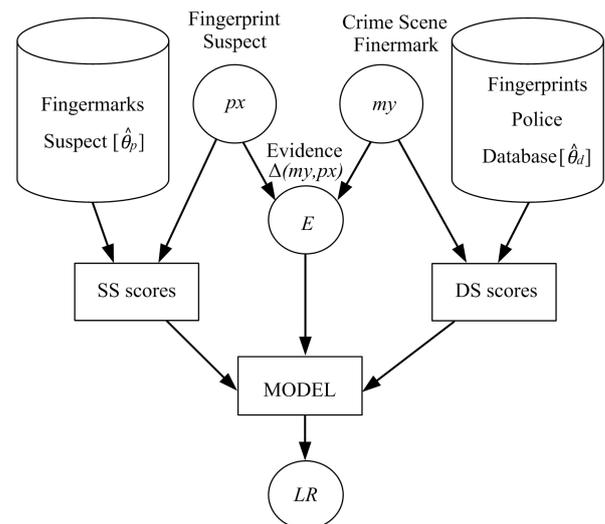
This method mimics the distortion and provides enough reference material for modeling the WS score distribution.

The fingerprints produced by this method are not completely equivalent to real crime-scene fingerprints but for the purpose of this article and based on the results published in [3], their similarity is considered as sufficient (see figure 2). The number of minutiae and the effect of distortion, present in the set of fingerprints used, represent the key elements of variability for the calculation of the evidential value.

Simulated fingerprints with 8 minutiae configurations were chosen for this article, as a majority of the fingerprints recovered as pieces of evidence contains less than 12 minutiae, which is the numerical standard in most countries using a numerical standard. In these countries fingerprints with less than 12 minutiae are currently not considered as evidence that can be presented at court and would primarily benefit from the approach described in this paper.

### 2.1 LR model and size of the dataset used

Figure 3 illustrates the LR model used in this article. The nomenclature used to describe the different datasets refers to the one used in [2].



**Figure 3.** – The LR model

The fingerprint police database consists of electronic copy of ten-print cards. For the purpose of this article we have selected a population of 20.000 individuals (200.000 fingerprints) to represent the BS population.

Since we aim to establish the stability of the LRs in the lower region of the WS score distribution, we will use data from four individuals, for which we have large quantity of simulated fingerprints available – ranging from 2.179 to 8.455. In practice, collecting a WS dataset counting 1000s' of fingerprints for a suspected individual is a time consuming procedure which largely depends on the willingness of the suspect to cooperate (in many cases impossible).

In the following section a forensic evaluation will be described together with the calculation of a likelihood ratio.



### 3. Evidence Evaluation

As indicated in figure 3, we proceed with evidence evaluation in multiple stages:

- Establish the value of the evidence (E) – a similarity score between a fingerprint or fingerprint
- Model the WS distribution based on the comparison of the marks and prints of the same individual (same finger)
- Model the BS distribution based on the comparison of the marks and prints of the different individual (different fingers)
- Calculate the Likelihood Ratio

According to [5] the LR is calculated in the following way:

$$LR = \frac{\Pr(E | H_p, \Delta_{SS}(m, p))}{\Pr(E | H_d, \Delta_{DS}(m, p))} \quad (2)$$

where:

$\Delta_{SS}(m, p)$  is the similarity score of the marks and print of the same source

$\Delta_{DS}(m, p)$  is the similarity score of the marks and prints of the different source

In order to obtain calculate the evidence same source in the same dataset, one of the simulated fingerprints (on a leave-one-out basis) will play the role of the crime scene mark and will be compared to the reference print of the same individual. If the total number of the simulated marks per individual is  $n$ , a total of  $n-1$  fingerprints will be available to form the WS score distribution.

As indicated earlier, for WS and BS score distribution modeling we will use the KDF function.

For measuring the stability of the LRs we will vary the number of the WS and BS scores using random sub-sampling. Ideally, with increasing number of the WS scores we should observe more stable LR. More data is in general more informative, especially in the tails of the WS and BS distributions.

In the following section we shall study the influence of the size of the WS and BS datasets on the stability of the LR.

### 4. Method used

Since we aim to examine the lower tail of the WS score distribution, we will focus on the similarity score interval 375 – 900 (shown in figure 1). The similarity scores are dimensionless, which advocates for the use of the LR framework. Simulated fingerprints of 4 individuals are used in this study.

**Table 1** – Proportion of simulated fingerprints

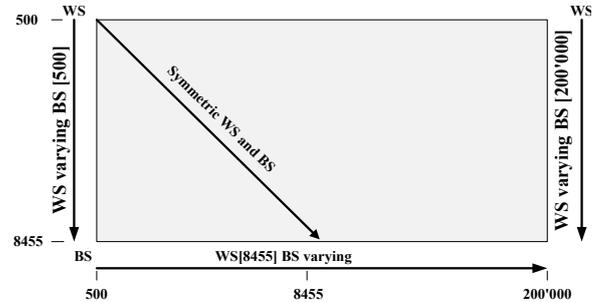
	No. of fingerprints
Individual 1	8455
Individual 2	4666
Individual 3	3179
Individual 4	3758

Individual 1 is used as a benchmark (largest number of simulated fingerprints available) to study the influence of the varying size of the simulated marks and police database datasets. We defined 5 experimental conditions:

1. Equal proportion of WS and BS scores (Symmetric)
2. WS[8455] and BS varying (WSmax)
3. WS[500] and BS varying (BSmin)
4. WS varying and BS[500] (BSmin)
5. WS varying and BS[200'000] (BSmax)

These conditions (where available) will be applied to all 4 individuals.

For all scenarios, the smallest number of WS scores tested counts 500 with 500 scores increments until the WSmax (where available). Similarly the smallest number of BS scores configuration counts 500 with 500 scores increments until BSmax. Since we have a lot more scores available for the BS, we will examine the influence of the amount of BS scores on the stability of the LR with 20,000, 50,000, 100,000 and 200,000 scores.



**Figure 4** – Four scenarios for LR stability analysis

Please recall that we selected the similarity score interval range from 375 to 900 (see Figure 1). Based on the initial assumption that the LRs in this region are of low order of magnitude, we will place the LRs into 4 bins ( $10^{-2} < LR < 10^{-1}$ ;  $10^{-1} < LR < 10^0$ ;  $10^0 < LR < 10^1$ ;  $LR > 10^1$ ) in order to analyze the LR behavior. We are particularly interested in observing the varying proportions of the LRs crossing the value of the neutral evidence ( $LR_E = 1$ ), changing the support of  $H_p$  to  $H_d$  and the actual value of the LRs (observation of the E at a fixed value with changing the experimental conditions). The influence of the varying sizes of the WS and BS datasets on the stability of the LR is presented in the following chapter.

### 5. Results

The experimental setup with most similarity scores (BSmax WSmax) was taken as ideal condition, which we aim to approach with increasing number of the similarity scores. In this sense, we want to get as close to the “best estimate” with the minimum number of scores. Reader should also keep in mind that our aim here is to understand the data rather than draw conclusions of the rather erratic behavior of the LRs produced.

Results are divided into two sections: firstly we will look at the stability of the LR for the individual 1 (counting the most WS scores), while in the second part we will attempt to replicate the results for the remaining individuals.



The sum of all the LRs in the 4 LR ranges is equal (126 – given by the total number of  $E$  scores for which the LRs have been calculated).

### 5.1 LR stability analysis

In figure 5 one of the populations (BS or WS) is fixed while other one varies from 500 to 8000 (however LRs have been analyzed on the whole range of BS 500 - 200000).

ranges.

Calculated LR values for each piece of evidence  $E$  under different experimental conditions are presented in figure 6 on the log-scale. For the experimental condition 1 (symmetric WS and BS) [1000] 85% of LRs support  $H_p$ , on contrary in the symmetric set WS and BS [4000] only 46% supports  $H_p$  (horizontal line in figure 6 indicates LR = 1 and demonstrates the LR shift in support of different hypothesis).

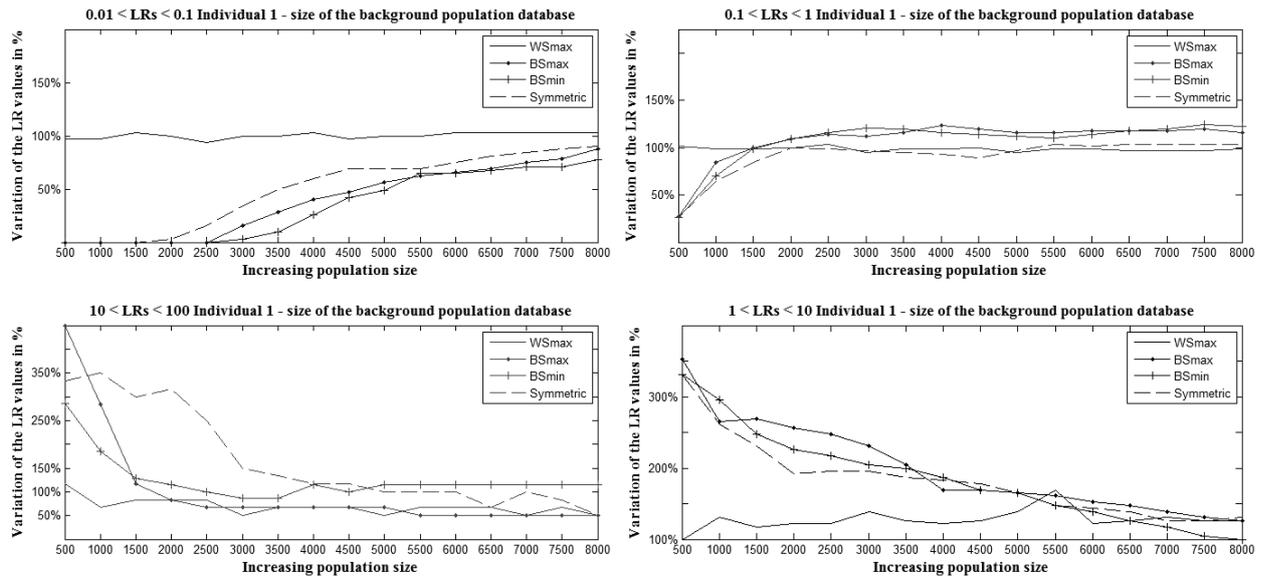


Figure 5 – Experimental setups results for individual 1

The stability of the LRs can be observed and compared with varying size of the BS population (BSmin, BSmax...) The experimental results for the individual 1 show that about 4000 scores (WS) are needed to obtain a stabile behavior of  $\pm 10\%$  of the LR values, for the selected LR bin

The size of the BS population does not have a significant influence on the overall stability of the LR. The symmetric experimental condition converges the fastest to the best estimate; therefore this condition will be replicated for the remaining individuals.

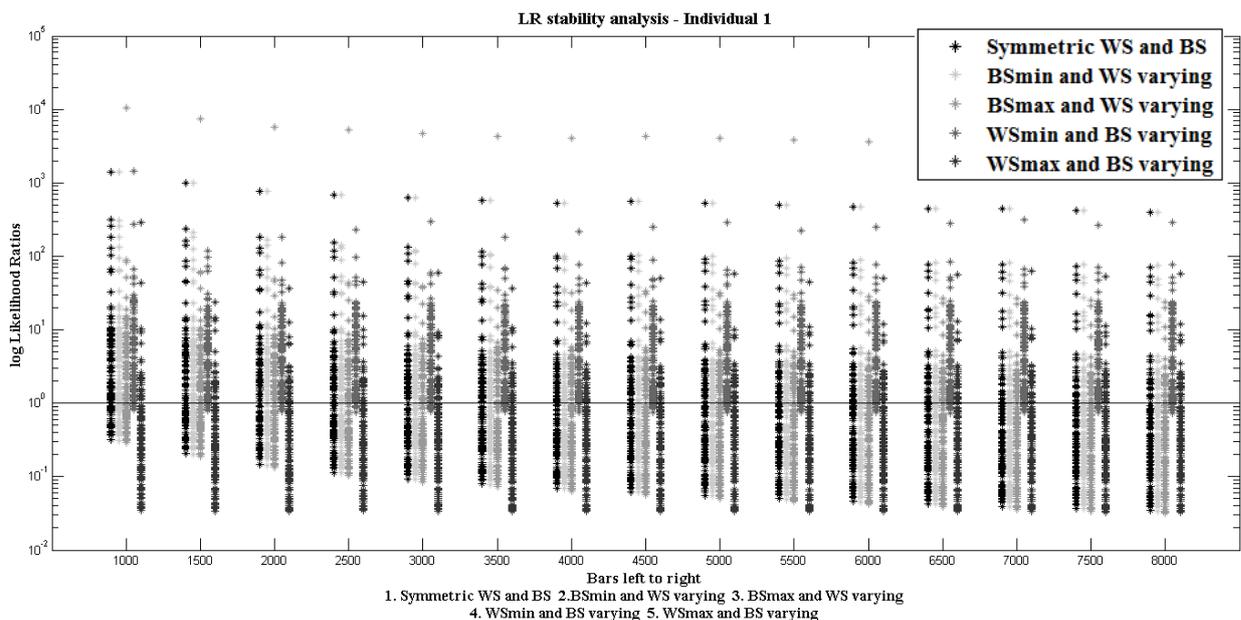


Figure 6 – log(LRs) presented with varying BS population



## 5.2 Replication for the remaining individuals

The stability of the LRs is analyzed using the experimental condition 1 (symmetric WS and BS). Figure 7 illustrates the experimental results for the individuals 2, 3 and 4.

The best estimate was calculated from the LRs in the configuration (BSmax and WSmax) of each individual. No LR lower than  $10^{-1}$  was recorded for individuals 2–4; hence this bin will remain empty.

The results observed advocates for using the LR calculation method as described in [5]. Despite of the different size of the within source dataset for the 4 individuals, the stabilizing effect of increasing the size of the datasets on the LRs (as observed in the benchmark) was replicated with amongst all four individuals. Analyzing the results separately, within source scores dataset counting 4500 seems sufficient to reach stability of  $\pm 10\%$  of the LR values for individual 2, 3000 for individual 3 and population size of 2000 for individual 4. More general conclusions cannot be drawn from such a limited number of individuals.

## 6. Discussion and conclusions

The aim of this article was to study the influence of the size of datasets on the stability of the LR. Judging from the experiments conducted, the increase in the between source population size does not seem to have much influence on the LR stability. The symmetric experimental setup has shown to produce the most stable LRs, while a significant variability was observed between the WSmin and WSmax experiments (see figure 6).

The stabilizing trend of the LR due to the increasing size of within source population was replicated for all four individuals, however the results show differences in the minimum number of the within source scores necessary to obtain a stable LR amongst the different individuals and call for further tests with datasets of comparable sizes before a generic threshold can be set.

The use of simulated fingerprints in the experiments show that they are a valuable evaluation tool, as they are relatively easy to produce in significant quantities and one can be “beyond any doubt” certain regarding their origin.

## 7. Future work

This article is intended as a preliminary study on the stability of the LRs and shows how the LRs behave with varying population sizes. The future work will focus on obtaining similarly large datasets of simulated fingerprints to individual 1 and extend the study for the  $E$  different source. Following research will be dedicated to non-parametric methods and model-based approaches.

## Acknowledgements

The research has been conducted on behalf of the BBfor2 project of the Marie Curie Actions FP 7 using the NFI data resources.

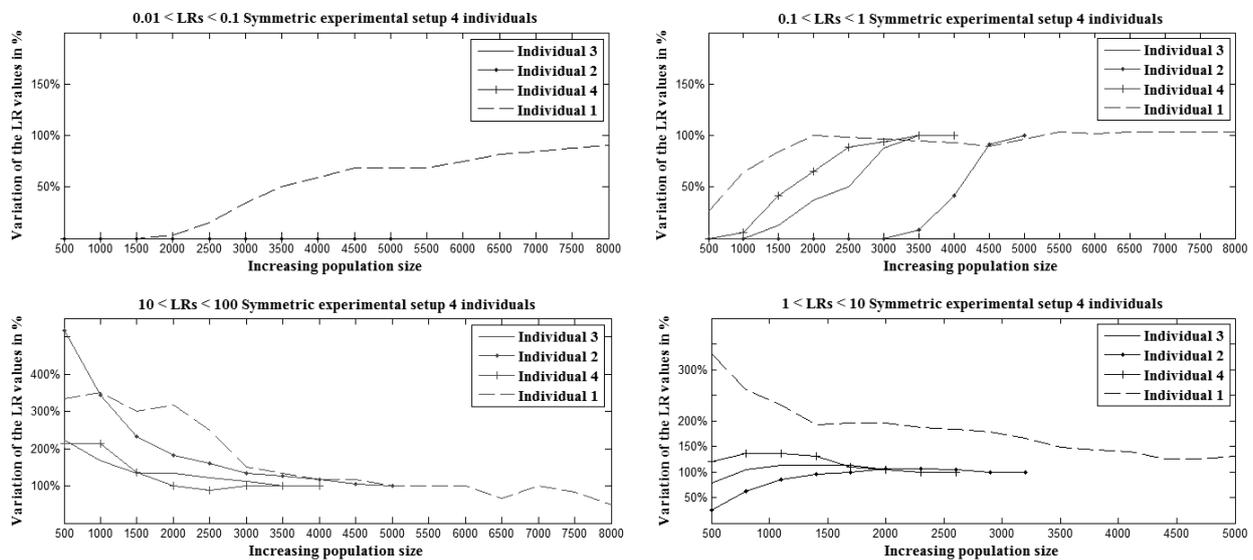


Figure 7 – Differences in stability of the LR amongst 4 individuals using symmetric experimental condition.

## References

- [1] D. Ramos, Forensic Evaluation of the Evidence Using Automatic Speaker Recognition Systems, Universidad Autonoma de Madrid, November 2007
- [2] C. Neumann, Quantifying the weight of evidence from a forensic comparison: a new paradigm, *RSS* 175(2), (2011) pp 1–26
- [3] C. M. Rodriguez, A. de Jongh, D. Meuwly, Introducing a semi-automated method to simulate large numbers of forensic fingerprints for research on fingerprint identification, *JFS* 57(2), (2012) pp. 334–342

- [4] N. Egli, C. Champod, P. Margot, Evidence evaluation in fingerprint comparison and automated fingerprint identification systems – Modelling within finger variability, *FSI* 167 (2007) 189–195

- [5] C. Neumann et al., Computation of Likelihood Ratios in fingerprint identification for configurations of three minutiae, *JFS* 51(6), (2006) 1255–1266.

- [6] T. Ali, L.J. Spreeuwiers, R.N.J. Veldhuis, A review of calibration methods for biometric systems in forensic applications, In: 33<sup>rd</sup> WIC Symposium on Information Theory in Benelux, Boekelo, Netherlands, (May 2012), pp. 126-133, WIC. ISBN 978-90-365-3383-6



## A human benchmark for automatic speaker recognition

Milou van Dijk<sup>1</sup>, Rosemary Orr<sup>1</sup>, David van der Vloed<sup>2</sup> and David A. van Leeuwen<sup>2,3</sup>

<sup>1</sup>University College Utrecht, The Netherlands

<sup>2</sup>Netherlands Forensic Institute, The Hague

<sup>3</sup>Radboud University Nijmegen, The Netherlands

### Abstract

Automatic Speaker Recognition has a potential to be used in Forensic Speaker Comparison. For the latter, forensic scientists agree that presentation of the comparison to court should be in terms of a calibrated likelihood ratio. In recent years the field of automatic speaker recognition has made significant progress in the analysis, evaluation and calibration of likelihood ratios. In this paper we investigate if speaker comparison by humans can be carried out using the same framework. For this, we use the US National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation 2010 material to measure the performance and calibrate the speaker comparison opinions of human subjects. Because empirical calibration needs a large collection of trials and a human judgment takes a substantial effort, the analysis is carried out for a collection of 40 subjects. From NIST SRE 2010 a subset of 1280 speaker comparison trials are selected. The selection is made using the scores from a state-of-the-art speaker recognition system, such that 1) the trials are representative of the overall performance, in terms of difficulty of the comparisons for the automatic system, and 2) they can be analyzed in three distinct classes ‘hard,’ ‘representative’ and ‘easy.’ Results show that this classification extends to the performance of the human collective, with an Equal Error Rate of 45 %, 25 % and 13 % respectively. Further, the overall human results can be calibrated using ROC convex hull analysis to show a nice linear relation between the 10-level similarity response and a log-likelihood-ratio scale.

### 1. Introduction

With the performance of automatic speaker recognition systems steadily increasing, in part driven by evaluation campaigns such as the NIST Speaker Recognition Evaluations and the various JHU and BOSARIS workshops, and commercial systems becoming readily available, the use of automatic speaker recognition systems for forensic speaker comparison purposes becomes viable. In both automatic speaker recognition and forensic speaker comparison, the task is to compute the likelihood ratio that, given two speech segments, these originate from the same speaker or from different speakers. In formula form, the

likelihood ratio  $r$  is

$$r = \frac{P(\text{speech segments} | H_p)}{P(\text{speech segments} | H_d)}, \quad (1)$$

where  $H_{p,d}$  are the *prosecutor’s* and *defense* hypotheses, stating that the speech segments are produced by the same, or different, speakers, respectively. In automatic speaker recognition, the likelihood ratio can be used to make an optimal Bayes’ decision given a cost function and a prior [1], whereas in forensic speaker comparison this can be used to express the weight of evidence in court [2].

In many countries, forensic speaker comparison is still carried out exclusively by human experts [3, 4], but in some countries forensic examiners are beginning to use automatic speaker recognition in certain cases [4, 5]. An argument, on the one hand, to be reluctant to use automatic systems in a forensic case is that the speech style and electro-acoustical recording conditions of the trace (the evidence) is often quite different from the reference recording in the speaker comparison, and no explicit performance characteristics of the system under those conditions are known. On the other hand, the reports of forensic speaker comparisons are seldom explicit in computation of the likelihood ratio for acoustic-phonetic or linguistic features that are marked as similar between the questioned recording and the reference [6].

The methods for computing likelihood ratios in the automatic speaker recognition domain and in the forensic speaker comparison domain are quite different. In the former, the approach is empirical and the raw discriminative scores of a system are taken as uncalibrated scores, and using a large collection of supervised trials (same-speaker and different-speaker comparisons) an empirical score-to-likelihood-ratio transformation is determined. So far, one of the most robust and effective score-to-likelihood-ratio functions has been an affine transformation of the score  $s$

$$\ell \equiv \log r = as + b, \quad (2)$$

effectively scaling the score by  $a$  and shifting it with  $b$  such that the resulting log-likelihood-ratio has good, probabilistically well interpretable, properties. In the latter, manually or semi-automatically obtained continuous features are directly modeled in same-speaker and



different-speaker distributions [7], or, in the case of discrete features, population frequencies can be used to compute the likelihood ratio, similar to how this happens with DNA. But often, an opinion is formulated where the similarity between the segments is expressed using a “verbal scale” [3,4,6,8]. How such a verbal scale maps to likelihood ratios is, however, a subject of debate. [9, 10]

If we want to get a better insight in how the automatic and the human methods compare we should probably let one do the task of the other, and see what the performance is. One way of doing this is by doing a “human benchmark”: giving a human exactly the same task as the system, and evaluate the performance in the same way. For automatic speaker recognition, such an experiment has been carried out by Schmidt-Nielsen and Crystal [11], and in the NIST Human Assisted Speaker Recognition (HASR) evaluations [12–14]. What the exact ‘human method’ is, is not so important for the performance evaluation and calibration method set forward in this paper; it could be a detailed acoustic-phonetic analysis as performed by forensic experts [3, 4], or holistic acoustic impressions as carried out in this study and others [11, 13, 14].

In this paper, we carry out a similar experiment to Schmidt-Nielsen in a somewhat different setting, and with the goal to investigate a method for determining a score-to-likelihood ratio mapping for human speaker comparison. This is in a way similar to the approach of ATVS-UAM to NIST HASR 2010 [13], where this mapping was taken a linear function. In this approach we study the shape of the mapping and the range of the resulting likelihood ratios.

## 2. Experimental design and details

With humans being limited in the amount of trials they can perform, we have designed an experiment with several goals in mind. The first is that the overall task should be of the same level of difficulty as the system test of NIST SRE 2010. Secondly, we want to study if trials that are hard for a system are also hard for the humans and vice versa. Finally, we want to find a relationship between a verbal scale of similarity and the likelihood ratio.

For empirical performance measurement and calibration we need many trials. Because a single trial takes a human subject a considerable time to complete—the subject should at least listen once to the segments—we decided to determine the performance of human subjects as a whole, effectively integrating out between-subject performance variation. Thus, every subject was exposed to their own set of trials, and the analysis is typically carried out over all subjects.

### 2.1. Trial selection

We used a different trial selection algorithm from what was used in HASR1 in NIST SRE 2010 [12], where pairs of similar speakers were sought for non-target trials and dissimilar target trials using a combination of machine and human judgments, leading to a set of very hard trials. In order to have a range of difficulties in this experiment we selected the trials as follows. We used speech material from the NIST 2010 SRE, telephone-telephone male English condition, a.k.a. ‘det 5’. Our Radboud University Nijmegen (RUN) automatic speaker recognition system [15] computed scores for all trials, i.e., the entire score matrix of train vs. test segments. Then we self-calibrated the scores using logistic regression, i.e., a linear transformation of the log-likelihood-ratio (LLR) scores (2) optimizing a cross-entropy objective function on the test data itself. The log-likelihood-ratio score distributions after calibration are shown in Fig. 1. We then selected trials from three regions, corresponding to difficulty classes: 1) around  $\ell = 0$ : these trials are “hard,” the recognizer cannot separate targets and non-targets; 2) around the modes in the distribution: these trials are “representative”; 3) high target and low non-target scores: these trials can be considered “easy.” These three classes are indicated as shaded bars in Fig. 1. The total amount of trials in each class was 160, 960 and 160 respectively. The trials were further distributed over 40 subjects in such a way that each subject had 4, 24 and 4 trials from each class, respectively, with equal amounts of target and non-target trials per class. This distribution guaranteed that a) each subject is exposed to approximately the same level of difficulty, according to the system, b) the target priors for each subject are the same, c) the overall difficulty is similar to the complete test. Since bias, the tendency of some subjects to find speakers more different where others may find them more the same, has an effect on the calibration of the speaker comparison opinion, we stressed that the target priors of the trials were 50%. This is different from the experiment by Schmidt-Nielsen, where the priors were only *approximately* 0.5, as we did not anticipate subjects to count their own decisions to match the given priors over the 32 trials. For any subject, trials were presented in random order.

### 2.2. Experimental interface

We used a similar experimental interface to what we had used in experiments in human language recognition [16], that had proved to be quite effective. The interface is shown in Fig. 2. Every trial is a comparison of two speech segments, where the task is to determine if the identity of the speaker is the same or not. For both “same” and “different”, five levels of confidence could be specified, named “very uncertain”, “uncertain”, “confident”, “very confident” and “certain”. This configuration is the same

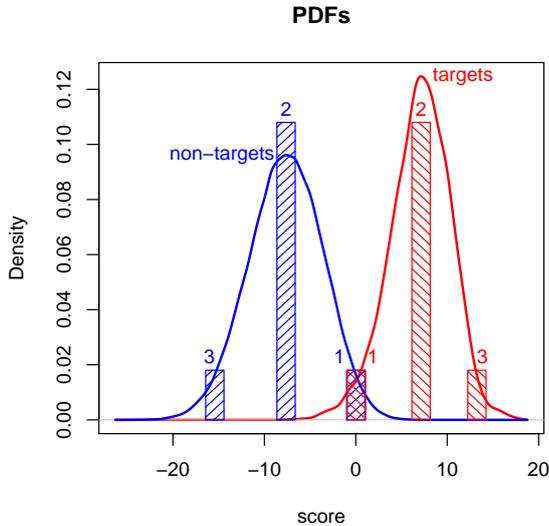


Figure 1: The Probability Density Functions (lines) for target (red) and non-target (blue) scores of the automatic speaker recognition system for NIST SRE 2010 ‘det-5’ male trials after self-calibration. The bars show the LLR score regions and relative quantities from which the trials for the human benchmark were drawn, with numbers indicating the difficulty class.

as in the Schmidt-Nielsen experiment, but the labels have somewhat different naming. The labels are different from the ‘verbal scale’ that is sometimes used in forensic evidence reporting [8], not only in wording (“... support for the prosecution hypothesis”) but also in the omission of an “indecision” option corresponding to a likelihood ratio (LR) of 1. Since the subjects were lay people w.r.t. forensic speaker comparison, we felt the wording in terms of a posterior probability was more intuitive, and is not incorrect given the explicit information about the prior. The subjects were in control of the playback of both segments, they could switch from one to another, and pause, at will. In order to maximize exposure to phonemic variability, playback of a segment would continue where it had been stopped the previous time. We did not provide the subjects with further control over the position of playback. A trial ended when any of the ten response buttons along the “certain: same”–“certain: different” scale was pressed. The interface did not allow for corrections after decisions. For comparison, in the experiment by Schmidt-Nielsen [11], a stimulus was presented as a test-target-test sequence of  $3 \times 3$  s, paced by the subjects. Trials were presented in blocks of 20 with the same target speaker, with 2 min. training of the target speaker before each block. The ten response buttons were ordered left to right, with *extremely certain* at the edges and *uncertain* in the middle, similar to our vertical lay-out (cf. Fig. 2).

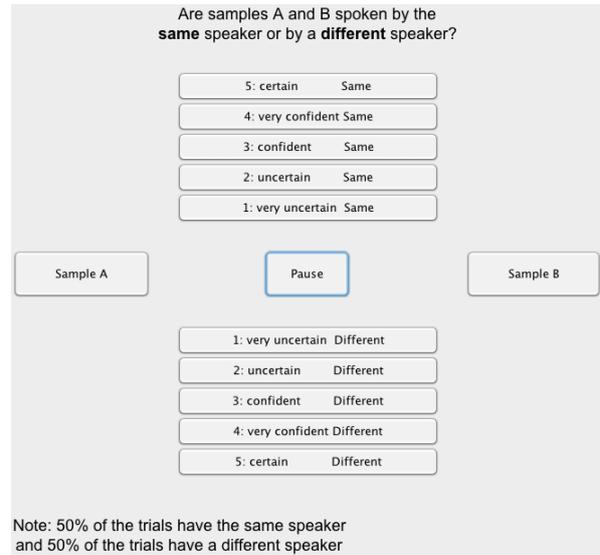


Figure 2: The experimental interface shown to the subjects

### 2.3. Recruitment of subjects

Because the focus of this study is on the evaluation and calibration method for human speaker comparison, we used naïve subjects rather than forensic experts—a much more scarce resource—in order to obtain more trials, similar to [13] and [14]. Forty subjects were recruited from the student community of the *University College Utrecht*, an international liberal arts and sciences establishment. The language of communication at the college is English, and most subjects are non-native English speakers but actively embedded in an English speaking community. Recruitment was carried out through contemporary social media, and subjects were not paid for their efforts. The typical session duration was 30–45 minutes, with some subjects requiring much more time. None of the subjects reported hearing problems, but their hearing abilities were not explicitly tested.

### 2.4. Experimental details

Experiments were conducted in a reasonably quiet environment. The software was run in a Java virtual machine in a Linux virtual environment on a laptop PC. Audio was presented through high quality headphones at a comfortable listening level. Longer periods of silence ( $> 0.5$  s) in the speech had been automatically removed in order to make the experiment more efficient. The subjects received a short introduction about the purpose of the experiment from the experiment leader, and further received instructions through information panels on the screen. One of the screens drew special attention to the fact that in 50% of the trials the speakers were, in fact, the same. This message was also always visible during the main



Table 1: Aggregate statistics for the responses, same/different versus response score. Scores are sorted from “certain: different” to “certain: same.”

trial	-4.5	-3.5	-2.5	-1.5	-0.5	0.5	1.5	2.5	3.5	4.5
diff	152	117	152	74	11	14	40	43	18	19
same	36	26	70	56	22	15	78	138	118	81

experiment (cf. Fig. 2). After this information, six otherwise unused trials were presented as training/habituation. No feedback towards the decision was given in the habituation period, but the trials were chosen according to the easiest selection criteria.

After the experiment of 32 trials, subjects received immediate feedback about their performance in terms of an Equal Error Rate (EER).

### 3. Results

#### 3.1. Overall performance

For analysis of the results, the response buttons are represented as scores, from +4.5 for “certain: same” to -4.5 for “certain: different”. The aggregate response statistics are tabulated in Table 1. The overall results can best be summarized in a Receiver Operating Characteristic (ROC), as in Fig. 3. This is a graph showing the trade off between the probabilities of false alarms (false positives,  $P_{FA}$ ) and misses (false negatives,  $P_{miss}$ ) if a threshold  $t$  for forcing decisions would have been ‘between the response buttons,’ effectively at  $t = -5, -4, \dots, 5$ . The circles correspond to these thresholds, and the line segments to response buttons. Because we have plotted the convex hull (CH) of the ROC, which has a special minimum-cost interpretation [17], some segments actually correspond to a group of adjacent buttons.

An often reported summary of the overall performance is the Equal Error Rate  $E_{=}$ , which we define as the point where  $P_{FA} = P_{miss}$  on the ROC-CH. For the overall data  $E_{=} = 26.5\%$ . We can compute  $E_{=}$  for the different subsets of the data, namely the difficulty classes 1–3 discussed in Section 2.1. From hard to easy, the results are  $E_{=} = 44.8\%, 25.5\%, 13.2\%$ .

We can use the ROC-CH to compute what the optimal log-likelihood-ratio score is corresponding to the buttons, assuming we can treat all subjects as a single ‘system.’ This implicitly assumes that subjects share the same ‘calibration,’ i.e., that one person means more-or-less the same with “confident” as the next. The optimal likelihood that can be associated with the subject’s judgment is just the negative slope of the corresponding ROC-CH line segment. Optimal in this sense means restricting the score-to-likelihood function to be a monotonously increasing function. The result of this operation is plotted in Fig. 4, as the heavy black line.

The likelihood ratio can also be computed by taking

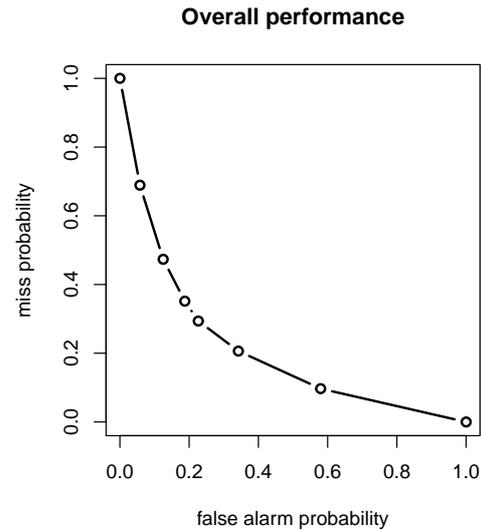


Figure 3: The overall performance, as Receiver Operating Characteristic, using the convex hull.

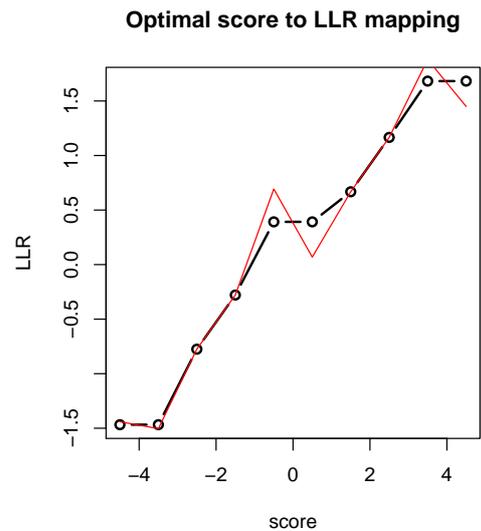


Figure 4: The optimal score-to-log-likelihood ratio function for the response of the human subjects (heavy line), and the LLRs found from ML ratios.



the Maximum Likelihood (ML) estimates of  $P(s | H_p)$  and  $P(s | H_d)$ . From Table 1 we would, e.g., find for the score  $-0.5$  (very uncertain: different) the ratio  $\frac{22}{640} / \frac{11}{640} = \frac{22}{11}$ , resulting in a log-likelihood ratio of  $\log 2$ . The values of such a computation are shown as the red, thin line in Fig. 4. It follows our earlier optimal LLR curve, but is not constrained by monotonicity. This way of computing LLRs directly from the probability density functions (PDFs) observed in development data sets [18] is popular in the forensic science community, but we feel that it has some undesirable properties. For one thing, it can associate a *higher* LR with a *lower* score, as can be observed from Fig. 4 at scores near  $-4$ ,  $0$  and  $4$ , but it can also lead to arbitrarily high and fluctuating LLRs if smoothing parameters that are needed to estimate the PDF for continuous features are chosen badly. The ROC-CH method is more robust in relation to these issues. Note that the ROC-CH method gives exactly the same LLRs as the PAV algorithm [1] does.

### 3.2. Per-subject calibration

As indicated above, the overall performance is expected to be a bit pessimistic because mis-calibration between subjects will lead to worse performance. We can try to compensate for this by calibrating the individual subject's responses using their own performance characteristic, and then pooling their calibrated scores. As a first, cheating, experiment, we use all 32 responses per subject to compute this subject's optimal score-to-likelihood-ratio mapping. This is very similar to the "likelihood-ratio" score combination method used in [11] to combine responses from different listeners,<sup>1</sup> although we use the ROC-CH derived LR (cf. the black line in Fig. 4) and they use the ML method (red line). In order to limit the magnitude of the LLRs, which can easily become  $\pm\infty$  for some trials, we used 'Laplace's rule of succession' [1], which effectively adds additional scores of  $+\infty$  and  $-\infty$  to target and non-target scores before calibration, to cater for scores potentially unobserved in training. The resulting ROC is indicated in Fig. 5 in black, and it shows a much lower overall  $E_{=} = 23.0\%$ . This way of calibrating each subject individually really is "cheating," because the information of the true hypothesis is used for each trial, albeit in a constrained way. If this monotonicity constraint were removed, such cheating would lead to LLRs of  $\pm\infty$ , giving rise to no errors.

A better approach to calibrating individual subjects is to use a cross-validation method. We used the chronological first half of each subject's trials to compute an optimal score-to-LLR transformation, and applied this to the second half of their scores. In order to obtain the same number of scores as before we also reversed this

<sup>1</sup>In [11], this was used to combine responses from subjects for the same trial, where we do it to pool different trials, but the idea is the same.

Individual calibration

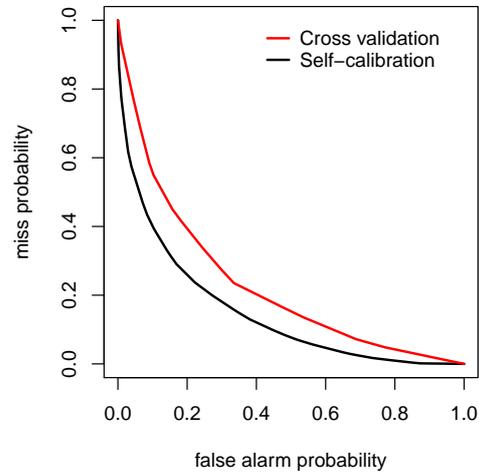


Figure 5: ROC after calibration of individual subject's responses.

operation, i.e., calibrating on the second half and applying this to the first, resulting in a 2-fold cross-validation setup. The results of this individual calibration is shown in red in Fig. 5, which has a lower discrimination performance than the cheating experiment shown in black, with  $E_{=} = 28.8\%$ . This is not better than the original, uncalibrated, pooled scores, with  $E_{=} = 26.5\%$ . Apparently, the 16 scores available per subject are not enough to calibrate an individual. In [13], a similar effect (individual vs. pooled) was observed w.r.t calibration performance.

## 4. Discussion and Analysis

The overall discrimination performance of the human subjects, evaluated as if the judgments are from a single, consistent, system for the SRE 2010 data in our experimental conditions can be summarized as  $E_{=} = 26\%$ . This may appear very high, at first, but we have to point out the caveat that these subjects are naïve w.r.t. the task, and predominantly not native speakers of English. Moreover, the subjects did not use the full content of the speech files, but made decisions after having listened to part of the files. We recorded the exact button-press times, so that we can analyze the average time a subject was exposed to speech. For the overall set, this was 17.7 s per speech segment, ranging per subject from 8.8–34.2 s. Only a small correlation effect ( $-0.4\%/s$ ,  $p = 0.02$ ) could be found between a per-subject  $E_{=}$  and this listening time. More interesting perhaps is that the average duration measured over trial difficulty classes (cf. Section 2.1) drops as 18.8, 17.8, 15.6 s with decreasing difficulty, showing that the easier trials took less effort.

If we want to compare these results to automatic sys-



Table 2: RUN Automatic Speaker Recognition system performance as a function of duration of the speech segments.

duration (s)	5	10	20	40	80
$E_{=}$ (%)	23.3	13.2	6.5	4.4	3.4

tem performance, we have to be very careful. First of all, system performance increases steadily over time, and specifically for this data, because researchers improve their systems using this data as evaluation material. Further, systems have access to the full utterances, and hence use more information per trial, and it is not trivial to change the experimental set-up to allow human subjects to utilize the same amount of information. Probably a set-up with detailed play-back control and spectrographic tools, much as the forensic speaker comparison examiner has, would come closer to these goals but require much more effort on behalf of the subject, making an experiment at this scale (1280 trials) almost impossible. Please note that the detailed Human Assisted Speaker Recognition evaluations of NIST SRE 2010 [12] and 2012 only contained 15 and 20 trials, respectively, and required substantial effort from participants. With these restrictions in mind, we can compare the results on the same trials to the RUN system<sup>2</sup>, which is not the best available system but probably has not yet been over-tuned to this data. We used several data sets where the utterances have been truncated in duration [19]. The discrimination performance results are in Table 2, from which we can conclude that the naïve human performs comparably to the RUN system using about 5 seconds per segment where the humans use 18 seconds.

Related to the relatively low discrimination performance, we find that the range of LLRs is fairly limited, roughly to a magnitude of 1.5 (cf. Fig. 4), corresponding to LRs in the range 0.23–5.4. This means that these subject’s opinion of “certain” (the most extreme confidence available to them) correspond to LRs of roughly  $\frac{1}{4}$  and 5 for this data, which is certainly far away from the ‘verbal likelihood ratio scales’ as used in the literature, which can range from  $10^{-4}$  to  $10^4$  [9]. One could argue that the data or the task simply is too difficult, and that there is the psychological effect that humans want to use the entire response scale for answers, and will scale accordingly. However, if we analyze the mean absolute score for the first and second halves of the trials per subject, we find the values 2.96 and 2.80 which shows, if anything, an opposite effect. One can also argue that a small trial set like this can’t produce any high magnitude LLR anyway, as by virtue of Laplace’s rule of succession values would be limited to  $\sim \log 640 \approx 6.5$ , in a case where all tar-

<sup>2</sup>Using a configuration of the system that was different from the system used in trial selection, so that scores for the selected trials are more evenly distributed than the shaded areas in Fig. 1

Optimal score to LLR mapping

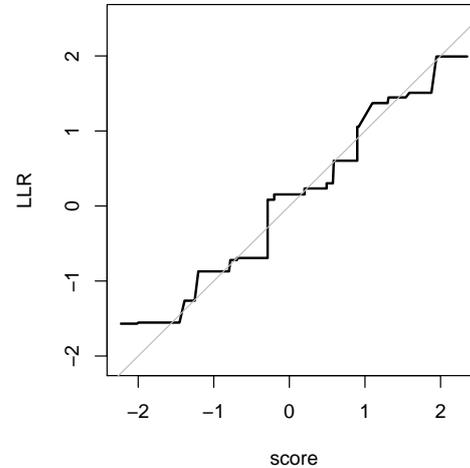


Figure 6: The score-to-LLR mapping after per-subject calibration, cross-validation experiment.

get trials are judged with the same, supportive, response, with no non-target trials with that response. The magnitude of LLRs corresponding to ‘certain’ suffer a bit from the mis-calibration between subjects. If we remove this by using the cross-validation calibration results from Section 3.2, as is shown in Fig. 6, the range of LRs becomes a little wider, about  $\frac{1}{5}$ –7.4. The finding of low magnitude LLR is consistent with [13] where calibrated LRs not exceeding the range 0.1–10 were reported for human speaker comparison.

Despite the low magnitude LLRs observed in the original experiment in Fig. 4, it is interesting that the (raw) score-to-LLR function is fairly linear. No where in the briefing or the experimental protocol a logarithmic relation of the levels of certainty was suggested. Perhaps this is consistent with the interpretation of “weight of evidence,” which, for being an additive quantity functioning on a prototypical weighing scale of justice, must be expressed as log likelihood ratios [20]. In [13], this linear relation between the response value and the LLR was taken as an assumption in the calibration function. It appears we have found support for this assumption in this work.

The choice of ten response values, similar to [11], with a forced decision and visually not linear may have affected the distribution of responses. From Table 1 it can be seen that response values in the middle of the range,  $-0.5$  and  $0.5$  receive less hits, and detailed analysis shows that this is true for all difficulty classes. We can only understand this as a psychological effect of evading ‘extreme’ responses, even though these middle values are not extreme in score value, but only in visual grouping. Such effects should be taken into account for



subsequent experiments, where we would advise to use a linear equidistant set of responses with a middle ground, “undecided” or “ $LR = 1$ .” Ramos [13] used a 7-point confidence scale with indeed a middle value indicating  $LR = 1$ . From Fig. 4, there is evidence that the correct interpretation of both “very uncertain” responses is  $\ell \approx 0$ , and it is probably better to allow explicitly for such a response.

## 5. Conclusions

The performance of the “human system,” whether internally calibrated or not, is not to be taken as representative for that of manual forensic speaker comparison: here, we work with lay listeners instead of experts, the exposure to the speech is very limited, the listeners are mostly non-native in the spoken language, long silences were removed, and the speech material is not taken from actual cases. However, the method of empirical calibration—as we are used to in automatic speaker recognition—should be possible to carry out with forensic experts. A practical problem, however, is the amount of effort that such an empirical calibration would take. If experts take, e.g., two weeks to form a well-founded opinion for speaker similarity for a single trial, a calibration at the scale of this experiment would take many person-years. Even if such an effort is taken, it is difficult to keep the ‘internal calibration’ of the expert constant over the entire period. Finally, the stimulus material must be made up of trials for which the true hypothesis is known, effectively ruling out real case material that includes the questioned recording. However, from ‘collateral’ case material it should be possible to generate trials where the hypothesis is known with negligible uncertainty. We would advocate that the proposed method of empirically calibrating opinions should somehow be carried out. Perhaps there are paradigms feasible in which the average time per trial is reduced, e.g., by grouping these for the same target speaker or by structurally including calibration trials in the standard operating procedure in forensic case work.

## 6. Acknowledgments

The research leading to these results has in part received funding from the European Community’s Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 238803.

## 7. References

- [1] N. Brümmner and J. du Preez, “Application-independent evaluation of speaker detection,” *Computer Speech and Language*, vol. 20, pp. 230–275, 2006.
- [2] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, and J. Ortega-Garcia, “Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2104–2115, September 2007.
- [3] T. Cambier-Langeveld, “Current methods in forensic speaker identification: Results of a collaborative exercise,” *The International Journal of Speech, Language and the Law*, vol. 14, no. 2, pp. 223–243, 2007.
- [4] E. Gold and P. French, “An international investigation of forensic speaker comparison practices,” in *The 17th International Congress of Phonetic Sciences (ICPhS)*, 2011, pp. 751–754.
- [5] H. J. Künzell, “Automatic speaker recognition with cross-language speech material,” *The International Journal of Speech, Language and the Law*, vol. 20, no. 1, pp. 21–44, 2013.
- [6] P. French *et al.*, “Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases,” *The International Journal of Speech, Language and the Law*, vol. 14, no. 1, pp. 137–144, 2007.
- [7] C. Aitken and D. Lucy, “Evaluation of trace evidence in the form of multivariate data,” *Applied Statistics*, pp. 109–122, 2004.
- [8] C. Champod and I. W. Evett, “Commentary on broeders (1999): “some observations on the use of probability scales in forensic identification,”” *Forensic Linguistics*, vol. 7, no. 2, pp. 238–243, 2000.
- [9] P. Rose, *Forensic Speaker Identification*. Taylor & Francis, 2002.
- [10] A. Nordgaard *et al.*, “Scale of conclusions for the value of evidence,” *Law, Probability and Risk*, vol. 11, no. 1, pp. 1–24, 2011.
- [11] A. Schmidt-Nielsen and T. H. Crystal, “Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 speaker evaluation data,” *Digital Signal Processing*, vol. 10, pp. 249–266, 2000.
- [12] C. Greenberg, A. Martin, L. Brandschain, J. Campbell, C. Cieri!, G. Doddington, and J. Godfrey, “Human assisted speaker recognition in nist sre10,” in *Proc. of Odyssey Speaker and Language Recognition Workshop*. ISCA, 2010, pp. 180–185.
- [13] D. Ramos, J. Franco-Pedroso, and J. Gonzalez-Rodriguez, “Calibration and weight of the evidence



## Assessing latent fingerprint distortion using forensic databases and minutiae paring by human experts

Ruifang Wang <sup>a</sup>, Daniel Ramos <sup>a</sup>, Didier Meuwly <sup>b</sup>, Raymond Veldhuis <sup>c</sup>,  
Julian Fierrez <sup>a</sup> and Rudolf Haraksim <sup>b</sup>

<sup>a</sup> Biometric Recognition Group - ATVS, EPS, Universidad Autonoma de Madrid, Spain

<sup>b</sup> WISK, Netherlands Forensic Institute, The Hague, The Netherlands

<sup>c</sup> Signals and Systems Group, EEMCS, University of Twente, the Netherlands

{ruifang.wang, daniel.ramos, julian.fierrez}@uam.es

{d.meuwly, r.haraksim}@nfi.minvenj.nl, R.N.J.Veldhuis@utwente.nl

### Abstract

Large non-linear distortion in a finger mark, introduced at the time when a latent print was left in crime scenes, is one main difficulty in forensic fingerprint matching [1] compared to plain or rolled fingerprint matching. It is significant and desirable that AFIS systems in forensic applications have much higher distortion tolerance compared to those systems for civil purposes. Moreover, recent work [2] has highlighted the importance of distortion models to assess the evidential value of fingerprint comparison. Previous works [3, 4, 5, 6] on studying distortion in fingerprints mainly described two aspects. On the one hand, in [6], the flexibility of finger skin under different stresses based on the knowledge of the anatomy of the hand and skin was illustrated. On the other hand, researchers from the biometrics area studied quantitative (measurable) aspects of finger skin deformation by proposing distortion models [4, 5] aimed to enhance matching algorithms for AFIS systems.

However, those distortion models, heavily depending on the completeness of fingerprint images, will face problems when applied to forensic scenarios, mainly due to the serious partial overlap found between a finger mark and its paired fingerprint. To avoid the dependence on the completeness of fingerprint images, we focus on distortion assessment at feature level (generally minutiae features). This is inspired by the fact that when non-linear distortion is introduced at image level, consequently, it is the spatial location of minutiae confounding the matching process. Furthermore, to deal with non-linear distortion at matching stage, existing algorithms [3] essentially present local structure matching, which was deemed to have high distortion tolerance. This motivates us to study distortion assessment through minutiae windows under different window shapes as a local structure.

Based on the motivations described above, in this work, we propose a method to assess distortion in latent fingerprints, which uses forensic databases of marks and mated prints. The method needs the information about minutiae paring provided by human examiners, i.e., which minutiae in the finger mark is paired to which minutiae in the fingerprint, and works with groups of minutiae in the mark and the print that define so-called minutiae windows, which are used for distortion assessment. We approach to the problem of distortion assessment in two stages. First, we compare different window shapes, selecting the one that presents less variation between the window in the mark and in the print. We call this measure of variation global distortion, or distortion at the window level. Second, we

compute the variation of the minutiae points within a given window, namely local distortion or feature-level distortion. Thus, global distortion is a criteria to select the most stable comparison window, whereas the ultimate measure of distortion is the local one. In the experimental study, two forensic fingerprint databases are used, i.e., NIST SD27 [7] including 258 pairs of finger marks and fingerprints, and a casework database collected by Netherlands Forensic Institute (NFI) [8] including 58 pairs of finger marks and fingerprints. The information about the paring of minutiae in the mark and those in the print is provided by forensic examiners. The results show that the distortion of a circle window is the smallest among various types of window shapes tested, namely, i.e., rectangular, elliptic, and circular. Also, the distortion assessment of minutiae points under circle window can show non-linear distortion quantitatively, which can help with the design of forensic fingerprint comparison algorithms.

### 1. References

- [1] A.K. Jain and J. Feng, "Latent fingerprint matching," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 88–100, 2011.
- [2] C. Neumann, I. W. Evett, and J. Skerrett, "Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm," *Journal of the Royal Statistical Society: Series A*, vol. 175, no. 2, pp. 371–415, 2012.
- [3] Davide Maltoni, Dario Maio, Anil K. Jain, and Salil Prabhakar, Eds., *Handbook of Fingerprint Recognition*, chapter 4, Springer, London, UK, 2009.
- [4] R. Cappelli, D. Maio, and D. Maltoni, "Modelling plastic distortion in fingerprint images," in *Advances in Pattern Recognition ICAPR 2001*, vol. 2013, pp. 371–378. Springer Berlin Heidelberg, Berlin, Germany, 2001.
- [5] Arun Ross, Sarat Dass, and Anil Jain, "A deformable model for fingerprint matching," *Pattern Recognition*, vol. 38, no. 1, pp. 95–103, 2005.
- [6] Alice V. Maceo, "Qualitative assessment of skin deformation: A pilot study," *Journal of Forensic Identification*, vol. 59, no. 4, pp. 390–440, 2009.
- [7] National Institute of Standards and Technology, "Nist special database 27," 2000.
- [8] Netherlands Forensic Institute, "Csi the hague project," 2013.



# Normalized Ordinal Distance; A Performance Metric for Ordinal, Probabilistic-ordinal or Partial-ordinal Classification Problems

Mohammad Hasan Bahari, Hugo Van hamme

Center for processing speech and images, KU Leuven, Belgium

{mohamadhasan.bahari, hugo.vanhamme}@esat.kuleuven.be

## Abstract

In many forensic scenarios, we deal with problems of ordinal nature, where there is intrinsic ordering between the categories. For example, in human age group recognition from speech or images, the categories can be child, young, middle-aged and senior. In detecting the level of intoxication, the categories can be low, medium and high. In this paper, a novel application-independent performance metric for ordinal, probabilistic-ordinal and partial-ordinal classification problems is introduced. Conventional performance metrics for ordinal classification problems, such as mean absolute error of consecutive integer labels and ranked probability score, are difficult to interpret and can be misleading. In this paper, first, the ordinal distance between two arbitrary vectors in Euclidean space is introduced. Then, a new performance metric, namely normalized ordinal distance, is proposed based on the introduced ordinal distance. This performance metric is conceptually simple, computationally inexpensive and application-independent. The advantages of the proposed method over the conventional approaches and its different characteristics are shown using several numerical examples.

## 1. Introduction

A large number of real world classification problems are ordinal, where there is intrinsic ordering between the categories. For example, in quality prediction systems, the task is to categorize the quality of a product into bad, good and excellent [1]. In human age group recognition from speech or images, the categories can be child, young, middle-aged and senior [2, 3]. In the classification of the therapeutic success, the classes are good recovery, moderate disability, severe disability, and fatal outcome [4]. In all ordinal classification problems ( $C_O$ ), the class labels are ordinal numbers, i.e. there is intrinsic ordering between the categories.

Probabilistic-Ordinal and Partial-Ordinal Classification problems, labeled as  $C_O^{\text{Pr}}$  and  $C_O^{\text{Pa}}$  respectively, are well-known generalizations of the  $C_O$ . In  $C_O^{\text{Pr}}$ , for a test datapoint, the classifier calculates the probability of belonging to each category. In  $C_O^{\text{Pa}}$ , instead of the crisp class labels each datapoint has a degree of membership

to every class [5]. These types of problems, explained in sections 2.2 and 2.3 in detail, can be found in many domains, such as natural language processing, social network analysis, bioinformatics and agriculture [5].

Scientists have proposed different methods to solve  $C_O$ ,  $C_O^{\text{Pr}}$  and  $C_O^{\text{Pa}}$  [5, 6, 7, 8, 9, 10]. For example, McCullagh introduced an ordinal classifier, namely the proportional odds model (POM), based on logistic regression [6]. In [7],  $C_O$  is addressed using a generalization of support vector machines (SVM) namely support vector ordinal regression (SVOR). A neural network approach for the  $C_O$  is suggested in [8]. [9] suggested Gaussian processes for  $C_O$ . In [5], kernel-based proportional odds models is introduced to solve the  $C_O^{\text{Pa}}$ .

To measure the performance of these classifiers, different approaches have been suggested. For example, mean zero-one error ( $E_{mzo}$ ) and mean absolute error of consecutive integer labels ( $E_{ma}^{\text{cil}}$ ) are widely applied to measure the performance of the classifiers in  $C_O$  [7, 8, 9, 10]. However, non of these methods are applicable to  $C_O^{\text{Pr}}$  and  $C_O^{\text{Pa}}$ . Percentage of correctly fuzzy classified instances ( $P_{\text{cfci}}$ ) and Average Deviation ( $E_{\text{ad}}$ ) have been suggested to measure the classifier performance in  $C_O^{\text{Pr}}$  and  $C_O^{\text{Pa}}$  [5, 11, 12, 13]. The main drawback of  $P_{\text{cfci}}$  is that it does not consider the order of categories [11, 12]. The  $E_{\text{ad}}$  suggests a simple idea to solve this problem [12, 13]. Although the  $E_{\text{ad}}$  is attractive from several aspects, the interpretation of its results is difficult, because the range of its output depends on the application. The same difficulty is observed in  $E_{ma}^{\text{cil}}$ . Application dependency makes the interpretation of  $E_{ma}^{\text{cil}}$  and  $E_{\text{ad}}$  very challenging. The average of ranked probability scores ( $E_{\text{rps}}$ ), is also applied as a performance metric in  $C_O^{\text{Pr}}$  and  $C_O^{\text{Pa}}$  [14, 15]. In this method, the order of categories is important and the range of the output is fixed between 0 and 1. This method can be applied to  $C_O$ ,  $C_O^{\text{Pr}}$  and  $C_O^{\text{Pa}}$ . However, analysis reveals that  $E_{\text{rps}}$  over estimates the performance of classifiers in many situations. This issue, which leads to a erroneous interpretation of classifier performance, is illustrated by some numerical examples in section 5.

In this paper, we investigate different characteristics of these performance metrics and finally a new application-independent performance metric, namely Normalized Or-



dinal Distance ( $E_{\text{nod}}^P$ ), is introduced. The Matlab code of the suggested approach, which can be applied to all three types of considered problems  $C_O$ ,  $C_O^{\text{Pr}}$  and  $C_O^{\text{Pa}}$ , can be downloaded from our website<sup>1</sup>.

This paper is organized as follows. In section 2, the mathematical formulations of  $C_O$ ,  $C_O^{\text{Pr}}$  and  $C_O^{\text{Pa}}$  are presented. In section 3, five different conventional performance metrics are explained. The proposed performance metric is elaborated in section 4. In section 5, the effectiveness of the proposed approach is illustrated using some numerical examples. The paper ends with a conclusion in section 6.

## 2. Problem Formulation

In this section, the ordinal, probabilistic-ordinal and partial-ordinal problems are formulated.

### 2.1. Ordinal Classification

Assume that we are given a training data set  $S^{\text{tr}} = \{(X_1, Y_1), \dots, (X_n, Y_n), \dots, (X_N, Y_N)\}$ , where  $X_n = [x_{n,1}, \dots, x_{n,i}, \dots, x_{n,I}]$  denotes a vector of observed characteristics of the data item and  $Y_n = [y_{n,1}, \dots, y_{n,d}, \dots, y_{n,D}]$  denotes a label vector. The label vector is defined as follows if  $X_n$  belongs to class  $C_d$ :

$$y_{n,j} = \delta_{j,d}. \quad (1)$$

where  $\delta$  denotes the Kronecker delta. In ordinal problems, there is an intrinsic ordering between the classes, which is denoted as  $C_1 < \dots < C_d < \dots < C_D$  like low, medium and high [5]. The goal is to approximate a classifier function ( $G$ ), such that for the  $m^{\text{th}}$  unseen observation  $X_m^{\text{tst}}$ ,  $\hat{Y}_m = G(X_m^{\text{tst}})$  is as close as possible to the true label. For a crisp classifier  $\hat{Y}_m$  is defined as follows if the  $d^{\text{th}}$  class is chosen for  $X_m^{\text{tst}}$

$$\hat{y}_{m,j} = \delta_{j,d}. \quad (2)$$

### 2.2. Probabilistic-Ordinal classification

Probabilistic-ordinal classification problem ( $C_O^{\text{Pr}}$ ) is a generalization of the  $C_O$ , where each element of the classifier output vector ( $\hat{Y}$ ) represents the probability of belonging to the corresponding category. In this type of classification,  $Y_n$  is defined by relation (1). However,  $\hat{Y}_m$  is defined as follows

$$\hat{Y}_m = \left\{ \begin{array}{l} [\hat{y}_{m,1}, \dots, \hat{y}_{m,d}, \dots, \hat{y}_{m,D}] \in \mathbb{R}^D \\ \hat{y}_{m,d} \geq 0; \sum_{d=1}^D \hat{y}_{m,d} = 1 \end{array} \right\}, \quad (3)$$

where  $\mathbb{R}$  denotes the set of real numbers.

### 2.3. Partial-Ordinal Classification

Partial-ordinal classification problem ( $C_O^{\text{Pa}}$ ) is another generalization of  $C_O$  [5]. In ordinal problems, each data object

is limited to belong to a single category, i.e. out of all  $D$  elements of  $Y_n$ , only one is nonzero. However, this is too conservative in the case of non-crisp or fuzzy classes. This limitation is relaxed in  $C_O^{\text{Pa}}$  by rephrasing  $Y_n$  as follows

$$Y_n = \left\{ \begin{array}{l} [y_{n,1}, \dots, y_{n,d}, \dots, y_{n,D}] \in \mathbb{R}^D \\ y_{n,d} \geq 0; \sum_{d=1}^D y_{n,d} = 1 \end{array} \right\}. \quad (4)$$

Therefore, each datapoint has a degree of membership to all classes. Like in ordinal problems, the final goal is to approximate a classifier function ( $G$ ), such that for an unseen observation  $X^{\text{tst}}$ ,  $\hat{Y}_m = G(X_m^{\text{tst}})$  is as close as possible to the true label. In this type of classification  $\hat{Y}_m$  is also defined by relation 3.

## 3. Conventional Performance Metrics

In this section, five widely-used conventional metrics, namely  $E_{\text{mzo}}$ ,  $E_{\text{ma}}^{\text{cil}}$ ,  $P_{\text{cfci}}$ ,  $E_{\text{ad}}$  and  $E_{\text{rps}}$  are introduced [5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17].

### 3.1. Mean Zero-One Error ( $E_{\text{mzo}}$ )

Performance metric  $E_{\text{mzo}}$  is the fraction of incorrect predictions, which is calculated as follows [7, 8, 9, 10]

$$E_{\text{mzo}} = \frac{1}{M} \sum_{m=1}^M (1 - \delta_{\hat{y}_m, y_m}), \quad (5)$$

where  $M$  is the total number of test set datapoints,  $\hat{y}_m$  is the predicted label of the  $m^{\text{th}}$  test set datapoint and  $y_m$  is the true label of the  $m^{\text{th}}$  test set datapoint. The main advantage of  $E_{\text{mzo}}$  is its simplicity. However, it does not consider the order of the categories. Furthermore, it is not applicable to measure the performance in  $C_O^{\text{Pr}}$  or  $C_O^{\text{Pa}}$ .

### 3.2. Mean Absolute Error of Consecutive Integer Labels ( $E_{\text{ma}}^{\text{cil}}$ )

To calculate the  $E_{\text{ma}}^{\text{cil}}$ , first, both true labels and predicted labels of the test set datapoints are transformed into consecutive integers so that if the  $d^{\text{th}}$  column of the label vector is 1 then the transformed label is equal to  $d$  [7, 8, 9, 10]. After label transformation the  $E_{\text{ma}}^{\text{cil}}$  is calculated as follows

$$E_{\text{ma}}^{\text{cil}} = \frac{1}{M} \sum_{m=1}^M |\hat{U}_m - U_m|, \quad (6)$$

where  $\hat{U}_m$  is the transformed predicted label of the  $m^{\text{th}}$  test set datapoint and  $U_m$  is the transformed true label of the  $m^{\text{th}}$  test set datapoint. The  $E_{\text{ma}}^{\text{cil}}$  enjoys the advantage of considering the order of categories into account. However, it cannot be applied to evaluate the classifiers in  $C_O^{\text{Pr}}$  or  $C_O^{\text{Pa}}$ . Moreover, the range of its output is application-dependent. Therefore, the interpretation of this metric is challenging. This is shown in section 5 using some numerical examples.

<sup>1</sup><http://www.esat.kuleuven.be/psi/spraak/downloads/>



### 3.3. Percentage of Correctly Fuzzy Classified Instances ( $P_{cfci}$ )

Performance metric  $P_{cfci}$  has been applied to measure the performance of probabilistic or fuzzy classifiers [11, 12]. It is calculated as follows

$$P_{cfci} = \frac{100}{M} \sum_{m=1}^M \left(1 - \frac{1}{2} \sum_{d=1}^D |\hat{y}_{m,d} - y_{m,d}|\right). \quad (7)$$

As it can be inferred from the above relation, the order of the categories is not considered in  $P_{cfci}$ .

### 3.4. Average Deviation ( $E_{ad}$ )

Performance metric  $E_{ad}$  was originally introduced by Van Broekhoven [12] to evaluate the classifiers in fuzzy ordered classification problems. It was also applied in different applications with other names [5, 13]. The  $E_{ad}$  is calculated as follows

$$E_{ad} = \frac{1}{M} \sum_{m=1}^M \left\{ \sum_{d=1}^{D-1} \left| \sum_{i=1}^d \hat{y}_{m,i} - \sum_{i=1}^d y_{m,i} \right| \right\}. \quad (8)$$

It can be interpreted from the above relation that the order of categories is important in  $E_{ad}$ . It is also useful for classifier evaluation in  $C_O^{Pr}$  or  $C_O^{Pa}$ . However, similar to  $E_{ma}^{cil}$ , the range of  $E_{ad}$  is application-dependent and hence difficult to interpret.

### 3.5. Average Ranked Probability Scores ( $E_{rps}$ )

The ranked probability score was originally introduced to score the output of probabilistic classifiers [14, 15]. It is defined as follows

$$RPS_Y(\hat{Y}) = \frac{1}{D-1} \left\{ \sum_{d=1}^{D-1} \left( \sum_{i=1}^d \hat{y}_i - \sum_{i=1}^d y_i \right)^2 \right\}. \quad (9)$$

This scoring rule can be easily extended to measure the performance of classifiers in  $C_O$ ,  $C_O^{Pr}$  and  $C_O^{Pa}$  using the following relation.

$$E_{rps} = \frac{1}{M(D-1)} \sum_{m=1}^M \sum_{d=1}^{D-1} \left( \sum_{i=1}^d \hat{y}_{m,i} - \sum_{i=1}^d y_{m,i} \right)^2 \quad (10)$$

As it can be interpreted from the above relation, the order and the number of categories are important in  $E_{rps}$ . It is assumed that the maximum of the nominator of  $E_{rps}$  is  $M(D-1)$ . Therefore, to fix the range of  $E_{rps}$  between 0 and 1 the nominator is divided to its maximum possible value  $M(D-1)$ . However, this assumption is very conservative so that in many practical cases the maximum of the nominator of  $E_{rps}$  is less than  $M(D-1)$ . Consequently, this assumption may lead to an erroneous interpretation of the classifier performance. Numerical examples of Section 5 reveal this issue clearly.

## 4. Proposed Performance Metric

In this section, first, Ordinal Distance (OD) of two vectors in Euclidean space is introduced. Then, a new performance metric, namely normalized ordinal distance ( $E_{nod}^p$ ), is developed based on the ordinal distance.

### 4.1. Ordinal Distance (OD)

In this section, the definition of a distance function is recaptured. Then, the Minkowski distance is described and finally, the ordinal distance is introduced as an extension of the Minkowski distance.

#### 4.1.1. Distance

By definition, a distance function of two points  $A = [a_1, \dots, a_d, \dots, a_D]$  and  $B = [b_1, \dots, b_d, \dots, b_D]$  is a function  $D: \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ , which satisfies the following three conditions [18]:

1.  $D(A, B) \geq 0$  and  $D(A, B) = 0 \Leftrightarrow A = B$
2.  $D(A, B) = D(B, A)$
3.  $D(A, C) \leq D(A, B) + D(B, C)$

A variety of distance functions have been introduced by scientists for different applications such as Minkowski distance, Mahalanobis distance, Chebyshev distance and Hamming distance [18].

#### 4.1.2. The Minkowski distance of order $p$

The Minkowski distance of order  $p$  or  $p$ -norm is a distance function, which satisfies all conditions of a distance function.

$$\|A - B\|_p = \left( \sum_{d=1}^D |a_d - b_d|^p \right)^{1/p} \quad (11)$$

where  $p$  is a real number not less than 1. As in can be interpreted from relation (11), in  $p$ -norm, the order of the elements of two points  $A$  and  $B$ , is not important.

#### 4.1.3. The Ordinal distance of order $p$

The notion of ordinal distance is previously used to measure the differences of two strings [19] or two histograms [20]. In this paper, an ordinal distance of two vectors in Euclidean space is introduced. The Ordinal Distance of order  $p$  between two points  $A$  and  $B$  is defined as follows

$$\|A - B\|_p^{OD} = \left( \sum_{d=1}^D |\bar{a}_d - \bar{b}_d|^p \right)^{1/p} \quad (12)$$

$$\bar{a}_d = \sum_{i=1}^d a_i$$

$$\bar{b}_d = \sum_{i=1}^d b_i$$



where  $p$  is a real number not less than 1. Since (12) is a Minkowski distance between  $\bar{A} = [\bar{a}_1 \dots \bar{a}_d \dots \bar{a}_D]$  and  $\bar{B} = [\bar{b}_1 \dots \bar{b}_d \dots \bar{b}_D]$ , it follows that the ordinal distance of order  $p$  satisfies the conditions of section 4.1.1.

#### 4.2. Normalized Ordinal Distance ( $E_{\text{nod}}^p$ )

In this section, a new performance metric, namely normalized ordinal distance ( $E_{\text{nod}}^p$ ), is introduced to measure the performance classifiers in  $C_O$ ,  $C_O^{\text{Pr}}$  and  $C_O^{\text{Pa}}$ .

$$E_{\text{nod}}^p = \frac{\sum_{m=1}^M \|Y_m - \hat{Y}_m\|_p^{\text{OD}}}{\sum_{m=1}^M \psi_{Y_m}^p}, \quad (13)$$

where  $\psi_{Y_m}^p$  is the upper bound of  $\|Y - \hat{Y}\|_p^{\text{OD}}$  for any possible  $\hat{Y}$  in its defined range.  $\psi_Y$  is defined as follows

$$\psi_Y^p \triangleq \max_T \|Y - T\|_p^{\text{OD}}, \quad (14)$$

where  $T = \{t_1, \dots, t_d, \dots, t_D\}$  is an arbitrary vector with the same specifications of  $\hat{Y}$  mentioned in relation (2), i.e.,  $T$  lies on a simplex.  $\psi_Y^p$  can be calculated using theorem 1.

##### Theorem 1:

The upper bound of  $\|Y - \hat{Y}\|_p^{\text{OD}}$  for any possible  $\hat{Y}$  can be obtained as follows

$$\psi_Y^p = \max(\|Y - L_1\|_p^{\text{OD}}, \dots, \|Y - L_d\|_p^{\text{OD}}, \dots, \|Y - L_D\|_p^{\text{OD}}) \quad (15)$$

or equivalently

$$\psi_Y^p = \max(\|Y - L_1\|_p^{\text{OD}}, \|Y - L_D\|_p^{\text{OD}}), \quad (16)$$

where  $L_d$  is a vector of size  $Y$ . The  $d^{\text{th}}$  element of  $L_d$  is equal to 1 and the rest of elements are zero. As it can be interpreted from relations (15) and (16), although the latter one is more restrictive, it provides an easier way to calculate  $\psi_Y^p$ .

##### Proof:

We first prove the relation (15), which help us to show the correctness of relation (16).

*Proof of relation (15):*

By definition

$$\|Y - T\|_p^{\text{OD}} = \|\Lambda(Y - T)\|_p, \quad (17)$$

where  $\Lambda$  is a lower triangular matrix of size  $D \times D$  with all diagonal and lower diagonal elements equal to 1. Since  $\|(Y - T)\|_p$  is a convex function of  $T$  and a convex function remains convex under an affine transformation,  $\|\Lambda(Y - T)\|_p$  is also convex.

On the other hand, a convex function on a compact convex set attains its maximum at an extreme point of the set [21]. In this problem  $T \in \{[t_1, \dots, t_d, \dots, t_D] \in \mathbb{R}^D | t_d \geq 0; \sum_{d=1}^D t_d = 1\}$ , i.e.,  $T$  lies on a simplex. The extreme points of this compact convex set are  $L_d$  with

$d \in \{1, \dots, D\}$ .

Therefore,

$$\max_T \|\Lambda(Y - T)\|_p = \max(\|\Lambda(Y - L_1)\|_p, \dots, \|\Lambda(Y - L_d)\|_p, \dots, \|\Lambda(Y - L_D)\|_p) \quad (18)$$

Consequently,

$$\max_T \|Y - T\|_p^{\text{OD}} = \max(\|Y - L_1\|_p^{\text{OD}}, \dots, \|Y - L_d\|_p^{\text{OD}}, \dots, \|Y - L_D\|_p^{\text{OD}}) \quad (19)$$

*Proof of relation (16):*

Relation (16) is now shown by contradiction. Suppose relation (15) is not equivalent with relation (16), then there must be a  $k \in \{2, \dots, D-1\}$  such that both Relations 20 nad 21 hold.

$$\|Y - L_k\|_p^{\text{OD}} > \|Y - L_1\|_p^{\text{OD}} \quad (20)$$

$$\|Y - L_k\|_p^{\text{OD}} > \|Y - L_D\|_p^{\text{OD}} \quad (21)$$

Expansion of relation (20) and (21) is

$$\sum_{d=1}^{k-1} (\sum_{i=1}^d y_i)^p + \sum_{d=k}^{D-1} (1 - \sum_{i=1}^d y_i)^p > \sum_{d=1}^{D-1} (1 - \sum_{i=1}^d y_i)^p \quad (22)$$

$$\sum_{d=1}^{k-1} (\sum_{i=1}^d y_i)^p + \sum_{d=k}^{D-1} (1 - \sum_{i=1}^d y_i)^p > \sum_{d=1}^{D-1} (\sum_{i=1}^d y_i)^p. \quad (23)$$

After some manipulations (22) and (23) lead to

$$\sum_{d=1}^{k-1} \left[ (\sum_{i=1}^d y_i)^p - (1 - \sum_{i=1}^d y_i)^p \right] > 0 \quad (24)$$

$$\sum_{d=k}^{D-1} \left[ (1 - \sum_{i=1}^d y_i)^p - (\sum_{i=1}^d y_i)^p \right] > 0. \quad (25)$$

If relation (24) holds,  $(\sum_{i=1}^d y_i) > (1 - \sum_{i=1}^d y_i)$  hence  $(\sum_{i=1}^d y_i) > 0.5$  for at least one  $d$  between 1 and  $k-1$ . Likewise, from (25),  $(\sum_{i=1}^d y_i) < 0.5$  for at least one  $d$  between  $k$  and  $D-1$ . This is impossible, since  $\sum_{i=1}^d y_i$  is an increasing function of  $d$  and hence (16) holds.

##### 4.2.1. Relation to $E_{\text{rps}}$

There is a close relationship between  $E_{\text{rps}}$  and  $E_{\text{nod}}^p$  specially for  $p = 2$ . In both  $E_{\text{rps}}$  and  $E_{\text{nod}}^p$ , denominators are assumed to be the upper bound of the nominator and are used to keep the range of performance metric between 0 and 1. In  $E_{\text{rps}}$ , it is assumed that the upper bound of the nominator is  $M(D-1)$  [15, 22]. However, this is a conservative bound in many situations. This is illustrated by some numerical examples in section 5. We will also show this conservative assumption can result in a misleading or erroneous interpretation of the classifiers performance. In  $E_{\text{nod}}^p$ , this upper bound is explicitly defined by relation (14) and calculated by relation (16).



## Towards Quantification of the Weight of Evidence with Partial Fingermarks on Real Forensic Casework

**Ram P. Krish, Julian Fierrez, Daniel Ramos, Ruifang Wang**

Biometric Recognition Group - ATVS, EPS - Univ. Autonoma de Madrid  
C/ Francisco Tomas y Valiente, 11 - Campus de Cantoblanco - 28049 Madrid, Spain  
Email: {ram.krish, julian.fierrez, daniel.ramos, ruifang.wang}@uam.es

In forensic fingerprint examinations, there is a need for statistical techniques to quantitatively assess the weight of the evidence and measure the performance of the fingerprint comparisons in more adequate ways that should be both empirical and scientific. The first step towards this evidence quantification for statistical reporting is to build evidence-evaluation methods that generate score from pairs of latent and impression images that the examiner has annotated with minutiae features manually. In real forensic casework scenario, the examiner manually compares the given latent fingerprint against a reference fingerprint and arrives at a logical conclusion of identification-exclusion decision based on examiner's training and experience following the ACE-V protocol, but recently this kind of procedures have been criticized, arguing for a more quantitative evaluation of the weight of the evidence to be produced in court.

There is no scientific framework in use at the criminal justice system to characterize the uncertainty involved in the ACE-V procedure, as well as to express the strength of opinion of the forensic examiner quantitatively. Such a requirement has been articulated in several influential reports like the NRC 2009 report and the NIST Human Factors report. The new paradigm coming forward in this regard avoids hard identification decisions by considering evidence reporting methods that incorporate uncertainty and statistics. Among all the methods of evidence evaluation, the likelihood ratio is receiving greater attention. To use any statistics-based framework for quantification of evidence, scores are required at the ACE-V stage in place of logical decision (*match*, *non-match* or *the comparison is inconclusive*). We proposed a framework to generate a score from the matched latent template and the reference impression template at the ACE-V stage [1] [2]. Such a score can be utilized to quantitatively express the strength of opinion of the forensic examiner using statistics-based framework.

Together with the description of the new realistic forensic casework driven score computation, we also exploited the developed framework to study the discriminating power of matched template [1] on the NIST Special Database (SD) 27 and the real forensic fingerprint database (GCDB) acquired from The Guardia Civil, the law enforcement agency of the Government of Spain. Along with the location and orientation attributes for minutiae, GCDB also consisted of type information for each minutiae. Apart from typical minutiae features (*ridge-ending and bifurcations*), GCDB also consisted of other rare minutiae features like *fragments, enclosures, points/dots, interruptions* etc. We also exploited the developed score computation framework to study the importance of rare minutiae features in the matched templates [2]. The results shows the feasibility of the developed approach towards quantification of the weight of evidence in forensic caseworks.

### References

- [1] R.P.Krish, J.Fierrez, D.Ramos. R.Veldhuis, R.Wang: "Evaluation of AFIS-ranked latent fingerprint matched template", *6<sup>th</sup> Pacific-Aim Symposium on Image and Video Technology*, Mexico, 2013.
- [2] R.P.Krish, J.Fierrez, D.Ramos. R.Wang: "On the importance of rare features in AFIS-ranked latent fingerprint matched templates", *47th International Carnahan Conference on Security Technology*, Colombia, 2013.



# Speaker recognition by means of Deep Belief Networks

*Vasileios Vasilakakis, Sandro Cumani, Pietro Laface,*

Politecnico di Torino, Italy  
 [{first.lastname}@polito.it](mailto:{first.lastname}@polito.it)

## 1. Abstract

Most state-of-the-art speaker recognition systems are based on Gaussian Mixture Models (GMMs), where a speech segment is represented by a compact representation, referred to as “identity vector” (ivector for short), extracted by means of Factor Analysis. The main advantage of this representation is that the problem of intersession variability is deferred to a second stage, dealing with low-dimensional vectors rather than with the high-dimensional space of the GMM means.

In this paper, we propose to use as a pseudo-ivector extractor a Deep Belief Network (DBN) architecture, trained with the utterances of several speakers. In this approach, the DBN performs a non-linear transformation of the input features, which produces the probability that an output unit is on, given the input features. We model the distribution of the output units, given an utterance, by a reduced set of parameters that embed the speaker characteristics.

Tested on the dataset exploited for training the systems that have been used for the NIST 2012 Speaker Recognition Evaluation, this approach shows promising results.

## 2. Introduction

Many resources have been devoted in the last few years to Auto Associative Neural Networks (AANNs), Bottleneck Networks, and Deep Belief Networks (DBNs) as possible frameworks for innovative solutions to speech and speaker recognition problems. DBNs have been successfully used for speech recognition [1], rising increasing interest in the DBNs technology [2]. AANNs, trained to reconstruct the input features, or even simple Neural Networks classifiers, have been used to compress in a bottleneck layer the information given by a window including a suitably wide context. The outputs of the bottleneck units are then used as new features for training traditional classifiers such as Hidden Markov Models for speech [3][4], or as features for GMM based speaker recognition systems [5].

AANNs have also been used, without exploiting a wide input context, but still using the compression layer as a feature extractor, for training an ivector based speaker recognition system [6]. In another approach, the weights of the bottleneck layer have been adapted to the speaker of the

current utterance and then used as ivectors [7]. Although some of the mentioned techniques use complex architectures and computationally expensive optimization procedures, none of them is able to reach state-of-the-art performance in speaker recognition. On the contrary, the published results lay well behind the ones obtained by the “standard” ivector systems using a Probabilistic Linear Discriminant Analysis (PLDA) classifier [8],[9]. Preliminary experiments with DBNs have been reported in [10],[11] using i-vectors as input features for DBN based classifiers.

In contrast with these approaches, which try to estimate a huge number of parameters with a relatively small number of ivectors, our approach tries to extract a pseudo-ivector directly from the frames of the speech segment, i.e. we leave to a PLDA classifier the task of discriminating among the speakers. In particular, our approach aims at extracting a compact information, which summarize the speaker characteristic given an utterance, directly from the output units of a stack of Restricted Boltzmann Machines (RBMs). In this paper, we will refer to this stack of RBMs as a DBN. The DBN performs a non-linear transformation of the input features, and produces the probability that an output unit is active, given a wide-context of input frames. We model the distribution of each output unit, given an utterance, by a few set of parameters which embed the speaker characteristics because they average the acoustic content of the utterance.

In this work we tested several techniques to model these distributions and to extract pseudo-ivectors. We trained and tested PLDA classifiers using these pseudo-ivectors on the dataset exploited for training the systems that have been used for the NIST 2012 Speaker Recognition Evaluation [12]. Although this approach does not obtain state-of-the-art results, it is an attempt to go beyond the use of the DBN as an ivector classifier, or as a bottleneck feature extractor.

The paper is organized as follows: Section 2 briefly recalls the GMM, the ivector and the PLDA models for speaker recognition. Section 3 introduces the Restricted Boltzmann Machine (RBM) model, and illustrates our approach for pseudo-ivector extraction based on the analysis of the probability distribution of the DBN output units. Section 4 details the computation of different pseudo-ivectors from the DBN output nodes probability distributions. Section 5 is devoted to the illustration of the training and test datasets, and to the experimental results. Finally, in Section 6 we draw our conclusions.



### 3. GMM Speaker models

Our reference models in this work are the state-of-the-art speaker Gaussian Mixture Models (GMMs), which are used to estimate statistics that allow obtaining a low-dimensional representation of a speech segment, the so-called “identity vector” or ivector [13][14]. An ivector is a compact representation of a GMM supervector [15], representing both the speaker and channel characteristics of a given speech segment, which captures most of the supervector variability. The ivector representation constrains the GMM supervector  $\mathbf{s}$  to live in a single subspace according to:

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{1}$$

where  $\mathbf{m}$  is the Universal Background Model (UBM) supervector,  $\mathbf{T}$  is a low-rank rectangular matrix, of  $C \times F$  rows and  $M$  columns, and  $C$  and  $F$  are the number of GMM components and feature dimensions, respectively. The  $M$  columns of  $\mathbf{T}$  are vectors spanning the subspace including important inter and intra-speaker variability in the supervector space, and  $\mathbf{w}$  is a latent variable of size  $M$  with standard normal distribution. A Maximum-Likelihood estimate of matrix  $\mathbf{T}$  is obtained by minor modifications of the Joint Factor Analysis approach [16][17].

Ivectors, in conjunction with a PLDA classifier, allow state of the art results to be obtained. A Gaussian PLDA system, implemented according to the framework illustrated in [8], has been used in this work.

The details of the features, and datasets used for training these models are described in Section 6.

### 4. Restricted Boltzmann Machines

A Boltzmann machine is a generative Neural Network that can learn a probability distribution over its set of binary inputs. A Restricted Boltzmann Machine is a variant of a Boltzmann machine with a hidden layer  $\mathbf{h}$  and a visible layer  $\mathbf{v}$ , and without hidden-to-hidden or visible-to-visible connections, as depicted in Figure 1.

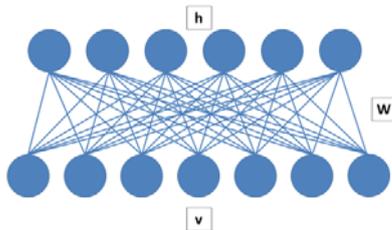


Figure 1. A Restricted Boltzmann Machine

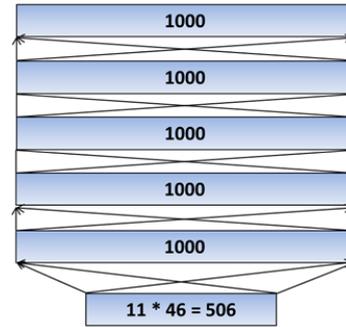


Figure 2: Topology of a 5 layer RBM. The visible layer of the first RBM takes a context of 11 frames consisting of 46 parameters. All the other layers have 1000 units.

Since the RBM has the shape of a bipartite graph, with no intra-layer connections, the hidden unit activations are mutually independent, given the visible unit activations, and vice-versa. RBMs are trained by means of the contrastive divergence technique [18], which is also able to deal with continuous input values, as the ones that have been used in our experiments.

The activation probability of a hidden unit  $j$  is given by:

$$p(h_j = 1|v) = \sigma(b_j + \sum_{i=1}^n W_{ij} v_i), \tag{2}$$

where

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{3}$$

denotes the sigmoid function.

Thus, after the network has been trained, the activation value of a unit of the first RBM hidden layer is its activation probability given the values plugged to its visible units. This single layer of binary features can be used as data for training a second RBM, and this procedure can be performed for the desired number of layers. Each layer of features captures higher-order correlations between the activities of the units of the previous layer.

The main idea in our proposal is to train a DBN consisting of a stacked set of RBMs using as input data a wide context of the frames of several hundred speakers segments, as shown in Figure 2, where the number of trained RBMs is 5, the input layer takes a context of 11 frames consisting of 46 parameters, and each hidden layer has 1000 units. The first layer is a Gaussian-Bernoulli RBM trained on acoustic features, whereas the others layers of the DBN are Bernoulli-Bernoulli RBMs. Since this DBN can be seen as a UBM with a very large number of mixture components, we can try to exploit the probability distributions from the activation probabilities of each hidden unit of the top RBM (we will refer to these units as the “output nodes”). Our assumption is that these probability distributions, and their shape, carry information about the speaker identity, because the phonetic content of the segment, for long enough utterances, is averaged. It is worth noting that the DBN is

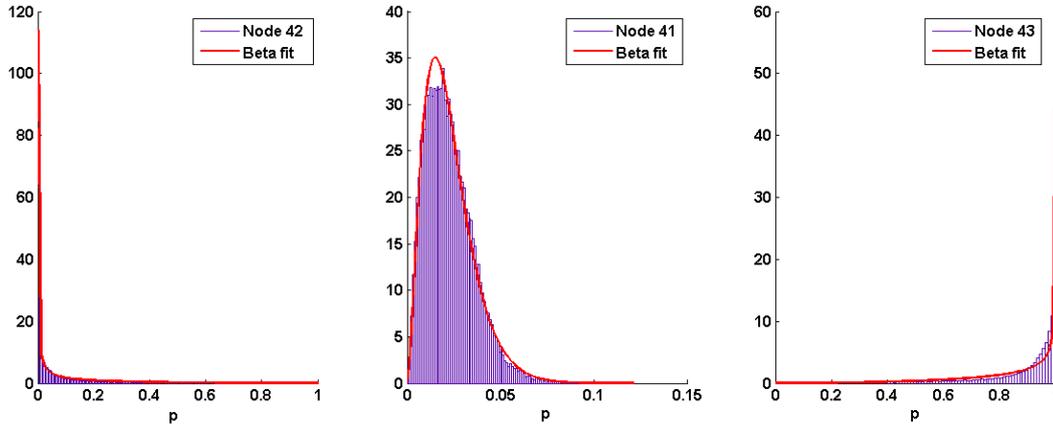


Figure 3: Distribution of the activation probability and the Beta fitting pdf for three output nodes.

not further retrained with a different objective function, leaving to the PLDA classifier the task of discriminating among the speakers.

Since the outputs of the net are highly correlated, Principal Component Analysis is performed to obtain a pseudo-ivector with low dimensions, to be used as observation sample for the standard PLDA model, which takes also care of the intersession variability.

## 5. Pseudo-ivectors

Three main sets of pseudo-ivectors have been extracted, based on the analysis of the probability distribution of the output units of the same DBN.

### 5.1. Empirical mean and variance

The simplest pseudo-ivector extraction approach computes the average value of the probability that a given output unit  $j$  is active:

$$\mu_j = \frac{1}{T} \sum_t (p_{tj}=1 | v), \quad (4)$$

where  $T$  is the number of frames of the speaker segment, without performing any decorrelation of the obtained  $\mu_j$ , i.e., using as pseudo-ivector a 1000-dimensional vector  $\boldsymbol{\mu}$ . The results obtained using these pseudo-ivectors were, as expected, not good compared with the pseudo-ivectors obtained by reducing their dimension by means of a PCA projection. Appending to  $\boldsymbol{\mu}$  the vector of the variances of each output node probability, and then performing PCA, gives far better results at the cost of increasing the dimension of the pseudo-ivectors.

### 5.2. Beta distribution fitting

Figure 3 shows the distribution of the probabilities of three output nodes, computed for a file including 28596 voice frames. Similar plots are obtained for the other output nodes, and for other files. These distributions are not at all Gaussian. This is not surprising because the output layer

produces activation probabilities, i.e., the probability that the output of node  $j$  is active, given input  $\mathbf{x}_t$ . The figure shows that the activation probability of a specific node is concentrated near 0 or 1, and that its distribution has a shape that can be better fit by a Beta distribution. In particular, given an input  $\mathbf{x}_t$ , the  $j$ -th output  $y_{t,j}$  of the network is active with a probability  $p_{t,j}$ , which depends on the input value and the network weights.

In the following, we assume that the activation probabilities  $p_{t,j}$  are realizations of a random variable  $P_j$ , and that each random variable  $P_j$  follows a Beta distribution with parameters  $\alpha_j, \beta_j$

$$P_j \sim B(\alpha_j, \beta_j),$$

and that the set of  $P_j$  are independent.

The pair of parameters of the Beta distribution  $(\alpha_j, \beta_j)$  can be estimated by maximizing the log-likelihood of the set of  $T$  observations  $\{p_{1,j}, \dots, p_{T,j}\}$  of a speech segment as:

$$\alpha_j, \beta_j = \operatorname{argmax}_{\alpha, \beta} \log L(p_{1j}, p_{2j}, \dots, p_{Nj} | \alpha, \beta). \quad (5)$$

The log-likelihood in (5) is given by:

$$\begin{aligned} \log L(p_{1j}, p_{2j}, \dots, p_{Nj} | \alpha, \beta) &= -T \log B(\alpha, \beta) \\ &+ (\alpha - 1) \sum_t \log p_{tj} \\ &+ (\beta - 1) \sum_t \log(1 - p_{tj}), \end{aligned} \quad (6)$$

which can be rewritten as:

$$\begin{aligned} \log L(p_{1j}, p_{2j}, \dots, p_{Nj} | \alpha, \beta) &= -N \log B(\alpha, \beta) + \\ &(\alpha - 1)L_p + \\ &(\beta - 1)L_m, \end{aligned} \quad (7)$$

where  $L_p$  and  $L_m$  are the sufficient statistics:

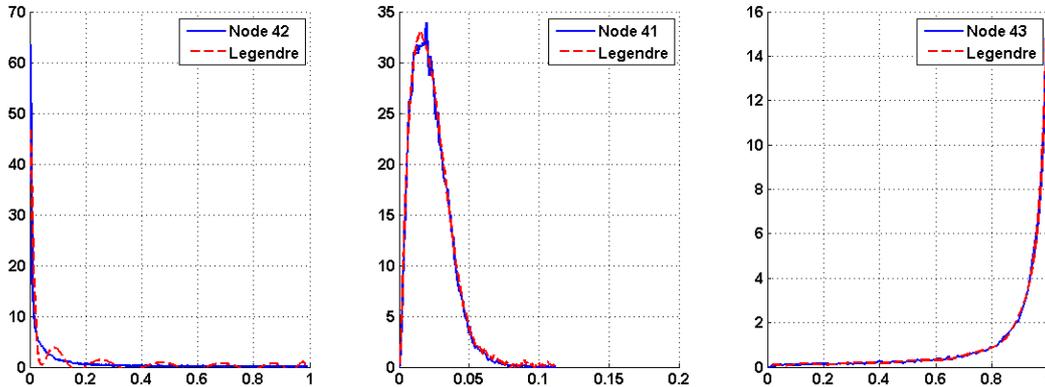


Figure 4: Distribution of the activation probability and Legendre regression for three output nodes.

$$\begin{aligned}
 L_p &= \sum_t \log p_{tj} \\
 L_m &= \sum_t \log(1 - p_{tj}).
 \end{aligned} \tag{8}$$

These sufficient statistics can be easily obtained from the network outputs. We can find a ML solution for obtaining every pair  $(\alpha_j, \beta_j)$  by maximizing the log-likelihood (7). These parameters can be estimated by means of a generic optimizer. In particular, we used the LBFG algorithm [19], a quasi-Newton optimizer, with a regularization which avoids that the mean of a Beta distribution, computed as:

$$\mu_j = \frac{\alpha_j}{\alpha_j + \beta_j}, \tag{9}$$

differs too much with respect to the empirical distribution mean. Then, three different sets of pseudo-vectors have been extracted from these statistics by performing a 400-dimensional PCA projection of three vectors consisting of:

- the concatenation of all the  $(\alpha_j, \beta_j)$  pairs,
- the vector of the mean of the Beta distributions  $\mu$ ,
- the vector of the mean of the log probability of each output unit  $L_p$  (8).

### 5.3. Legendre Polynomial fitting

As shown in Figure 3, for the output nodes 42 and 43, the Beta fitting pdf does not approximate very well actual distributions near the extreme values (0 or 1). Thus, a third pseudo-ivector extractor has been devised, based on Legendre polynomial fitting. In this approach, the distribution of the probability of each output node is approximated by taking the first  $K$  terms of a Legendre polynomial expansion estimated to fit the distribution. In our experiments we used Legendre polynomial expansions with  $K=13$  terms. Figure 4 shows the distribution of the

activation probability and the Legendre regression for the same output nodes illustrated in Figure 3. Compared to Beta fitting, Legendre regression is more accurate, although it has some difficulty with very sharp distributions, such as the distribution of output node 42. Legendre polynomial fitting requires a substantially larger number of coefficients compared to the first and second order statistics, or to the concatenation of the  $(\alpha_j, \beta_j)$  pairs.

## 6. Experimental settings and results

The features used in this work for training the models consist of 46 parameters, 19 MFCC coefficients (c1-c19), obtained from the output of a 300-3400Hz Mel Filterbank, 19 delta ( $\Delta c_0$ -  $\Delta c_{18}$ ), and 8 double-delta ( $\Delta \Delta c_0$ -  $\Delta \Delta c_7$ ). These features were computed with a frame rate of 100 observation vectors per second, and were subject to short term gaussianization [20] computed on a 3 sec sliding window applied to speech frames only.

Our reference system uses GMMs consisting of 2048 diagonal Gaussian mixtures. Gender dependent UBMs were trained using the conversations of the NIST SRE 2006, 2008 and 2010. The training set includes 737 hours of speech selected from the 21780 conversations of 1095 female speakers and 512 hours from 15726 conversations of 723 male speakers. Matrix  $\mathbf{T}$  has been obtained using the same dataset. The dimension of the ivectors has been set to 400.

We trained PLDA models with full-rank channel factors, using 200 dimensions for the speaker factors. The ivectors used for the PLDA models are L2 normalized. PLDA training was performed using utterances without any added noise. The training and test datasets for development were selected from the male SRE 2012 training data of the target models, eliminating the 10sec and the summed conversation utterances, and taking care that highly correlated segments (e.g. same interview from different microphones) were all assigned either to the training or to the test set. The development training set finally included



Table 1: Performance of the reference GMM ivector PLDA system, and of three other types of pseudo-ivectors extracted from the statistics of the output nodes of a DBN.

System	% EER	DCF08	DCF10
GMM (reference)	<b>0.45</b>	<b>18</b>	<b>87</b>
1000 means no PCA	2.91	105	237
Dim. 400 PCA projection	<b>0.81</b>	<b>40</b>	<b>190</b>
Dim. 400 PCA projection from the means of the hidden units of the third DBN layer	1.03	47	211

572 hours of speech selected from the 16850 conversations of 1095 female speakers and 391 hours from 12070 conversations of 723 male speakers. The partition of the SRE 2012 training data not used as PLDA training set was used as development test set, for a total of 8204 true and more than 15 million impostor trials.

It is worth noting that in this work we scored every male test segment against all the others, irrespective of the channel or noisy conditions defined in the NIST 2012 SRE [21]. The reported results are obtained without any score normalization

Although the distribution of each output node probability is not Gaussian, the information carried by the distribution mean is the easiest to be computed, thus Table 1 shows the results of experiments aiming at evaluating pseudo-ivectors extracted from the means only, in terms of %EER, minDCF08, and minDCF10 (x 1000). In particular, the first line of the table shows the performance of a GMM system, quite well aligned with state-of-the art systems, which is our reference. The results in the second and third line of the Table 1 make evident the importance of the decorrelation of the mean output probabilities. The performance of the mean based pseudo-ivector system, highlighted in the third row of the table, is the reference for all the other pseudo-ivector systems. The results of the last row support the hypothesis that deeper networks are able to capture more information with respect to shallow networks. Here, the performance using the output of the 5 layer DBN is better than the one obtained from the output of the third layer of the same DBN.

As far as the statistics obtained by estimating the Beta fitting distribution is concerned, the results are summarized in Table 2. Results are reported for pseudo-ivectors obtained projecting to 400 dimensions the concatenation of

Table 2: Performance of 400-dimensional pseudo-ivectors obtained using Beta distribution statistics.

PCA Dim 400	% EER	DCF08	DCF10
$\alpha + \beta$	0.91	44	181
$L_p$	0.89	45	190
$\mu(\alpha_i, \beta_j)$	0.77	39	170

all pairs  $(\alpha_j, \beta_j)$  estimated by the regularized LBFG algorithm, the average of the log probabilities  $L_p$ , and finally the values of the means obtained as a function of the pairs  $(\alpha_j, \beta_j)$ .

The first two set of parameters are similar to the ones obtained with the means only, reported in Table 1, whereas an improvement is achieved by using the means computed from the pairs  $(\alpha_j, \beta_j)$  estimated by the regularized LBFG algorithm.

Again, although the distribution of the output probabilities is not Gaussian, rough information about the shape of the each distribution can be given by its variance. Since in this case the number of parameters doubles, the dimension of the PCA projections has been increased. The first column of Table 3 shows the results obtained by using the concatenation of means and variances, and PCA projection with increasing dimensions. The performance keeps improving up to 1200-dimensional pseudo-ivectors.

Since obtaining the pairs  $(\alpha_j, \beta_j)$  is more expensive than just concatenating the means and variances, and the LBFG algorithm requires setting the regularization parameter, the pseudo-ivector based on Beta distributions were no more exploited in the remaining experiments. Table 3 reports in its second column the results for the pseudo-ivectors extracted from the vector concatenating the set of 13 Legendre coefficients fitting each output node distribution. It can be observed that the system based on Legendre polynomial coefficients uses a much larger number of parameters, but gives similar or slightly worse results with respect to system based on the concatenation of means and variances for every dimension of the pseudo-ivectors. However, as shown in the third column of the Table 3, the concatenation of all these parameters improves the performance, mostly the % EER. Compared to the means alone, this combination improves the performance by 41%, 25%, and 19% for the % EER, minDCF08, and minDCF10, respectively, but is still well behind the GMM reference system.

Table 3: Performance of three types of pseudo-ivector extractors.

PCA Dim.	means + variances (1000 + 1000)			Legendre coefficients (13000)			means + variances + Legendre coefficient (15000)		
	% EER	DCF08	DCF10	% EER	DCF08	DCF10	% EER	DCF08	DCF10
800	0.7	38	170	0.75	37	182	0.6	33	162
1000	0.68	37	160	0.72	37	176	<b>0.58</b>	<b>32</b>	<b>160</b>
1200	0.71	36	156	0.75	37	172	0.64	33	154



## 7. Conclusions

We have proposed to use a stack of Restricted Boltzmann Machines Deep Belief Network as a pseudo-vector extractor. The distribution of the output units, given an utterance, have been modeled by a reduced set of parameters that embed the speaker characteristics. The reported results are not comparable to the state-of-the-art, still we consider them promising considering that they have been obtained using a not yet mature technology. Our results are comparatively similar to the ones obtained by the systems so far proposed in the literature because the latter compare their performance only with results obtained by classifiers based on LDA and Cosine Distance scoring, which are known to perform worse than the state-of-the-art PLDA systems.

## 8. References

- [1] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep Neural Networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 82–97, 2012.
- [2] L. Deng, G. Hinton, B. Kingsbury, "New types of Deep Neural Network Learning for speech recognition and related applications: An overview", in *Proceedings ICASSP 2013*, pp. 8599–8603, 2013.
- [3] F. Grezl and P. Fousek, "Optimizing bottleneck features for LVCSR," in *Proceedings of ICASSP 2008*, pp. 4729–4732, 2008.
- [4] C. Plahl, R. Schluter, and H. Ney, "Hierarchical bottleneck features for LVCSR," in *Proceedings of Interspeech 2010*, pp. 1197–2000, 2010.
- [5] S. Yaman, J. Pelecanos, and R. Sarikaya, "Bottleneck features for speaker recognition," in *Proceedings of Odyssey 2012*, pp. 105–108, 2012.
- [6] S. Garimella and H. Hermansky, "Factor Analysis of Auto-Associative Neural Networks with application in speaker verification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 4, pp. 522–528, 2013.
- [7] S. Thomas, H. Mallidi, S. Ganapathy, and H. Hermansky, "Adaptation transforms of Auto-Associative Neural Networks as features for speaker verification," in *Proceedings of Odyssey 2012*, pp. 98–104, 2012.
- [8] N. Brummer and E. de Villiers, "The speaker partitioning problem," in *Proc. Odyssey 2010*, pp. 194–201, 2010.
- [9] P. Kenny, "Bayesian speaker verification with Heavy-Tailed priors," in *Keynote presentation, Odyssey 2010*, Available at [http://www.crim.ca/perso/patrick.kenny/kenny\\\_\\\_Odyssey2010.pdf](http://www.crim.ca/perso/patrick.kenny/kenny\_\_Odyssey2010.pdf).
- [10] T. Stafylakis, P. Kenny, M. Senoussaoui, and P. Dumouchel, "Preliminary investigation of Boltzmann Machine classifiers for speaker recognition," in *Proceedings of Odyssey 2012*, pp. 109–116, 2012.
- [11] M. Senoussaoui, N. Dehak, P. Kenny, R. Dehak, and P. Dumouchel, "First attempt at Boltzmann Machines for speaker recognition," in *Proceedings of Odyssey 2012*, pp. 1117–121, 2012.
- [12] D. Colibro, C. Vair, K. Farrell, N. Krause, G. Karvitsky, S. Cumani, P. Laface, "Nuance - Politecnico di Torino's 2012 NIST Speaker Recognition Evaluation System", in *Proc. Interspeech 2013*, pp. 1599–1603, 2013.
- [13] N. Dehak, R. Dehak, P. Kenny, N. Brummer, and P. Ouellet, "Support Vector Machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proceedings of Interspeech 2009*, pp. 1559–1562, 2009.
- [14] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
- [15] D. A. Reynolds, "Speaker identification and verification using Gaussian Mixture Speaker Models," *Speech Communications*, vol. 17, no. 1–2, pp. 91–108, August 1995.
- [16] P. Kenny, "Joint factor analysis of speaker and session variability : Theory and algorithms". Technical report CRIM-06/08-13, CRIM, 2005
- [17] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.
- [18] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, Vol. 14, pp. 1771–1800, 2002.
- [19] R. H. Byrd, P. Lu., J. Nocedal, C. Zhu, "A Limited Memory Algorithm for Bound Constrained Optimization". *SIAM J. Scientific Computation*, n.16, vol. 5, pp.1190–1208, 1995.
- [20] J. Pelecanos, and S. Sridharan, "Feature Warping for Robust Speaker Verification," in *Proc. 2001: A Speaker Odyssey*, pp. 213–218, 2001.
- [21] National Institute of Standards and Technology, "NIST speech group web," [http://www.nist.gov/itl/iad/mig/upload/NIST\\_SRE12\\_evalplan-v17-r1.pdf](http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf)



## Frequency and ridge estimation using structure tensor

Anna Mikaelyan, Josef Bigun

School of Information Science, Computer and Electrical Engineering  
Halmstad University, Sweden

{anna.mikaelyan, josef.bigun}@hh.se

Computing a reliable orientation map is a critical step in automatic fingerprint analysis and especially for analysis of fingermarks obtained at crime-scenes especially. Being the initial step of processing image information it may influence the further operations: registration, enhancement and matching. We suggest a new way of automatic frequency estimation improving the state-of-the-art results [1], [2] favourably for noisy images. We suggest using frequency to steer Structure Tensor [3] which is used to obtain refined orientation maps.

Fingermarks collected from crime scenes are usually low quality images, therefore lacking the support of automatic image analysis methods. Forensic expertise is utilized for the major part of the image analysis: registration of minutia, ridge frequency count, etc. It is desirable to support an expert by providing reliable orientation maps to ease the work. Having orientation maps estimated we can further provide suggestions for an expert (possible minutia location and orientation) to increase his/her efficiency. A structure tensor is a symmetric positive semi-definite matrix that can be utilized for building orientation maps of fingerprints. Response of the gradient filter, upon convolution with the image, is used for estimating frequency of the image. We provide mathematical descriptions to support the method. Suggested method provides maps of frequency that are continuous in the mathematical sense.

We have tested suggested frequency estimation on the NIST SD27 database to perform ridge counting. The fingerprint ridge count procedure is desirable but time consuming for an expert as it grows quadratically with the number of minutia we want to count ridges of. Also it is not fully reliable to be done automatically. The ridge count does not only depend on the frequency but also on the wave vector direction connecting two minutia points. We model a fingerprint image in the neighbourhood  $R$  around two minutiae with the planar wave  $\cos(\omega_0 r)$ . Here,  $\omega_0$  is a known constant wave vector (frequency of the image ridge flow calculated a forehead) and  $r$  is the line joining two minutia points.

$$N_L = \frac{L}{T} = \frac{L}{2\pi/(\omega_0 |\cos(\varphi_0 - \varphi_L)|)}, \quad (1)$$

where  $L$  is the distance between minutiae,  $\varphi_0 - \varphi_L$  is the

direction of the line joining two minutiae. For infinitesimally small length  $L$  and image area  $R$  and normalisation function  $q$  ( $q > 0, \int q = 1$ ) we will get the final formula:

$$Nr = \frac{L}{2\pi} \int q(\tau) \omega(\tau) |\cos(\varphi(\tau) - \varphi_L(\tau))|. \quad (2)$$

The quality metrics  $q(\tau)$  is introduced to increase the weight of confident ridge count measurements and decrease it for bad quality ridge counts. The metrics is used as well for extrapolation of ridge counts for areas where ridge information is absent.

In order to overcome noise on the forensic dataset we have applied the above-mentioned procedure iteratively. First, we estimated frequency based on the initial image, then we used the obtained frequency for tuning structure tensor. Finally, Gabor filter were used to enhance the image. Initial image is replaced with the enhanced image and the procedure is repeated unless the global average of frequency estimation converges. As a result, we obtain frequency and dense orientation maps providing enhanced image and automatic ridge count.

For the pair of fingerprint images ridge estimation is within 1 ridge for 78% of cases. For some minutiae pairs error in estimation is explained by other minutia on the line connecting the pair. Ridge count for closer minutia reduces this bias and raises the count performance to 83%. Performance of the automatic ridge count algorithm greatly depends on the precision of the minutia placement done by forensic experts (in case of bifurcation it may introduce a false ridge). In 50 randomly selected by human minutiae ridge distances 47 were within 1 ridge with automatic fingerprint ridge count (see fig 1). Remaining 3 cases of falsely estimated number of ridges are made for fingerprints with incipient ridges (see fig. 1) and fingerprint with scar (machine error).

### 1. References

- [1] Tony Lindeberg. Feature detection with automatic scale selection. *International journal of computer vision*, 30(2):79–116, 1998.
- [2] Dario Maio and Davide Maltoni. Ridge-line density estimation in digital images. In *Pattern Recognition*,

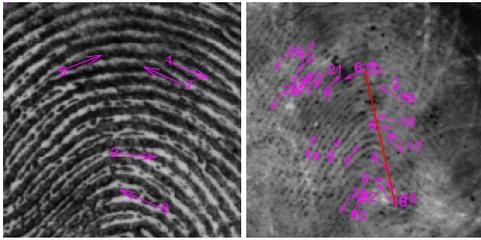


Figure 1: Examples of the fingerprints having erroneous and correct (minutiae 16-20) ridge count

1998. *Proceedings. Fourteenth International Conference on*, volume 1, pages 534–538. IEEE, 1998.

- [3] Josef Bigun, Tomas Bigun, and Kenneth Nilsson. Recognition by symmetry derivatives and the generalized structure tensor. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(12):1590–1605, 2004.