BA-LR: Binary-Attribute-based Likelihood Ratio estimation for forensic voice comparison

Imen Ben Amor & Jean-François Bonastre
Laboratoire Informatique d'Avignon, EA 4128, Avignon Université, France
{firstname.lastname}@univ-avignon.fr

***Abstract*—Likelihood ratio (LR) is a widely adopted paradigm in forensic science to represent the conclusion of a practitioner report. However, with existing estimation methods, the LR does not fully facilitate the decision making by judges and juries. With an explained decomposition of the LR value together with the case information, the judge can more easily take in hands the weight of evidence in the final decision. To that end, we propose the Binary-Attribute-based LR estimation approach (BA-LR) for Forensic Voice Comparison (FVC) where the LR is obtained as the composition of partial LRs, each dedicated to an attribute. An attribute is expressed by the presence or absence of a speaker voice characteristic. The partial LRs are directly computed following a formulation of prosecution and defence hypotheses using three probabilities that describe explicitly the behavior of the considered attribute. An implementation of our approach is evaluated on VoxCeleb1&2. The results show the effectiveness of our BA-LR approach and give hope on a better handling of LR-based FVC conclusions.**

***Index Terms*—Evidence interpretability, Forensic voice comparison, LR estimation, Strength of evidence, Uncertainty.**

I. Introduction

Forensic voice comparison (FVC) is the process of analyzing and comparing a voice trace and a suspect's voice recording under two competing hypotheses: the prosecution's hypothesis (the trace sample belongs to the suspect) and the defense's hypothesis (the trace sample belongs to someone other than the suspect). Within the forensic science community, Likelihood Ratio (LR) is commonly adopted as the "logically and legally correct" framework to present the strength of evidence to the court [1]–[3]. The LR is the statement of the support degree for the prosecution hypothesis against the defence hypothesis.

Two main approaches are used to calculate the LR called score- and feature-based methods [4]. As said in [4]–[6], the LR in feature-based methods is derived directly from density functions of the features while in score-based methods [4], [7]–[9], the LR is estimated from the similarity distance between two samples in the multidimensional feature space.

Modern machine learning approaches, as used for face and voice forensic analysis, belong mainly to score-based methods. However, the usual score-based approaches do not estimate or clearly present the relevant information's elements that contribute to the value of the LR. For instance, representing the weight of evidence by a LR equal to 50 only indicates that the evidence supports the prosecution hypothesis 50 times more than the defense hypothesis. This says nothing about the various factors that contribute to the LR estimate, their intrinsic characteristics, and the reliability of their individual estimates in this case. As a result, this lack of explicability complicates the work of the decision makers (e.g. district attorneys, prosecution or defense attorneys or jurors) fueling the still-active debate over the use of LR to effectively reflect the weight of evidence in courts [10]–[12].

Returning to the previous example, saying that the LR equal to 50 is obtained from a composition of two LRs, each dedicated to one factor and worth 2 and 25 respectively, not only indicates the same support for the prosecution hypothesis but it also provides a detailed composition of this support. Here, it is clear that this support comes mainly from the second factor ($LR = 25$). Adding to this, the intrinsic characteristics of the factors, such as their discriminant power and their expected reliability of estimation, would certainly provide more information about the final LR. Referring to the previous example, the first factor ($LR = 2$) could thus be highly discriminating but with low estimation reliability or moderately discriminating but associated with very good estimation reliability. The importance of uncertainty and quantification of estimation reliability is emphasized in [10]: "With estimation comes always a degree of uncertainty, and it should be acknowledged that this uncertainty induces a risk of jumping to the wrong conclusions at court [...] Nevertheless, having the uncertainty in mind when reporting the likelihood ratio will reduce the risk of making an erroneous conclusion at court". We believe that providing this level of explanation when a LR is proposed can make it easier for decision makers to understand the existing information in the evidence, in light of other information in the record, especially in a complex area such as voice comparison.

Toward this direction, we propose in this work a Forensic Voice Comparison (FVC) approach able to decompose the LR into factors and to explicitly describe the behavior of each factor as well as its individual role in the LR estimation. Our approach is called "Binary-Attribute-based LR" (BA-LR). It is firstly based on a binary attribute vector where each component indicates the presence/absence of a given voice characteristic, taken from a finite set of characteristics, in a given utterance. Secondly, it models an explicit representation of the behavior of an attribute by three probabilities; The first probability is

This work is funded by LIAvignon AI Chair
https//liavignon.fr/

the frequency of occurrence of the attribute in the reference population (i.e. the attribute typicality). The second is the probability of missing the attribute in a given speech sample when that attribute is present in other samples of the speaker's voice. The last probability measures the risk to mistakenly detect the presence of the attribute in a given speech sample. The final pillar of BA-LR is the direct formulation of attribute-level partial LRs using only the detection of the presence/absence of the attribute in the utterances of a pair-wise comparison and its behavioral probabilities.

In this paper, we start with a brief literature review (Section II) on the use of Bayes paradigm in FVC and existing LR estimation methods. Section III describes the proposed BA-LR framework, with a focus on the third pillar, the partial LRs computation. Next, Section IV describes a first implementation of the proposed method, developed on a x-vector baseline system using a "minimum effort" rule, and an evaluation protocol defined on the well known VoxCeleb 1&2 databases. In Section V, experimental results are presented and discussed, followed by some conclusions and description of future work in the Section VI.

II. Bayes paradigm in FVC and LR estimation

This section presents the use of Bayes' paradigm in FVC context, like introduced in [1], showing the role of LR in assessing the evidence. The two classical LR estimation approaches used in different forensic disciplines are then described, including a discussion on their relative pros and cons.

A. Bayes' paradigm in Forensic identification

In a FVC case, the judge needs to know how much likely the differences/similarities between two speech samples prove that the suspected speaker has/hasn't produced the trace sample. This is expressed as a ratio between the probability of prosecution hypothesis $\mathbf{H}_p$ given the evidence **E** divided by the probability of defence hypothesis $\mathbf{H}_d$ given the evidence **E** [1], [13]. This statement is solved using Bayes' theorem as follows:

$$\underbrace{\frac{P(H_p|E)}{P(H_d|E)}}_{\text{Posterior odds}} = \underbrace{\frac{P(E|H_p)}{P(E|H_d)}}_{\text{Likelihood ratio}} \cdot \underbrace{\frac{P(H_p)}{P(H_d)}}_{\text{Prior odds}} \tag{1}$$

The prior odds represent the view on the prosecution (H$_p$) and the defence (H$_d$) hypotheses before the scientific evidence is presented. The posterior odds are an update of the prior odds in light of knowledge of the scientific evidence. This update is done by multiplying the prior odds by the Likelihood Ratio (LR) corresponding to the evidence. The LR is the ratio of the probability that the trace and suspect speech samples have the same source (quantifying the degree of similarity between them) and the probability that they come from different sources (quantifying the degree of typicality of the sample characteristics in the relevant population). The LR estimate is the responsibility of the forensic expert and it summarizes his statement. Then, it is up to the court to evaluate its worth and decide whether to take it as an aid to their decision or not.

B. feature- and score-based LR estimation methods

Across multiple forensic disciplines, feature- and score-based methods are used to estimate the LR of forensic evidence, including DNA [14], MDMA tablets [4], fingerprints [8], handwriting [7], text [6] and voice [5].

For Score-based methods, the LR is estimated using scores, $s(x, y)$, measured as the similarity distance (e.g. cosine similarity) between the suspect sample (represented by a feature vector x) and the trace sample (represented by a feature vector y). The LR is the ratio of the two probability densities of the $s(x, y)$ scores under the prosecutor's (H$_p$) and defense's (H$_d$) assumptions respectively (2).

$$LR = \frac{f(s(x, y)|H_p)}{f(s(x, y)|H_d)} \tag{2}$$

Whereas feature-based methods estimate the LR as a ratio of two density functions of the feature vectors x and y directly under H_p and H_d (3).

$$LR = \frac{f(x, y|H_p)}{f(x, y|H_d)} \tag{3}$$

Each of those methods has advantages and shortcomings. The widespread use of score-based methods in different forensic branches is due to its ease of implementation using machine learning approaches and its robustness face to feature variations compared to feature-based methods [4], [6]–[8]. However, unlike feature-based methods that preserve the multivariate structure of feature vectors, score-based methods perform a dimensionality reduction from multivariate feature vectors to a one-dimensional score, resulting in information loss. In addition, the score-based LR estimation does not account for feature sparsity, which may make it unsuitable for correct quantification of the strength of evidence [4], [6]. Conversely, feature-based methods consider the similarity as well as the typicality of feature vectors under comparison expressed by the numerator and the denominator in (3) respectively. However, feature-based methods always consider the entire distribution of feature vectors [15] and do not separately analyze the contribution of each feature, nor the reliability estimate of that contribution.

Both methods have advantages and disadvantages, but what they have in common is that they do not provide technical explainability elements of the LR estimation that can be easily understood by the decision makers. This lack can clearly have an impact on the presentation of the strength of scientific evidence to the court [10]–[12].

III. Binary-Attribute-based LR estimation

The main idea behind this work is to address the lack of explanation of conventional LR estimation methods, highlighted in the previous section, for FVC. We propose a new LR estimation method, called Binary Attribute-based LR (BA-LR) estimation, that decomposes the LR into multiple factors and explicitly uses the behavior of each factor for the estimation of the corresponding partial LRs.

The decomposition of the LR into partial LRs expressing different contributing factors was firstly performed in forensic DNA identification [16]. A trace extracted from a crime scene and a suspect profile, both represented by a finite set of allele pairs at pre-defined loci [14], [17], are compared. For each locus, the presence/absence of alleles in both parts of the comparison is used to estimate a partial LR. The global LR is then obtained as the product of the partial LRs thanks to the independence between the loci involved. In this context, due to some sources of ambiguity [18], uncertainty by locus exists and is quantified by drop-in and drop-out probabilities of alleles [19], [20]. In speech context, speech representation dedicated to speaker recognition using large binary vectors has been proposed in [21], [22]. This binary representation offers several advantages, such as the ability to work with a high dimension while maintaining a compact representation and the ability to understand the main pieces of information considered in the final score or decision.

The remainder of this section presents the general idea of the proposed method, inspired by the research cited above, as well as some highlights of the method.

A. BA-LR method overview

The goals of the proposed BA-LR method are to decompose the LR into partial LRs and to use an explicit method to estimate the value of the partial LR. Fig. 1 presents an overview of BA-LR approach: the speech utterances are represented by Binary Attribute (BA) vectors where each attribute indicates whether a given voice characteristic is present (1) or absent (0) in the utterance; the partial LRs are computed for each attribute as the ratio of the probability of the attribute given H_p and the probability of the attribute given H_d; the global LR is computed as the product of the partial LRs per attribute.

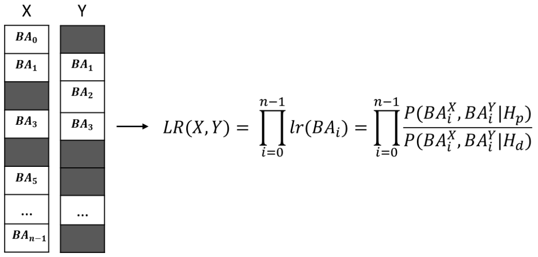
X
Y
BA_0
BA_1
BA_3
BA_5
...
BA_{n-1}
BA_1
BA_2
BA_3
...
$LR(X,Y) = \prod_{i=0}^{n-1} lr(BA_i) = \prod_{i=0}^{n-1} \frac{P(BA_i^X, BA_i^Y | H_p)}{P(BA_i^X, BA_i^Y | H_d)}$

Fig. 1. Overview of BA-LR approach

The BA-LR three main elements: the representation of speech in BA vectors, the behavioral parameters of BAs and the computation of partial LRs, are presented in the rest of this section.

B. Binary Attributes (BA)

In BA-LR approach, a speech utterance is represented by a ”BA-vector” which is a vector of n binary attributes (Fig. 1). A reference utterance BA-vector will be denoted as X = $\{BA_i^X, i \in [\![0 ; n-1]\!]\}$ (i.e. suspect recording). A trace utterance BA-vector will be denoted as Y = $\{BA_i^Y, i \in [\![0 ; n-1]\!]\}$ (i.e. trace recording).

C. The behavioral parameters of the binary attributes

The BAs behave differently depending on the considered voice characteristic and their estimation may suffer from uncertainty. These two aspects are modeled in the BA-LR approach by three parameters per BA, used for the calculation of partial LRs and for explicability reasons (abusively, we adopt for the BA parameters the same naming of the parameters as for DNA):

- Typicality of an attribute (T): defined as its presence frequency in the reference population. The more typical an attribute is (the more frequent it is in the reference population) the less discriminating it is and the lower the corresponding LR [23].
- Drop-out probability (Dout): defined as the probability that an attribute is not present in a given utterance while it is present in at least one other utterance of the speaker.
- Drop-in probability (Din): defined as the probability of encountering an artifact leading to the false detection of the presence of an attribute in a speech utterance.

Drop-out is natural in FVC because we only have speech utterances, which reflect a partial view only of the speaker's voice. Uncertainty at the attribute level in FVC comes mainly from intrinsic speech variabilities, session variability and the attribute extraction process which is also error prone.

D. Partial and global LR estimation

The estimation of the partial LRs is inspired from [18], [24]–[26] and explicitly involves the three attribute parameters. For the resolution of a partial LR, lr(BA_i), four potential cases are possible. The first case is when the attribute is absent in Y but present in X (e.g. BA_0, BA_5 and BA_{n-1} in Fig. 1). The second case is when the attribute is absent in X and Y (e.g. BA_4 in Fig. 1). The third case is when the attribute is present in X and Y (e.g. BA_3 and BA_1 in Fig. 1). The last potential case is when the attribute is present in Y but absent in X (e.g. BA_2 in Fig. 1). In the four cases just described, we consider that drop-in and drop-out could occur only in the trace; the comparison piece is thus considered as the reference of the suspect's voice. Note that in a pairwise comparison, apply the inverse process is possible to decrease the weight of this hypothesis. The elaboration of partial LRs based on the four cases is described as follows:

- *Y: (BA_i=0) , X: (BA_i=1)*: Under H_p the prosecutor says that it's just a drop-out that occurred at BA_i of the trace but actually the trace comes from the suspect. Under H_d the defence would quote all possibilities of random men. Therefore, the defense says that this could be a drop-out that occurred in the trace for this BA_i but it could also correspond to another speaker in the population for whom the BA_i was not observed so that no drop-out occurred.

- *Y: (BA$_i$=0) , X: (BA$_i$=0)*: Under H$_p$ there is 100% match at that BA$_i$. Under H$_d$, the defense accepts that the BA$_i$ doesn't exist in the trace and he adds that the BA$_i$ could also be present in the trace but a drop-out occurs.

- *Y: (BA$_i$=1) , X: (BA$_i$=1)*: Under H$_p$ there is 100% match at that BA$_i$. Under H$_d$ may be there is no drop-out in the trace and may be the BA$_i$ is absent in the trace but with a drop-in it appeared.

- *Y: (BA$_i$=1) , X: (BA$_i$=0)*: Under H$_p$, there was a drop-in which made the BA$_i$ appear in the trace. But under H$_d$, may be there was a drop-in and may be also there was no drop-in and instead the trace belongs to someone else in the population.

Equation system (4) presents the corresponding equations where $\overline{Din} = 1 - Din$ and $\overline{Dout} = 1 - Dout$. The case in which a drop-in could occur is reflected by a factor Din multiplied by T, the typicality of the attribute, which means that a drop-in has happened in BA$_i$ with probability T as done in [18] for DNA case: "A drop-in event (allele c) has happened with probability Pr(C) pc. This represents the idea that a drop-in event has happened, that the allele that has dropped in is a c allele, and that the allele and the drop-in are independent of each of other".

$$lr(\mathrm{BA}_i) = \begin{cases} \dfrac{\mathrm{Dout}}{T \cdot (\mathrm{Dout} + \overline{\mathrm{Dout}})} & \text{if } \mathrm{Y}(\mathrm{BA}_i = 0), \mathrm{X}(\mathrm{BA}_i = 1) \\[2ex] \dfrac{1}{T \cdot (\overline{\mathrm{Din}} + \mathrm{Dout})} & \text{if } \mathrm{Y}(\mathrm{BA}_i = 0), \mathrm{X}(\mathrm{BA}_i = 0) \\[2ex] \dfrac{1}{T \cdot (\overline{\mathrm{Dout}} + \mathrm{Din} \cdot T)} & \text{if } \mathrm{Y}(\mathrm{BA}_i = 1), \mathrm{X}(\mathrm{BA}_i = 1) \\[2ex] \dfrac{\mathrm{Din} \cdot T}{T \cdot (\mathrm{Din} \cdot T + \overline{\mathrm{Din}})} & \text{if } \mathrm{Y}(\mathrm{BA}_i = 1), \mathrm{X}(\mathrm{BA}_i = 0) \end{cases} \tag{4}$$

To be able to provide a global LR as the product of the partial LRs (Fig. 1), we assume that the attributes are independent, even if we know it is not always possible to fully respect this assumption.

IV. Experimental validation

This section describes the experimental validation of the BA-LR approach. To this end, we first build a preliminary version of the BA extractor as close as possible to the state-of-the-art approach, as explained in (subsection IV-A), to facilitate performance comparison. Second, we compute the BA parameters on a dataset, representing the reference population, following the recipe defined in (subsection IV-C). Then, we evaluate the performance of BA-LR approach on a different dataset using the protocol described in (subsection IV-B).

A. BA-vectors Extraction

Recent speaker recognition systems are mainly based on the notion of speaker embedding, such as x-vector [27]. In these approaches, a speech utterance is represented by a continuous feature vector of fixed length, optimized for speaker recognition tasks. In order to build our first BA extractor, we modify a baseline x-vector[1] system that extracts vectors of 256 dimensions [28] by adding a threshold layer that dynamically replaces the negative features of the x-vector by 0. To achieve this, the modified extractor is trained from scratch to classify speakers of the reference population. The extracted vectors "Relu-vectors" are transformed to BA-vectors, of 128 BAs, by replacing the non-zero values by 1. The dimensionality reduction of BA-vectors is due to deleting the null BAs for all test speech utterances, and to excluding the BAs that have been activate for less than 10% of the reference population. In the resulting BA-vectors, we need to point out that the BAs components are not perfectly respecting the independence assumption.

B. Data set and protocol

In a first attempt to validate the proposed method, we choose to work with VoxCeleb[2] [29], a general English dataset. It contains extracts from various celebrity interview videos on YouTube. The dataset is composed of two parts, VoxCeleb1 and 2, having no overlap between them in terms of speakers.

Table I describes VoxCeleb samples divided in two phases: train (VoxCeleb2) and test (VoxCeleb1) phase. VoxCeleb2, consists of around 6000 speakers that present the reference population of this experiment. In the train phase, the BA-vectors are extracted using the modified BA extractor trained for a speaker classification using VoxCeleb2. BA parameters, such as typicality and drop-out probability, are estimated on the reference population. In the test phase, only VoxCeleb1 is used. It consists of more than 1000 speakers unseen in the reference population. Pair-wise comparisons are done using 56,295 target and 56,295 non-target trials from VoxCeleb1; The target trials are combined from the first 10 utterances of each speaker. The non-target trials are combined from different-speakers utterances where each speaker has 10 utterances. A random selection of non-target trials is done to balance the number of target and non-target trials. Those trials

TABLE I
Data set and protocol description

	Train phase ***VoxCeleb2***	**Test phase** ***VoxCeleb1***
# of speakers	5,994	1,251
# of utterances	1,092,009	153,516
# of speaker couples	17,961,021	781,875
Modified Extractor	Speaker classification	
BA parameters	T + Dout	Din
# Target trials		56,295
# Non-target trials		56,295

[1] https://github.com/Chaanks/stklia

[2] https://www.robots.ox.ac.uk/ vgg/data/voxceleb/

are then used to evaluate and validate the BA-LR approach. The drop-in factor Din is set manually based on test trials.

C. Empirical estimation of attribute parameters

In this experimental protocol, we propose an estimation of the BA parameters as follows:

- Typicality probability: is calculated as the number of speaker couples sharing that attribute divided by the total number of couples in the reference population N_c (5):

$$T(\text{BA}_i) = \frac{\sum^{N_c} \text{P}_{S1} \cap \text{P}_{S2} = \{\text{BA}_i = 1\}}{N_c} \quad (5)$$

  Where P_{S1} is the set of attributes present in at least one of the S_1 speaker's utterances, similarly for P_{S2} but for speaker S_2 (Table I).

- Drop-out probability: is calculated firstly per each speaker (Dout_S) as the disappearance of the attribute in an utterance (utt_j) given that it is present in (P_S) divided by the number of speaker's utterances N_S (6):

$$\text{Dout}_S(\text{BA}_i) = \frac{\sum_{j=1}^{N_S} utt_j(\text{BA}_i=0) | P_S(\text{BA}_i=1)}{N_S} \quad (6)$$

  Then, an attribute drop-out is calculated (Dout_{Pop}) as the average of drop-out probabilities of the N speakers having that attribute in at least one utterance (Table I) (7):

$$\text{Dout}_{Pop}(\text{BA}_i) = \frac{\sum_{j=1}^{N} \text{Dout}_{S_j}(\text{BA}_i)}{N} \quad (7)$$

- Drop-in factor: is fixed and identical for all attributes. For this work, its value is determined empirically a posteriori. Note that the drop-in factor is used in conjunction with typicality, $\text{T}(\text{BA}_i)$, which induces variation by attribute.

V. Results and Discussions

Fig. 2 shows the distribution of Log(LR) scores (LLR) of target (i.e. blue) and non-target (i.e. orange) trials calculated using the proposed LR framework described in (section III-D). A discrimination between both LLR distributions is observed by a separation between target and non-target trials.

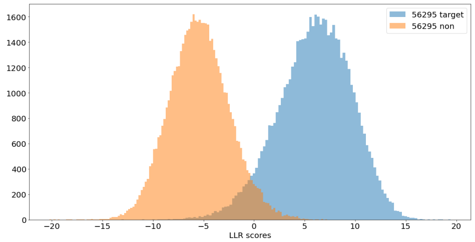

Fig. 2. Log Likelihood ratio distribution of target and non-target comparisons with a drop-in factor Din=0.56

The performance of the BA-LR method and the x-vector baseline system are presented in Table II, in terms of equal error rate (EER) and calibrated LLR (Cllr). For comparison reasons, we added the performance obtained using the Relu-vectors. The BA-LR method gives an EER of 4.79%, to be compared with 1.37% for the baseline. This loss was expected because first, contrarily to the baseline, the BA-LR takes into account only the part of information it is able to explain. Second, the BA-vector extractor was not modeled specifically for this purpose but adapted based on the x-vector baseline. At the decision level, it is interesting to note that the BA-LR approach obtained slightly better results in terms of EER than the "Relu-vector" (of the same size) for test trials, despite the binarization step and the associated loss of information. This is must be moderated by the fact that a simple cosine-based decision was used for the Relu-vector system. The results confirm the correctness of the proposed LR estimation scheme. In term of calibration, by setting a posteriori the factor, D-in, the LRs are automatically calibrated using BA-LR. For the two other systems, a further step of calibration is needed to obtain calibrated LRs.

TABLE II
Performance of BA-LR approach in terms of EER and CLLR.

	X-vectors	**Relu-vectors**	**BA-vectors**
Evaluation metric	Cosine	Cosine	BA-LR
EER	1.37	5.3	4.79
Cllr$_{min/act}$	0.056/0.819	0.19/0.95	0.180/0.186
Explainability	No	No	Yes

VI. Conclusion

The main contribution of the BA-LR method presented in this work is that it explains the origins of the LR value estimated within a pair-wise voice comparison, with elements that are easy to grasp for non-experts in the field. The BA-LR approach represents a speech utterance by a binary attribute vector expressing the presence or absence of identified voice characteristics in the speech utterances. The behavior of each attribute is explicitly described by three probabilities that are used to compute a partial LR per attribute. The BA-LR method then provides a clear and explainable view of the global LR, obtained as the product of the partial LRs.
A preliminary implementation of BA-LR approach was realized and evaluated on VoxCeleb database, based on slight transformation of a baseline x-vector system and empirical estimation of BA parameters on the training database (representing the reference population). The performance in terms of EER is very encouraging, even if a loss compared to the baseline system is noticed.

The encouraging results presented in this paper give hope that the BA-LR method can provide significant help in the field of forensic voice comparison. Some known limitations of the current implementation still need to be addressed in the near future, such as the unguaranteed respect of the independence assumption between BAs or the setting of the drop-in probability We also want to further analyze the individual contribution of partial LRs to the final evidence value, as well as the behavior of the BA-LR approach on other datasets.

Acknowledgment

The authors wish to thank Prof. D. Meuwly for his valuable advice.

References

[1] C. Champod and D. Meuwly, "Inference of identity in forensic speaker recognition," *Speech Communication*, pp. 193–203, 2000.

[2] P. Rose and G. S. Morrison, "A response to the uk position statement on forensic speaker comparison," *International Journal of Speech, Language and the Law*, p. 139–163, 2009.

[3] G. S. Morrison, E. Enzinger, D. Ramos, J. Gonzalez-Rodriguez, and A. Lozano-Diez, "Statistical models in forensic voice comparison," *Handbook of forensic statistics*, pp. 451–479, 2020.

[4] A. Bolck, H. Ni, and M. 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.

[5] J. Franco-Pedroso and J. Gonzalez-Rodriguez, "Feature-based likelihood ratios for speaker recognition from linguistically-constrained formant-based i-vectors," *Odyssey*, 2016.

[6] M. Carne and S. Ishihara, "Feature-based forensic text comparison using a poisson model for likelihood ratio estimation," *ACL Anthology*, pp. 32–42, 2020.

[7] X.-H. Chen, C. Champod, X. Yang, S.-P. Shi, Y.-W. Luo, N. Wang, Y.-C. Wang, and Q.-M. Lu, "Assessment of signature handwriting evidence via score-based likelihood ratio based on comparative measurement of relevant dynamic features," *Forensic science international*, pp. 101–110, 2018.

[8] A. J. Leegwater, D. Meuwly, M. Sjerps, P. Vergeer, and I. Alberink, "Performance study of a score-based likelihood ratio system for forensic fingermark comparison," *Journal of forensic sciences*, 2017.

[9] A. B.Hepler, C. P.Saunders, L. J.Davis, and J. Buscaglia, "Score-based likelihood ratios for handwriting evidence," *Forensic Science international*, pp. 129–140, 2007.

[10] A. Nordgaard and B. Rasmusson, "The likelihood ratio as value of evidence—more than a question of numbers," *Law, probability and Risk*, 2012.

[11] S. P. Lund and H. Iyer, "Likelihood ratio as weight of forensic evidence: A closer look," *Journal of Research of National Institute of Standards and Technology*, 2017.

[12] D. M. Daniel Ramos, Rudolf Haraksim, "Likelihood ratio data to report the validation of a forensic fingerprint evaluation method," pp. 75–92, 2017.

[13] P. Rose, "Technical forensic speaker recognition: Evaluation, types and testing of evidence," *Computer speech and language*, 2006.

[14] B. Budowle, A. M. Giusti, J. S. Waye, F. S. Baechtel, R. M. Fourney, D. E. Adams, L. A. Presley, H. A. Deadman, , and K. L. Monson, "Fixed-bin analysis for statistical evaluation of continuous distributions of allelic data from vntr loci, for use in forensic comparisons," *American Journal of Human Genetics*, pp. 841–855, 1991.

[15] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," *Interspeech*, 2019.

[16] A. Collins and N. E. morton, "Likelihood ratios for dna identification," *Medical sciences*, 1994.

[17] J. Brookfield, "The effect of population subdivision on estimates of the likelihood ratio in criminal-cases using single-locus dna probes," *Heredity*, 1992.

[18] P. Gill, J. Curran, and C. Neumann, "Interpretation of complex dna profiles using tippett plots," *Forensic science international: Genetics supplement series 1*, pp. 646–648, 2007.

[19] F. V. Nieuwerburgh, E. Goetghebeur, M. Vandewoestyne, and D. Deforce, "Impact of allelic dropout on evidential value of forensic dna profiles using rmne," *Oxford journal bioinformatics*, 2010.

[20] F. R.Bieber, john S. Buckleton, bruce Budowle, J. M.Butler, and M. D. Coble, "Evaluation of forensic dna mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion," *BMC genetics*, pp. 679–688, 2016.

[21] J.-F. Bonastre, P.-M. Bousquet, D. Matrouf, and X. Anguera, "Discriminant binary data representation for speaker recognition," *ICASSP*, 2011.

[22] G. Hernández-Sierra, J.-F. Bonastre, and J. R. C. de Lara, "Speaker recognition using a binary representation and specificities models," *Iberoamerican Congress on Pattern Recognition*, 2012.

[23] V. Hughes, A. Brereton, and E. Gold, "Reference sample size and the computation of numerical likelihood ratios using articulation rate," *York papers in Linguistics Series 2*, 2013.

[24] P. Gill, L. Gusmão, H. Haned, W. Mayr, N. Morling, W. Parson, L. Prieto, M. Prinz, H. Schneider, P. Schneider, and B. Weir, "Dna commission of the international society of forensic genetics: Recommendations on the evaluation of str typing results that may include drop-out and/or drop-in using probabilistic methods," *Forensic Sci Int Genet*, pp. 679–688, 2012.

[25] T. Tvedebrink, P. S. Eriksen, H. S. Mogensen, and N. Morling, "Estimating the probability of allelic drop-out of str alleles in forensic genetics," *Forensic Sci Int Genet*, pp. 679–688, 2012.

[26] D. J. Balding and J. Buckleton, "Interpreting low template dna profile," *National library of medicine*, 2010.

[27] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanp, "x-vectors: robust dnn embeddings for speaker recognition," *The international Conference on Acoustics, Speech, Signal Processing, ICASSP*, 2017.

[28] H. Zeinali, S. Wang, A. Silnova, P. Matejka, , and O. Plchot, "But system description to voxceleb speaker recognition challenge 2019," *Audio and Speech Processing*, 2019.

[29] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *Interspeech*, 2017.