



# Forensic speaker recognition with BA-LR: calibration and evaluation on a forensically realistic database

Imen Ben-Amor<sup>1</sup>, Jean-François Bonastre<sup>2,1</sup>, David van der Vloed<sup>3</sup>

<sup>1</sup>Laboratoire Informatique d'Avignon, EA 4128, Avignon Université, France,

<sup>2</sup>Inria, Defense&Security, France,

<sup>3</sup>Netherlands Forensic Institute, The Hague, The Netherlands

## Abstract

The Likelihood Ratio (LR) is fundamental in presenting forensic speaker recognition (FSR) results. Despite its theoretical benefits, conventional LR estimation lacks transparency, impeding courtroom reliability assessment. In response, the Binary-Attribute-based Likelihood Ratio (BA-LR) framework models speech extracts based on the presence or absence of a set of speaker-specific attributes. It estimates the LR as a function of attribute-based LRs. Previous works demonstrated BA-LR's three levels of interpretability: explicit computation of attribute-based LRs, explicit contribution of these LRs to the final LR and phonetic description of the attributes, promising a fully transparent FSR solution. This work adds an examination of LR calibration using a forensically realistic database. Logistic regression is used for calibration purposes, as well as for a regularized fusion of attribute-Log LRs. Results highlight robustness and generalization ability of BA-LR, particularly in forensics.

**Index Terms:** Forensic Speaker Recognition, Interpretability, Calibration, Fusion, Likelihood ratio.

## 1. Introduction

Forensic speaker recognition (FSR) aims to automatically determine whether two voice recordings originate from the same speaker. This determination is quantified through the Likelihood Ratio (LR), commonly used to evaluate the strength of evidence [1]. The LR represents the ratio between two likelihoods corresponding to two opposing hypotheses. The first hypothesis, referred to as the prosecution hypothesis  $H_p$ , posits that the two voice samples were spoken by the same individual. Conversely, the defense hypothesis  $H_d$  assumes that each voice sample was spoken by a different person. The predominant approach for estimating LR is score-based approach [2, 3, 4], which converts a similarity score obtained from a DNN-based speaker recognition model into LR.

Even though the LR is meaningful and self-sufficient by nature [1], the lack of explainability and transparency linked to presenting a single number as the output of an automatic system is becoming a serious weakness with respect to regulatory compliance and ethical considerations [5]. This is particularly pronounced in high-risk fields such as forensics [6, 7]. To overcome this lack of explainability in FSR systems while retaining the LR paradigm, [8] recently introduced a novel interpretable and explainable framework for FSR, known as Binary-Attribute-based Likelihood Ratio (BA-LR). This framework represents a speech excerpt by a binary vector, thanks to a deep neural network model. Each coefficient  $i$  in this vector, denoted as  $BA_i$ , signifies the presence or absence of a specific speech attribute in the speech extract. For each attribute, an

Attribute-LR is estimated in a forensically interpretable manner under prosecution and defence hypotheses as inspired from forensic DNA identification [8]. This estimation process is only based on both the presence/absence of the attribute in the two speech extracts and the pre-trained behavior of this attribute, defined by three explicit parameters. The final LR is therefore calculated as the product of attribute-LRs, assuming independence between them. Augmented with a description of attributes in terms of phonetic information as proposed in [9], this approach establishes a broadly interpretable framework for FSR. With these advantages, it can become a useful tool for forensic practitioners to better understand how a FSR system work and where its outputs come from. At the end of the process, it can grandly aid the court in decision-making.

In this work, we aim to better evaluate the potential of BA-LR in the forensic context. For that, we apply the BA-LR framework on a forensically realistic database, NFI-FRIDA issued from Netherlands Forensic Institute (NFI) [10, 11]. In forensic context, using a calibration step is essential [3] to handle the mismatch between the training conditions and a real-world scenarios. Previous work [8] neglected calibration, prompting its integration into this work. The traditional approach to calibration involves employing a linear function with trainable parameters [12, 13]. Logistic Regression is frequently employed for calibration in speaker recognition [14, 12, 13, 15, 3] and we selected it for this work. This method presents an affine transformation, shifting and scaling non-calibrated scores to obtain well-calibrated LLRs.

Through the application of BA-LR framework on NFI-FRIDA dataset, we aim to achieve four main objectives:

- Propose a more interpretable version of attribute-LRs estimation, well suited to FSR.
- Assess the generalization capability, in terms of language, recording conditions and linguistic content, and the robustness of BA-LR in a forensic context.
- Propose a calibration step of the final LLR using logistic regression.
- Extend this calibration process by incorporating a fusion approach of attribute-LLRs to compute the final LLR, with the goal of enhancing both performance and calibration.

This paper is organized as follows; Section 2 provides a brief description of BA-LR framework. Section 3 proposes an improved version of attribute-LR estimation. Next, Section 4 defines the calibration approach, along with a fusion proposal of attribute-LLRs that improves performance and calibration. Subsequently, a description of the experimental protocol, including

the NFI-FRIDA dataset as well as the applied methodology is outlined in Section 5. Section 6 presents the results in terms of SR performance and calibration. Finally, some conclusions and description of future work are summarized in section 7.

## 2. BA-LR framework

This section is devoted to the description of BA-LR framework introduced in [8]. Given a comparison pair  $(X1, X2)$ , this framework involves three phases to report the final LR value, as illustrated in Figure 1. Here, we provide an overview of the three phases and propose a new version for the third phase in the next section.

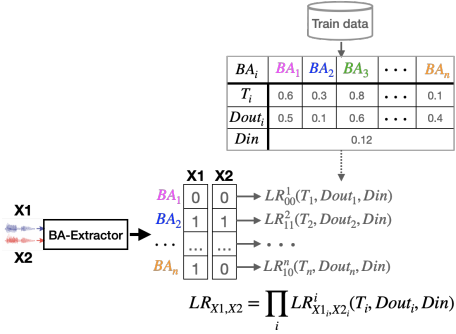


Figure 1: Overview of BA-LR framework

1. In BA-LR, the two speech extracts are represented by binary attribute vectors denoted as BA-vectors, as illustrated in Figure 1. Each dimension in the vector indicates whether an assumed attribute is present (i.e. 1) or absent (i.e. 0) in the utterance. This vector is inspired from work [16, 17, 18] on binary representations for speech and speaker recognition. The BA-vectors are extracted using a BA-extractor optimized towards generating binary representations.
2. The behavior of an attribute,  $BA_i$ , is described by three parameters, as shown in Figure 1, estimated on a train set representing the reference population [8].  $T_i$  represents the typicality of the attribute, or how frequently it occurs among speakers (i.e., its discriminative power).  $Dout_i$ , namely Drop-out, is the probability that an attribute is absent from a speech extract while present in other extracts of the same speaker.  $Din_i$ , namely Drop-in, is the probability of falsely detecting an attribute in an extract, due to noise, for example.  $Din_i$  is computed as the product of a fixed factor  $Din$ , and  $T_i$ . This represents the idea that a drop-in occurs in an attribute along with its presence frequency. We refer to [8] for a proposed estimation of  $T_i$  and  $Dout$ .
3. During a test comparison of a given pair  $(X1, X2)$ , an attribute-LR is computed separately for each attribute  $BA_i$ , using both its behavioral parameters (i.e.,  $T_i$ ,  $Dout_i$ ,  $Din_i$ ) and its value in  $X1$  and  $X2$ . Based on binary values of  $BA_i$ , four cases are considered: 00, 11, 01, 10. Equation (1) expresses the attribute-LR of a  $BA_i$  under prosecution and defence hypotheses. A detailed estimation of these cases is provided in the next section.

$$LR_{X1_i, X2_i}^i = \frac{P(X1_i, X2_i|H_p)}{P(X1_i, X2_i|H_d)} \quad (1)$$

## 3. New speech-oriented attribute-LR estimation

The initial formulation of attribute-LR in [8]<sup>1</sup> draws inspiration from the context of forensic DNA identification. However, dealing with trace and suspect samples in DNA analysis is asymmetrical since a complete DNA profile, treated as the ground truth, is readily available for the suspect. This results in distinct estimations of attribute-LR for cases 01 and 10. For a speech comparison, both samples are subject to errors or gaps, as the corresponding ground truth for a suspect is not available or does not exist. In response, we propose a more suitable adaptation for FSR by refining the LR estimation method introduced in [8]. This refined version considers the following assumptions:

- Drop-out and drop-in phenomena could occur in both,  $X1$  and  $X2$  recordings.
- The event of an occurring phenomenon in  $X1$  is independent of its happening in  $X2$ .
- $\overline{Din}$  indicates absence of drop-in, whereas  $\overline{Dout_i}$  means the absence of drop-out.
- $(X1_i, X2_i)$  is represented by an observed state in the time of comparison, and an actual state without any misleading phenomenon.

The forensic hypothetical rationale for formulating the attribute-LR in the four cases, is described in the following. Equation (2) presents these interpretations mathematically, based on Equation (1).

- $X1: (BA_i=0)$ ,  $X2: (BA_i=0)$ : The observed state is  $(0, 0)$ . **Under  $H_p$** , the prosecution considers two possibilities: either the true state is also  $(0, 0)$ , or it's  $(1, 1)$  but with drop-out on both sides resulting in  $(0, 0)$ . **Under  $H_d$** , the defense presents various scenarios. He argues that the true state could be either  $(0, 1)$  or  $(1, 0)$ , but with drop-out on one side, leading to  $(0, 0)$ . Thus, this possibility is counted twice. Moreover, it is possible that with drop-out on both sides,  $(0, 0)$  is observed from the true state  $(1, 1)$ . Additionally, if there are no drop-ins on either side,  $(0, 0)$  is obtained from the true state  $(0, 0)$ .
- $X1: (BA_i=1)$ ,  $X2: (BA_i=1)$ : **Under  $H_p$** , there is a 100% match. Alternatively, if the true state is  $(0, 0)$  but experiences drop-in on both sides, this would lead to observe  $(1, 1)$ . **Under  $H_d$** , the true state could be either  $(0, 1)$  or  $(1, 0)$ , but with a drop-in on one side or the other, resulting in  $(1, 1)$ . Additionally, it's possible that with a drop-in on both sides, we observe  $(1, 1)$  from the true state  $(0, 0)$ . Furthermore, there may be no drop-outs on either side.
- $X1: (BA_i=1|0)$ ,  $X2: (BA_i=0|1)$ : **Under  $H_p$** , the observed state is  $(1, 0)$  or  $(0, 1)$ , but they should belong to the same speaker. Thus, it is possible that the true state was  $(0, 0)$ , experiencing a drop-in on one side but not the other. Alternatively, the true state could be  $(1, 1)$ , but a drop-out occurred on only one side, not the other. **Under  $H_d$** , since both samples belong to different speakers, it's conceivable that the true state is  $(0, 1)$  or  $(1, 0)$ . There could be a drop-in on one side and not the other. A dropout on one side but not the other. A simultaneous drop-in on one side and a dropout on the other side.

<sup>1</sup>As underscored by the authors, this inspiration is tied to the estimation method rather than the identification media. Consequently, it should be approached with considerable caution.

$$\begin{cases} \frac{1 + \text{Dout}_i^2}{T_i \cdot (2 \cdot \text{Dout}_i \cdot \overline{\text{Din}} + \text{Dout}_i^2 + \overline{\text{Din}}^2)} & \text{if } (0, 0) \\ \frac{1 + (\text{Din} \cdot T_i)^2}{T_i \cdot (2 \cdot \text{Din} \cdot T_i \cdot \overline{\text{Dout}}_i + (\text{Din} \cdot T_i)^2 + \overline{\text{Dout}}_i^2)} & \text{if } (1, 1) \\ \frac{\overline{\text{Din}} \cdot \text{Din} \cdot T_i + \text{Dout}_i \cdot \overline{\text{Dout}}_i}{T_i \cdot (1 + \overline{\text{Din}} \cdot \text{Din} \cdot T_i + \text{Dout}_i \cdot \overline{\text{Dout}}_i + \text{Din} \cdot T_i \cdot \text{Dout}_i)} & \text{Otherwise} \end{cases} \quad (2)$$

The final LR is calculated as the product of the  $m$  attribute-LLRs following Equation (3) and assuming independence between attributes.

$$LR = \prod_{i=1}^n LR_{X1_i, X2_i}^i \quad (3)$$

## 4. Calibration and fusion

In this section, we firstly define the calibration approach applied on the final LLR values. Then, we introduce a fusion approach that extends this calibration to efficiently combine attribute-LLRs.

### 4.1. Global calibration

Several factors could lead to the miscalibration of LLRs in evaluation datasets, including:

- The attribute behavioral parameters used in attribute-LLR computation are estimated based on the train dataset population.
- Dutch language usage is limited in the train recordings.
- The forensic conditions of the evaluation dataset differ significantly in terms of quality and environment.
- The independence assumption is hard to fully achieve.

To address this mismatch, we employ a logistic regression model mathematically defined as follows.

Let's consider a dataset  $\{S_i, Y_i\}_{i=1}^N$ , where  $S_i = (s_1, s_2, \dots, s_n)$  is a  $n$ -dimensional variable, and the target variable  $Y_i$  is a binary variable, being 0 or 1. The general logistic regression model is as follows:

$$\log\left(\frac{P(y_i = 1|s_i)}{1 - P(y_i = 1|s_i)}\right) = \alpha + \sum_{j=1}^n \beta_j \cdot s_{ij} \quad (4)$$

Where  $P(y_i = 1|s_i)$  is the probability of  $Y$ .  $\alpha$  represents the intercept.  $\beta = (\beta_1, \dots, \beta_n)^T$  is the regression coefficient vector. The logarithmic likelihood function is therefore expressed as follows:

$$l(\beta, \alpha) = \sum_{i=1}^n [y_i \cdot (\alpha + \sum_{j=1}^n \beta_j \cdot s_{ij}) - \log(1 + \exp(\alpha + \sum_{j=1}^n \beta_j \cdot s_{ij}))] \quad (5)$$

Given a set of  $N$  comparison pairs  $(X1_i, X2_i)$  where  $i = 1 \dots N$ ,  $S_i$  is a 1-dimensional variable representing the final  $LLR_{X1_i, X2_i}$  scores, and  $Y_i$  represents the ground truth of scores being target (i.e 1) or non-target (i.e. 0). The logistic regression model for global calibration is a univariate model with  $n = 1$ . The obtained  $LLR'_{X1_i, X2_i}$  represents therefore the calibrated LLR expressed as follows:

$$LLR'_{X1_i, X2_i} = \alpha_G + \beta_G * LLR_{X1_i, X2_i} \quad (6)$$

Where  $\alpha_G$  and  $\beta_G$  are scalars.

### 4.2. Fusion of attribute-LLRs

In the BA-LLR framework, the LLR of a comparison pair is calculated by summing attribute-LLRs, assuming independence between attributes. Figure 2 illustrates the distribution of Pearson correlation values among the dimensions of BA-vectors before binarization. The relatively low correlation between attributes ensures, to some extent, the decomposition of the LLR into a direct sum of attribute-LLRs. However, any remaining correlation could still lead to an overestimation of the final LLR.

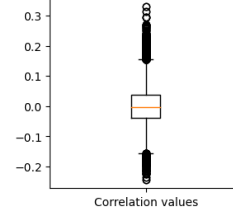


Figure 2: Pearson correlation values between attributes before binarization

To address this, we propose a fusion approach of attribute-LLRs. Logistic regression fusion is a process widely used to combine parallel sets of scores from different SR systems, yielding into more accurate and well calibrated LLRs [3, 4, 12, 14, 19, 20, 21]. Fusion could be applied between LLRs issued from an automatic FSR system with those derived from a semi-automatic FSR system [22], or between systems that use diverse signal processing and modelling techniques, or even distinct acoustic-phonetic systems where each system tackle information from a phonetic unit within the same data [4]. A regularization of logistic regression is often employed to enhance the robustness of the calibration model and mitigate the risk of overfitting [23, 13, 24]. It consists in incorporating a term into the objective function, penalizing the distance from the estimated parameters to a default set of parameters. In the following, we introduce a weighted fusion of attribute-LLRs, instead of a straightforward sum of all attribute-LLRs. This fusion incorporates a sparsity regularization to retain only relevant attributes while discarding irrelevant ones.

#### 4.2.1. Logistic regression fusion

In BA-LLR context, each attribute is considered as a sub-system that outputs a score, namely an attribute-LLR. We represent each comparison pair by  $n$ -dimensional variable  $S$ , which comprises  $n$  attribute-LLRs. As illustrated in Equation 7, logistic regression is then modeled to obtain a well-calibrated  $LLR''_{X1_i, X2_i}$ .

$$LLR''_{X1_i, X2_i} = \alpha + \sum_{j=1}^n \beta_j * LLR_{X1_i, X2_i}^j \quad (7)$$

#### 4.2.2. Regularization and selection of attributes

In order to push the logistic regression fusion to select the best subset of attribute-LLRs in terms of interpretation (but able to help also for performance), a L1 regularization term [25] is added to the log-likelihood function of the logistic model as expressed in Equation.8. This regularization encourages *sparsity* and compression [26, 27, 24], pushing the weights of some attributes to be exactly zero. The regularization parameter  $\lambda$  controls the strength of the penalty applied to the coefficients.

By increasing the value of  $\lambda$ , the penalty on large coefficients becomes stronger.

$$(\hat{\beta}, \hat{\alpha}) = \underset{\beta, \alpha}{\operatorname{argmin}} \left( \frac{-l(\beta, \alpha)}{n} + \lambda \sum_{j=1}^m |\beta_j| \right) \quad (8)$$

## 5. Experimental protocol

This section outlines the experimental protocol. Firstly some details about the NFI-FRIDA database are provided. Subsequently, the experimental setup for BA-LR application is described.

### 5.1. NFI-FRIDA description

NFI-Forensically Realistic Inter-Device Audio (NFI-FRIDA) [10, 11] is a Dutch database comprising of 302 male speakers representing a specific reference population. In the following, we describe the database in terms of recording devices and sessions.

#### 5.1.1. Recording devices

The speech was simultaneously recorded with 3 devices, namely 1, 4 and 5 [10], in each session type. These devices are chosen to reflect conditions encountered in NFI casework. The description of these devices is provided in Table 1 and is as follows:

- **Recording device 1 (d1):** a headset microphone that exhibits a high quality recording.
- **Recording device 4 (d4):** Recordings contain considerable reverberation and have a higher noise level. It represents low quality police interview recordings.
- **The intercepted telephone recordings (d5):** Extracted through a police telephone interception system that is used in actual criminal investigations. Either an iPhone 4 or a Nokia 1280 telephone was used as shown in Table 2, according to the session.

Table 1: Recording devices description

	Recording device	Session
<b>Device 1</b>	Shure WH20 HQ Headset	1,2,3,4,5,6,7,8
<b>Device 4</b>	Shure SM58 far	1,2,3,4
<b>Device 5</b>	Intercepted telephone	1,2,3,4,5,6,7,8

#### 5.1.2. Sessions

Speakers were recorded in 16 sessions, spread across two days, with a minimum interval of one week between the two days. Each day comprised eight sessions recorded in diverse locations, using different telephones, and varying in environmental noise, as detailed in Table 2. Each session lasts approximately 5 minutes, featuring telephone conversations between participants. In indoor sessions, a noisy environment included static radio noise, while outdoor sessions alternated between quiet and noisy street locations.

Table 2: Sessions description

Session	Location	Environment	Telephone
<b>1&amp;2</b>	Inside	Silent	Nokia 1280 & iPhone 4
<b>3&amp;4</b>	Inside	Noisy	Nokia 1280 & iPhone 4
<b>5&amp;6</b>	Outside	Calm	Nokia 1280 & iPhone 4
<b>7&amp;8</b>	Outside	Busy street	Nokia 1280 & iPhone 4

## 5.2. Experimental setup

This section is dedicated to the experimental setup.

### 5.2.1. BA-LR framework setting

The application of the BA-LR framework<sup>2</sup> requires the extraction of BA-vectors and the computation of attribute behavioral parameters. Both aspects are executed as follows:

- **BA-vectors extraction:** BA-vectors are derived from a ResNet speaker embedding extractor [28], where the last layer is replaced by a Softplus activation to dynamically force negative neurons to be deactivated (i.e. 0). This extractor is trained on VoxCeleb2 [29] of  $\sim 6000$  speakers. During inference, a sparse embedding is extracted and transformed into a binary vector, by simply replacing non-zero values by 1 (values  $< 10^{-4}$  are considered as 0). During this process, we remove BAs with zero activity resulting in BA-vectors of 205 BAs retained out of the 256.
- **Estimation of behavioral parameters:** The attribute behavioral parameters, namely  $T_i$ ,  $Dout_i$  and  $Din$  factor, are all estimated on VoxCeleb2.  $T_i$ ,  $Dout_i$  are computed following [8].  $Din$  factor is estimated by composing a set of comparison pairs of VoxCeleb2, calculating the LR with BA-LR, and searching for the optimal factor that yields the best performance and calibration. This search is described in Figure 3. The convergence is quite regular, with an optimum in a flat region around [0.24,0.27], giving an optimum value for  $Din$  of  $\sim 0.26$ .

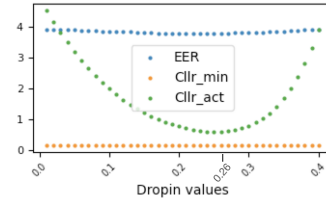


Figure 3: Estimation of the optimal value of Din

### 5.2.2. Data preparation

In this experiment, we combine for the same device, the data of two sessions sharing the same location and environment into one session. Table 3 provides details on the number of utterances and speakers used in each experiment<sup>3</sup>. All experiments use raw speech recordings under real conditions.

### 5.2.3. Protocol description

To apply the calibration and fusion approaches on NFI-FRIDA data, we establish the protocol illustrated in Figure 4. For a given device<sub>i</sub>-session<sub>j</sub>, Dev and Test sets of utterances are selected and defined with 15-fold cross-validation. In each fold, utterances are randomly selected for the Dev and Test sets, ensuring that speakers are randomly assigned to each set with no overlap between them. For Dev and Test sets, the BA-vectors are firstly extracted. Then target (tar) and non-target (non) comparison pairs are composed using all data of speakers. The BA-LR framework is thus applied on these pairs to compute the

<sup>2</sup><https://github.com/LIAvignon/BA-LR>

<sup>3</sup>Contrary to [10], no editing is applied.

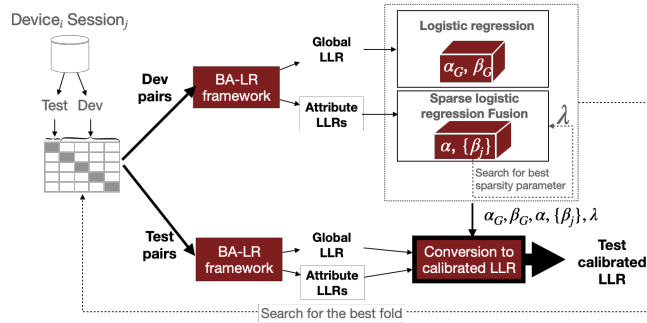


Figure 4: Description of experimental protocol using calibration and fusion approaches on BA-LR

Table 3: Experiment data description

Device-Sessions	#Utterances	#Speakers
d1-1&2	1,190	302
d1-5&6	1,186	302
d1-3&4	1,187	302
d1-7&8	1,184	302
d4-1&2	1,190	302
d4-3&4	1,183	302
d5-1&2	772	202
d5-5&6	766	203
d5-3&4	765	203
d5-7&8	768	204

Table 4: Description of Dev set composed of 150 speakers and Test sets of the best accurate fold

Device	Dev		Test	
	#speakers	#tar/non	#speakers	#tar/non
d1	150	~870/ 30K	152	~884/ 30K
d4	150	~860/ 30K	152	~890/ 30K
d5	150	~550/ 30K	~53	~550/ 30K

attribute-LLRs and the global LLR. The Dev pairs are employed for training the logistic regression models, while the Test pairs are utilized for evaluation of SR performance and calibration. Details are provided in Figure 4 and as follows:

- **Training phase:** In the global LLR calibration, the Dev global LLRs are employed to train the logistic regression model, determining the optimal shifting and scaling parameters,  $\alpha_G$  and  $\beta_G$ . In the selective fusion, the Dev attribute-LLRs are firstly standardized, then used to train sparse logistic regression model, finding the intercept  $\alpha$  and the optimal fusion coefficients of attribute-LLRs  $\{\beta_j\}_{j=0}^{m-1}$ . During the training of the latter, a grid search is conducted to identify the optimal sparsity parameter  $\lambda$  for each best fold, ensuring the best discrimination and calibration on Dev set.
- **Testing phase:** In evaluation, we use the learned parameters  $\alpha_G$ ,  $\beta_G$  and we apply global calibration on LLRs as in Equation (6). For a more calibrated LLR and more accurate fusion of attribute-LLRs, we use the learned parameters  $\alpha$  and  $\beta_j$  as in Equation (7).

These experiments finally yield, for each approach, into 4 models for d1, 2 models for d4 and 4 models for d5. Details about the Dev and Test sets are provided in Table 4.

#### 5.2.4. Baseline

[10] proposed to assess NFI-FRIDA data for a FSR task using the VOCALISE [30]. As we don't have access to this software, we use in this work the modified ResNet x-vector system presented in [9].

## 6. Calibration and speaker recognition performance

In this section, we present the experimental results in terms of calibration and speaker recognition error rates. Due to the 15-fold protocol, we have 15 sets of results for each (device,sessions pairs). As the only parameters tuned on the dev data are the calibration/fusion's ones, we present the results obtained on the best fold only, for each couple (device,sessions pairs)<sup>4</sup>.

### 6.1. BA-LR generalisation ability

Table 5 presents the EER of the application of BA-LR on Test sets, before and after fusion approach. Before applying the fusion, the overall performance of BA-LR in all experiments proves its discrimination capability and its generalization ability to the Dutch data. The superior discrimination performance observed on d1 data, in contrast to d4 and d5, can be explained by the higher quality of recordings in d1. Furthermore, d4 and d5 represent forensic conditions and telephone intercepts, respectively, which are not covered in neither the training data of the BA-extractor nor the attribute behavioral parameters used in BA-LR framework.

### 6.2. BA-LR Vs. baseline X-vector

For comparison reasons, results from the x-vector baseline are also presented in Table 5<sup>5</sup>. Before fusion calibration, BA-LR exhibits a marginal decline in performance for d1, particularly more pronounced for d4 and d5, when compared with x-vectors. This loss, though indicative, is believed to be compensated by two key factors: the dimensionality reduction of BA-vectors by  $\approx 40$  times, and the interpretability aspect inherent in the LR computation offered by BA-LR, which is highly appreciated in a forensic context.

<sup>4</sup>The complete results are available on demand, but not add useful information, except the variability due to random selection of target pairs inside a quite limited cohort.

<sup>5</sup>The results of the baseline X-vector are calculated using all the data, thus breaking the k-fold protocol. Care should therefore be taken when comparing results.

Table 5: Speaker recognition performance of BA-LR on Test set before and after fusion, for the best fold. X-vector performance is also provided for comparison.

Device-Sessions	BA-vectors			X-vectors <sup>1</sup>
	BA-LR EER (205 BAs)	BA-LR Fusion EER #BAs		Cosine EER
d1-1&2	1.0%	1.87%	132	1.02%
d1-5&6	0.96%	1.2%	139	0.85%
d1-3&4	1.22%	1.83%	149	0.74%
d1-7&8	0.43%	0.5%	159	0.28%
d4-1&2	2.07%	2.37%	119	1.59%
d4-3&4	4.27%	2.82%	144	1.43%
d5-1&2	10.05%	7.31%	101	8.16%
d5-5&6	11.2%	7.84%	128	9.53%
d5-3&4	10.72%	7.18%	127	9.7%
d5-7&8	12.61%	7.59%	124	11.1%

<sup>1</sup> These results are indicative only, as they are calculated based on all comparison pairs corresponding to each device-session.

Table 6:  $Cllr_{min/act}$  computed with BA-LR before (Non-Calibrated) and after (Calibrated) applying calibration and fusion approaches (results for the best fold)

Device-Sessions	Non-Calibrated		Calibrated			
	$Cllr_{min}$	$Cllr_{act}$	Global $Cllr_{min}$	$Cllr_{act}$	Fusion $Cllr_{min}$	$Cllr_{act}$
d1-1&2	0.04	0.60	0.04	0.08	0.07	0.10
d1-5&6	0.04	0.64	0.04	0.06	0.05	0.078
d1-3&4	0.04	0.64	0.04	0.06	0.07	0.08
d1-7&8	0.01	0.59	0.01	0.03	0.02	0.02
d4-1&2	0.08	1.71	0.08	0.10	0.10	0.10
d4-3&4	0.16	8.26	0.16	0.16	0.1	0.12
d5-1&2	0.35	8.78	0.36	0.38	0.26	0.30
d5-5&6	0.41	10.2	0.41	0.45	0.28	0.30
d5-3&4	0.35	10.0	0.35	0.38	0.26	0.27
d5-7&8	0.42	10.1	0.42	0.43	0.27	0.28

### 6.3. Calibration and fusion results

Table 6 shows the calibration performance of BA-LR scores before and after calibration in terms of  $Cllr_{min/act}$ . The LLRs obtained with BA-LR are initially miscalibrated, which is particularly noticeable for d4 and d5. After calibration, the global calibration approach effectively converts these miscalibrated LLRs into well calibrated. Interestingly the selective fusion approach improves also the overall discrimination performance of BA-LR, especially on d4 and d5, as shown in Table 5. It also outperforms the X-vector system, although this comparison is not completely fair, as the selective fusion takes advantage of in-domain data when the X-Vector system is strictly trained on out-of-domain data. Nevertheless, for d1, where the recordings are of high quality, the fusion approach shows a slight performance loss compared to the EER calculated using all BAs. This might be due to an overfitting of the model.

### 6.4. EER and Cllr Vs. Number of selected attributes

Using the selective fusion model, each experiment results in the selection of a subset of attributes, as illustrated in Table 5. On average, the number of attributes selected represents  $\sim 67\%$  of the initial set (i.e., 205 BAs), for both d4 and d5 experiments. For more insights into this selection process, Figure 5 illustrates an example of the evolution of both EER and  $Cllr_{cal}$  (i.e.,  $Cllr_{act}-Cllr_{min}$ ) with respect to increasing number of attributes, for the optimal fold. As the values of  $\lambda$  decrease, and consequently, the number of attributes increases, the EER consistently decreases until reaching a certain number of attributes, after which it starts to rise again. The  $Cllr_{cal}$  exhibits a parallel behavior to the EER, with the optimal EER aligning with the

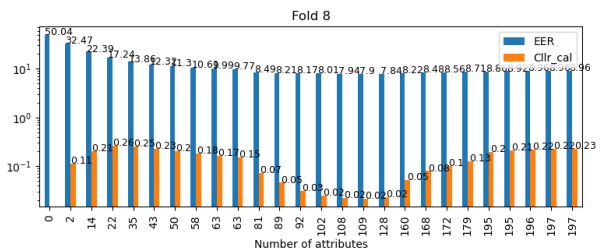


Figure 5: An example of EER and  $Cllr_{cal}$  evolution using BA-LR, along with the number of attributes, for fold 8 and d5-5&6

minimum  $Cllr_{cal}$ . This observation facilitates the identification of the optimal number of attributes that ensures both efficient discrimination and calibration performance.

## 7. Conclusion

In this article, we set out to bring the BA-LR approach closer to the needs of forensic applications and to evaluate its capabilities in this application context. To this end, we focused on a few aspects, including LR calculation and calibration, and conducted validation using the forensically realistic NFI-FRIDA database. We’ve maintained all other components of our pretrained system without any domain tuning or adaptation. We first proposed a method for calculating the attribute-LLRs that is more rational and easier to interpret forensically, compared with the original BA-LR. We then applied a calibration solution for the global LLRs using logistic regression. Additionally, we proposed an approach for merging attribute LLRs, also utilizing logistic regression, capable of calibrating the final LLR.

The overall performance obtained on NFI-FRIDA proved the generalization power of BA-LR, even though the BA-extractor model and BA-related parameters were trained on a different language and condition far from the forensic ones. Compared to baseline x-vector, an average slight increase in EER of 0.85% for all devices is observed using BA-LR, except for d4 (1.66% average EER increase). The global Logistic-Regression based calibration approach showed its abilities to produce well calibrated LLRs, even when the mismatch with the training set was particularly large. A potential limitation of BA-LR lies in assuming independence between the BAs used to compute the global LLR as a sum of the attribute-LLRs, which is not theoretically guaranteed and is challenging to achieve in practice. The fusion approach we proposed enabled us to regulate the remaining potential correlation between attributes. It offered significant performance gains in difficult scenarios, occasionally surpassing x-vectors. This was achieved thanks to its ability to completely eliminate the influence of certain BAs, particularly affected by domain mismatches. As expected, this Logistic Regression based fusion also provided a level of calibration equivalent to the global calibration.

However, while these results are indeed promising, it is crucial to approach the forensic application of SR with caution [31]. Further research is necessary for real-world deployment. Specifically, our aim is to continue experimenting with larger and more diverse databases. This will help us understand the influence of the selected training database’s significance and the extent to which our findings can be generalized to specific cases commonly encountered in forensic contexts.

## 8. Acknowledgements

This work is funded and supported by the LIAvignon AI Chair. We express our appreciation to the Netherlands Forensic Institute for providing the dataset. Also, we would like to acknowledge the suggestions from Lukas Burget. Finally, we are grateful to the reviewers for their valuable suggestions and constructive feedback.

## 9. References

- [1] Christophe Champod and Didier Meuwly, “Inference of identity in forensic speaker recognition,” *Speech Communication*, pp. 193–203, 2000.
- [2] Annabel Bolck, Haifang Ni, and Martin Lopatka, “Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic mdma comparison,” *Journal of Law, Probability and Risk*, pp. 243–266, 2015.
- [3] Daniel Ramos, Ram P. Krish, Julian Fierrez, and Didier Meuwly, *From Biometric Scores to Forensic Likelihood Ratios*, 02 2017.
- [4] Geoffrey Stewart Morrison, “Tutorial on logistic-regression calibration and fusion:converting a score to a likelihood ratio,” *Australian Journal of Forensic Sciences*, vol. 45, no. 2, pp. 173–197, June 2013.
- [5] Brandon L. Garrett and Cynthia Rudin, “Interpretable algorithmic forensics,” *Proceedings of the National Academy of Sciences*, vol. 120, no. 41, pp. e2301842120, 2023.
- [6] Adrien Bibal, Michael Lognoul, Alexandre de Streel, and Benoît Frénay, “Impact of legal requirements on explainability in machine learning,” 2020.
- [7] Th. Kirat, O. Tambou, V. Do, and A. Tsoukiàs, “Fairness and explainability in automatic decision-making systems. a challenge for computer science and law,” *EURO Journal on Decision Processes*, vol. 11, pp. 100036, 2023.
- [8] Imen Ben-Amor and Jean-François Bonastre, “BALR: Binary-attribute-based likelihood ratio estimation for forensic voice comparison,” in *International Workshop on Biometrics and Forensics (IWBF)*, 2022.
- [9] Imen Ben-Amor, Jean-François Bonastre, Benjamin O’Brien, and Pierre-Michel Bousquet, “Describing the phonetics in the underlying speech attributes for deep and interpretable speaker recognition,” in *Proc. INTERSPEECH 2023*, 2023, pp. 3207–3211.
- [10] David van der Vloed, Finnian Kelly, and Anil Alexander, “Exploring the effects of device variability on forensic speaker comparison using vocalise and nfi-frida, a forensically realistic database,” 04 2020.
- [11] David van der Vloed, “Data strategies in forensic automatic speaker comparison,” *Forensic Science International*, vol. 350, pp. 111790, 2023.
- [12] Niko Brümmer and Johan A. du Preez, “Application-independent evaluation of speaker detection,” in *Computer Speech and Language*, 2006.
- [13] Luciana Ferrer, Mahesh Kumar Nandwana, Mitchell McLaren, Diego Castan, and Aaron Lawson, “Toward fail-safe speaker recognition: Trial-based calibration with a reject option,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 140–153, 2019.
- [14] Niko Brümmer, Lukas Burget, Jan Cernocky, Ondrej Glembek, Frantisek Grezl, Martin Karafiat, David A. van Leeuwen, Pavel Matejka, Petr Schwarz, and Albert Strasheim, “Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.
- [15] Niko Brümmer and George Doddington, “Likelihood-ratio calibration using prior-weighted proper scoring rules,” 2013.
- [16] Faheem Khan and Ben Milner, “Speaker separation using visually-derived binary masks,” in *Auditory-Visual Speech Processing (AVSP) 2013*, 2013.
- [17] Li Deng, Michael L Seltzer, Dong Yu, Alex Acero, Abdel rahman Mohamed, and Geoff Hinton, “Binary coding of speech spectrograms using a deep auto-encoder,” in *Eleventh annual conference of the international speech communication association*, 2010.
- [18] Gabriel Hernández-Sierra, Jean-François Bonastre, and José Ramón Calvo de Lara, “Speaker recognition using a binary representation and specificities models,” *Iberoamerican Congress on Pattern Recognition*, 2012.
- [19] Stéphane Pigeon, Pascal Druyts, and Patrick Verlinde, “Applying logistic regression to the fusion of the nist’99 1-speaker submissions,” *Digital Signal Processing*, vol. 10, no. 1, pp. 237–248, 2000.
- [20] Andrzej Drygajlo, Michael Jessen, Stefan Gfroerer, Isolde Wagner, Jos Vermeulen, and Tuija Niemi, “Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition including guidance on the conduct of proficiency testing and collaborative exercises,” 2016.
- [21] Erica Gold, *Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters*, Ph.D. thesis, 01 2014.
- [22] Cuiling Zhang, Geoffrey Morrison, and Tharmarajah Thiruvaran, “Forensic voice comparison using chinese /iaul,” *Proceedings of the 17th International Congress of Phonetic Sciences*, 01 2011.
- [23] Luciana Ferrer, Martin Graciarena, Argyris Zymnis, and Elizabeth Shriberg, “System combination using auxiliary information for speaker verification,” in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4853–4856.
- [24] Ville Hautamäki, Kong Aik Lee, Tomi Kinnunen, Bin Ma, and Haizhou Li, “Regularized logistic regression fusion for speaker verification,” in *Proc. Interspeech 2011*, 2011, pp. 2745–2748.
- [25] Mee Young Park and Trevor Hastie, “L1-Regularization Path Algorithm for Generalized Linear Models,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 69, 08 2007.
- [26] Mengyuan Zhang and Kai Liu, “On regularized sparse logistic regression,” *ArXiv*, vol. abs/2309.05925, 2023.
- [27] Mattia Zanon, Giuliano Zambonin, Gian Antonio Susto, and Seán McLoone, “Sparse logistic regression: Comparison of regularization and bayesian implementations,” *Algorithms*, vol. 13, 2020.

- [28] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.
- [29] Joon Son Chung, Arsha Nagrani, and Andrew Senior, “VoxCeleb2: Deep Speaker Recognition,” in *Interspeech*, 2018.
- [30] Finnian Kelly, Oscar Forth, Samuel Kent, Linda Gerlach, and Anil Alexander, “Deep neural network based forensic automatic speaker recognition in vocalise using x-vectors,” in *Audio Engineering Society Conference: 2019 AES International Conference on Audio Forensics*, Jun 2019.
- [31] Jean-Francois Bonastre, Frédéric Bimbot, Louis-Jean Boe, Joseph P. Campbell, Douglas A. Reynolds, and Ivan Magrin-Chagnolleau, “Person authentication by voice: a need for caution,” in *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, 2003, pp. 33–36.