Evolution of BA-LR and application to explainable cross-domain speaker verification*

Raphaël Duroselle
Inria Défense & Sécurité
LR2
France
raphael.duroselle@inria.fr

Yosra Jelassi
Inria Défense & Sécurité
LR2
France
yosra.jelassi@inria.fr

Jean-François Bonastre
Inria Défense & Sécurité
LR2
France
jean-francois.bonastre@inria.fr

***Abstract*—The Binary-Attribute Likelihood-Ratio (BA-LR) method has been proposed as an explainable speaker verification system focusing on forensic applications. BA-LR represents a speech utterance by a binary vector, indicating the presence or absence of speech attributes. In this work, we introduce a better founded formulation of BA-LR that can handle more naturally enrollments with multiple recordings. In addition this new formulation allows incorporating into the weight of evidence of each attribute its robustness to mismatch between enrollment and test conditions, leading to cross-domain scoring.**

***Index Terms*—speaker recognition, explainability, BA-LR, domain adaptation, NIST SRE24**

I. Introduction

The speaker verification task (SV) consists in deciding whether one test utterance was pronounced by a given speaker, represented by one or more enrollment recordings. Variability between enrollment and test conditions is a key factor that can limit system performance. This variability also adds a potential level of uncertainty to system reliability. When both conditions are known, this variability is denoted "domain mismatch" or "cross domain condition". For instance, recent NIST SRE campaigns [1] have focused on two challenging cross-domain tasks: cross-language and cross-source speaker verification.

Numerous methods have been proposed to model variability between conditions and improve performance accordingly [2]–[6]. Between them, the 4-cov PLDA [7] introduces an explicit model of the dependence between speaker embedding distributions over two domains and can be used for cross-domain scoring. These approaches are complementary to normalizations of the embeddings to limit variability between genders, languages or channels [8].

Recent SV systems [9]–[11] are based on high-dimensional embeddings, similar to x-vectors [12]. These systems use a large neural model with millions of parameters trained on large, poorly controlled databases. Thanks to their ability to exploit this considerable amount of data and parameters, they are able to discriminate between speakers and manage session variability, thus achieving cutting-edge performance. However, these systems produce a single score per trial and are unable to link this score, or parts of it, to a subset of input features, certain training examples, or certain parts of the model. They also often show a significant loss of performance when a domain mismatch occurs, i.e., when real-life conditions differ from training conditions, and this loss is difficult to predict. These two aspects lead to a lack of explainability and reliability that can significantly limit the deployment of practical solutions based on these systems. Furthermore, explainability itself is becoming increasingly necessary due to regulations such as the EU's GDPR or AI laws, and is mandatory in areas such as forensic and investigative speaker recognition [13]. Explainability is also necessary for reliability, because understanding and describing how a system works is essential for certifying its performance under specific working conditions.

Among several publications on explanability in speaker recognition [14]–[16], the BA-LR method [17], [18] has some interesting and specific features. First, BA-LR is an intrinsically explainable approach that offers three levels of explainability/interpretability: modelling, scoring and drivers. BA-LR models a speech utterance using a binary vector (BA) where a coefficient explicitely indicates the presence or absence of a given speech attribute. Each of the several hundreds attributes is shared among a group of speakers. The BA extractor is trained similarly to recent speaker embedding extractors, without requiring additional labels. It takes full advantage of x-vector state-of-the-art approaches but differs in that it produces binary embeddings and in the behavior of a coefficient: here, $BA(i) = 1$ means that the utterance contains the attribute i. The second level of explainability is scoring. BA-LR generates log-likelihood ratios (LLR) for each attribute (BA coefficient) and the final LLR is a simple combination of them. Finally, the third level is an explanation of the attributes, providing an analysis of the underlying phonetic factors. This explainability phase is done after training the system [19]. BA-LR also differs from the main part of other approaches in that it was introduced for forensic applications and has been evaluated in this context using a realistic database [20]. Recently, it was also applied to explainable spoofed speech characterization [21].

In this work, we do not address the tasks of extracting [18] or characterizing binary attributes [19] but focus on formulating the BA-LR speaker verification scoring based on statistics

This work has been partly funded by the Convention IA DGA - project VoxProfil.

of activation of binary attributes, with the aim of improving performance in cross-domain trials. Indeed, standard feature-based adaptation methods [22] are not meant to preserve the binary structure of embeddings, while model-based cross-scoring methods do not meet the explainability standards of BA-LR [5], [7]. Consequently we propose a new formulation of the BA-LR approach, which takes domain modeling into account.

In this paper, we make several contributions to the BA-LR method for speaker verification. We introduce BA-LR-v2, a new probabilistic formulation of the BA-LR model, and implement it with a Beta-Bernoulli model [23], [24]. We propose a simple cross-domain extension of BA-LR scoring and implement it with a Gaussian copula. This new BA-LR-v2 with cross-domain modeling is experimentally validated, both on a controlled experiment on VoxCeleb with simulated channel degradation and on the challenging NIST SRE24 corpus with cross-source trials. It is particularly effective at handling multiple enrollment utterances and cross-domain trials and incorporates the sensitivity of each attribute to domain mismatch into the weight of evidence.

II. BA-LR-v2 scoring

The BA-LR scoring model can be summarized by three hypotheses.

1) The test and enrollment sets of utterances x_t and x_e are represented as counts of activations a_i and non-activations n_i of N binary attributes.

$$x_t = (a_1^t, n_1^t \dots a_N^t, n_N^t) \quad x_e = (a_1^e, n_1^e \dots a_N^e, n_N^e) \tag{1}$$

2) The binary attributes are independent, which implies that the LLR can be decomposed into contributions LLR_i from each attribute. This assumption has been partially verified in [18].

$$LLR(x_t, x_e) = \sum_{i=1}^{N} LLR_i[(a_i^t, n_i^t), (a_i^e, n_i^e)] \tag{2}$$

3) For a given attribute i, LLR_i depends on statistics of activation of the attribute among a reference population.

In this work, we propose a new formulation of the LLR. We call it BA-LR-v2. We refer to [25] for an in-depth description of the original BA-LR model. In the two following sections, we work with a single binary attribute and omit the attribute index i.

A. Proposed formulation: BA-LR-v2 scoring

We assume that a speaker is represented by a latent variable p corresponding to the frequency of activation of the attribute in an utterance. The activation of the attribute for this speaker follows a Bernoulli distribution with parameter p.

We define the distribution of this probability of activation among a reference population of speakers and call its density f. According to this model, the likelihood of a sequence of observations of the attribute for a given speaker, with counts of activations (a, n), is given by:

$$L(a, n) = \int_{p=0}^{1} p^a (1-p)^n f(p) dp \tag{3}$$

The speaker verification LLR is obtained by grouping differently counts of activation of enrollment and test observations of the attribute, in a similar way to [7] for PLDA.

$$LLR = \log \frac{L(a_t + a_t, n_e + n_t)}{L(a_e, n_e) L(a_t, n_t)} \tag{4}$$

The spirit of the original BA-LR model can be found by selecting a distribution f with only two possible values of the probability of activation, interpreted as the groups of speakers with and without the attribute.

B. Implementation of BA-LR-v2 with a Beta-Bernoulli model

Similar to [23], [24], choosing a Beta distribution, with parameters α and β, simplifies the computation of the likelihood, since it is the conjugate of the Bernoulli distribution.

$$f(p|\alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)} \tag{5}$$

where $B(\alpha, \beta) = \int_{p=0}^{1} p^{\alpha-1}(1-p)^{\beta-1} dp$ is the Beta function.

$$L(a, n) = \frac{B(\alpha + a, \beta + n)}{B(\alpha, \beta)} \tag{6}$$

Consequently the LLR is given by:

$$LLR = \log \frac{B(\alpha + a_t + a_e, \beta + n_t + n_e) B(\alpha, \beta)}{B(\alpha + a_t, \beta + n_t) B(\alpha + a_e, \beta + n_e)} \tag{7}$$

III. Cross-domain scoring with BA-LR-v2

The distribution of attribute activation varies depending on the domain. This may be due to a domain mismatch with the training corpus of the attribute extractor, to noise levels that erase the attribute in an utterance, or to attributes that disappear under some conditions. For example, the fundamental frequency is outside of the telephone bandwidth for most speakers [26].

A. Cross-domain model

We propose to model the variability between probabilities of activation of an attribute under different conditions. We note 1 and 2 the two domains. A sequence of utterances from the same speaker is represented by counts of activation and non activation of the attribute for each condition.

$$x = \begin{pmatrix} a^1 & n^1 \\ a^2 & n^2 \end{pmatrix} \tag{8}$$

We now assume that each speaker is characterized by the probabilities p_1 and p_2 of activating the attribute on each condition. We assume a relationship between these two latent variables similar to the 4-cov PLDA model [7] and denote $f(p_1, p_2)$ the joint density of these probabilities among a reference population of speakers. Then, the likelihood of a

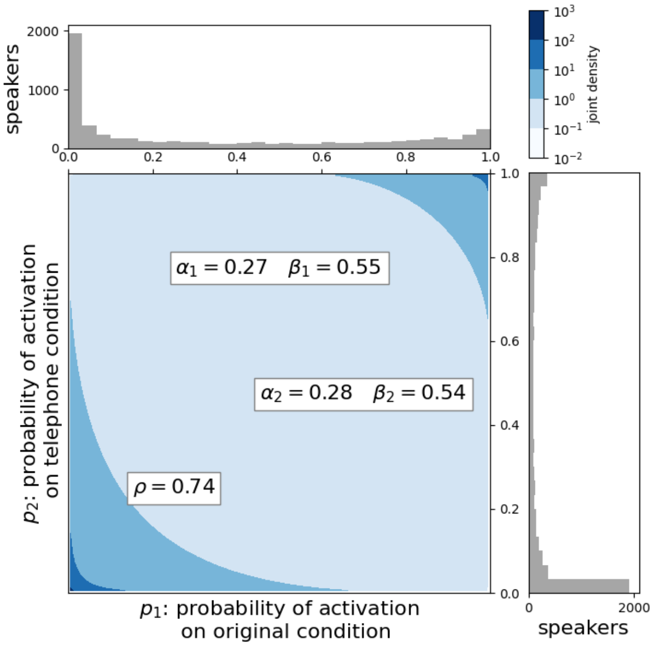

Fig. 1: Histogram of per-speaker probability of activation on each domain and estimated joint density for attribute BA386 (VoxCeleb protocol).

LLR		x_t	
		0	1
x_e	0	0.24	-0.80
	1	-0.80	0.75

mono-domain: original

LLR		x_t	
		0	1
x_e	0	0.25	-0.80
	1	-0.80	0.72

mono-domain: telephone

LLR		x_t	
		0	1
x_e	0	0.17	-0.47
	1	-0.47	0.55

cross-domain: telephone/original

Fig. 2: LLR values with cross-domain BA-LR-v2 for attribute BA386 (VoxCeleb protocol).

sequence of utterances belonging to the same speaker is given by:

$$L = \iint p_1^{a^1}(1-p_1)^{n^1} p_2^{a^2}(1-p_2)^{n^2} f(p_1,p_2)dp_1dp_2 \qquad (9)$$

B. Implementation with copula

The likelihood can be computed directly, using the empirical joint distribution of a training population of speakers. In practice, it may be convenient to consider modeling of the marginal distributions of p_1 and p_2 separately from the dependence structure between the two variables. For example, the marginal distributions may be estimated on two large corpora representative of each condition whereas we need a corpus with observations of the same speakers on both conditions to estimate the dependence between the two variables. The dependence structure between the two variables can be modeled with a copula [27]. In our context it is a bivariate cumulative density function defined on $[0,1]^2$ with uniform marginal distributions. We model the joint density by a copula of density $c(u,v)$ and the marginal distributions with cumulative density functions F_1, F_2 and densities f_1, f_2.

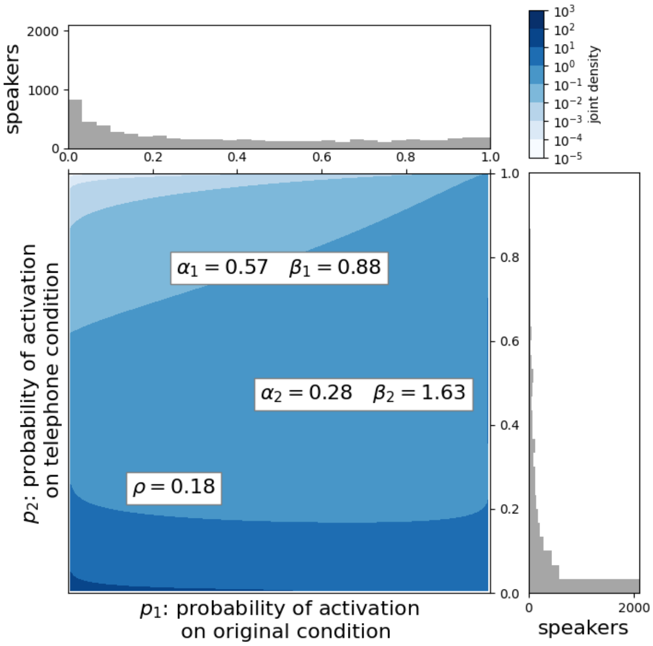

Fig. 3: Histogram of per-speaker probability of activation on each domain and estimated joint density for attribute BA220 (VoxCeleb protocol).

LLR		x_t	
		0	1
x_e	0	0.23	-0.52
	1	-0.52	0.49

mono-domain: original

LLR		x_t	
		0	1
x_e	0	0.06	-0.42
	1	-0.42	1.10

mono-domain: telephone

LLR		x_t	
		0	1
x_e	0	0.02	-0.03
	1	-0.12	0.16

cross-domain: telephone/original

Fig. 4: LLR values with cross-domain BA-LR-v2 for attribute BA220 (VoxCeleb protocol).

$$f(p_1,p_2) = c[F_1(p_1),F_2(p_2)]f_1(p_1)f_2(p_2) \qquad (10)$$

In this first implementation, we retain Beta distributions for f_1 and f_2 and select a Gaussian copula [28] for c. It has a single parameter $\rho \in]-1,1[$ which encodes the correlation between the two distributions. ρ can be estimated by maximum likelihood on a set of utterances with the same speakers on both conditions. A better model could be selected by an analysis of the actual joint distributions.

$$c(p_1',p_2'|\rho) = \frac{\mathcal{N}((\Phi^{-1}(p_1'),\Phi^{-1}(p_2'))|0,R)}{\mathcal{N}(\Phi^{-1}(p_1')|0,1)\mathcal{N}(\Phi^{-1}(p_2')|0,1)} \qquad (11)$$

where Φ^{-1} is the inverse of the cumulative densitiy function of a standard normal distribution $\mathcal{N}(.|0,1)$, and $\mathcal{N}(.|0,R)$ is a multivariate normal distribution with covariance matrix R:

$$R = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \tag{12}$$

The Gaussian copula has been recently applied to speaker recognition for score-level system fusion [29]. In [29], it models the dependence between the distributions of scores of different systems, whereas in our work it models the dependence between the distributions of a latent variable on two domains.

With the cross-domain model, likelihood values are no longer in closed form and require more computation. In a practical scenario with a large numbers of trials, the maximum number of observations of the attribute is often known in advance (one utterance for each enrollment on VoxCeleb1, three utterances maximum on NIST SRE24) and all required likelihood values can be precomputed once before inference.

Figures 1, 2, 3, 4 illustrate the impact of the latent variable joint distribution model on scoring, for attribute BA386 (high correlation between the two domains) and attribute BA220 (low correlation). Histograms of activation probabilities per speaker on each domain are shown on the x and y axes, while the estimated joint density is shown in the center, along with the corresponding values of the α_1, β_1, α_2, β_2 and ρ parameters. The impact on LLR values is shown in Figures 2 and 4. A weak correlation between the domains results in lower absolute LLR values for cross-domain trials.

IV. Experiments

To assess the validity of the proposed BA-LR scoring, we conduct two sets of experiments, first on a simulated and controlled corpus and then on the challenging NIST SRE24 corpus. For both sets of experiments, we use a baseline speaker verification system based on 256-dimensional embeddings. We extract vectors of activation of 512 attributes from these embeddings with a binary autoencoder (BAE) trained with the protocol described in [18]. We then apply the original BA-LR (speech-oriented model in [20]), as well as the proposed BA-LR-v2 scoring method.

A. VoxCeleb protocol with simulated cross-domain trials

The goal of this first experiment is to check the validity of the proposed cross-domain scoring with a controlled mismatch between the two conditions. We train the systems on VoxCeleb2 [30] and evaluate them on VoxCeleb1 [31]. The first condition consists of the original VoxCeleb utterances, corresponding to audio from video, with a sampling rate of 16kHz. For the second condition we apply a telephone bandpass filter (300-3400Hz) to the original files. Cross-condition trials are constituted of an enrollment utterance from the simulated telephone condition and an original test utterance.

The baseline system is the Wespeaker ResNet34 model with large-margin finetuning, cosine similarity scoring, and without AS-norm [11]. The BAE and BA-LR parameters are estimated on VoxCeleb2.

B. NIST SRE24 protocol

The NIST SRE24 corpus [1] focuses on challenging conditions: cross-lingual trials, among Tunisian Arabic, French and English, and cross-source trials, including conversational telephone speech (CTS) and audio from video (AfV). We evaluate the proposed cross-domain scoring method on cross-source trials. The two conditions differ not only by bandwidth but also by the level of noise, the speech content, and even the number of speakers contained in the test utterance. In addition, the NIST SRE24 corpus contains trials with multiple enrollment utterances.

The baseline speaker verification system is a ResNet101 model developed with the kiwano toolkit[1]. It is trained on the CTS superset corpus [32] (only on telephone condition), and all data is downsampled to 8 kHz. The speaker verification system is trained with the Jeffreys loss [33]. At inference, non speech regions are removed with rVAD [34]. For evaluation of the baseline system, a simple preprocessing step is applied to the embeddings before cosine similarity scoring (centering, reduction to 100 dimensions with LDA, and length normalization), whereas the original 256-dimensional embeddings are used for the extraction of binary attributes. The BAE is trained on the CTS superset corpus. BA-LR parameters are estimated on the SRE21-eval audio corpus which contains both conditions (CTS and AfV) but different languages (Cantonese, English, and Mandarin). For all systems, a per-condition logistic regression is trained on the SRE24-dev corpus. We use six conditions corresponding to the columns of Table II and defined by the enrollment and test channels (CTS or AfV) and the number of enrollment utterances (1 or 3).

V. Results

A. VoxCeleb experiments

Evaluation of the systems is reported in terms of equal error rate (EER) in Table I. We report performance on three conditions corresponding to matched conditions with original or simulated telephone utterances, and to cross-domain scoring with telephone enrollment utterances and original test utterances[2].

The original Wespeaker model achieves competitive performance on the original VoxCeleb1 [11]. It suffers from a significant but limited performance drop on the unseen simulated telephone condition. This performance drop is more moderate for cross-domain trials (twice the error rate of *original*).

The other systems correspond to different scoring methods with the same binary attributes. Using binary attributes for scoring with cosine similarity produces a performance drop,

[1] https://github.com/mrouvier/kiwano

[2] For consistency with published results, we report performance with three digits. However, a calculation of 95 % confidence intervals using a bootstrap (1000 samplings) taking into account speaker labels [35] provides the following intervals for the baseline *ResNet34-LM* model on *original/original* condition: $[0.51, 1.13]$ on *O-clean*, $[0.88, 1.09]$ on *E-clean*, and $[1.63, 1.94]$ on *H-clean*.

TABLE I: Performance on VoxCeleb1 (best BA-LR performance in bold). BA refers to binary attributes.

	System	**EER (%) on VoxCeleb1** *[enrollment condition]/[test condition]*								
		original/original			**telephone/telephone**			**telephone/original**		
		O-clean	E-clean	H-clean	O-clean	E-clean	H-clean	O-clean	E-clean	H-clean
	ResNet34-LM + cosine similarity [11]	0.814	0.933	1.695	1.803	2.123	4.031	1.462	1.881	3.488
BA	cosine similarity	1.212	1.322	2.278	2.978	3.293	5.815	2.579	3.198	5.155
	BA-LR *original*	1.255	1.443	**2.418**	3.286	3.645	6.247	2.973	3.662	5.677
	BA-LR-v2 *original*	**1.234**	**1.343**	2.445	2.755	3.007	5.674	2.345	2.792	5.027
	BA-LR-v2 *telephone*	1.404	1.523	2.756	**2.462**	**2.738**	**5.101**	2.489	2.962	5.202
	cross-domain BA-LR-v2	-	-	-	-	-	-	**2.287**	**2.613**	**4.879**

TABLE II: Performance on NIST SRE24-eval audio (best BA-LR performance in bold). BA refers to binary attributes.

	System	**SRE24 eval audio**			**EER (%) on subset of trials** *[enrollment condition]-[# enrollment utterances]/[test condition]*					
		$C_{Primary}$		EER						
		min	*act*	(%)	CTS-1/CTS	CTS-3/CTS	AfV-1/AfV	AfV-1/CTS	CTS-1/AfV	CTS-3/AfV
	ResNet101 + cosine similarity	0.698	0.830	10.47	4.73	2.48	7.29	8.06	8.77	6.85
BA	cosine similarity	0.778	0.861	12.90	5.65	2.93	8.77	10.74	11.41	9.01
	BA-LR	0.794	**0.801**	12.90	5.37	4.43	8.44	10.18	10.74	10.21
	BA-LR-v2	**0.768**	0.812	12.62	**5.32**	**2.69**	8.55	9.96	10.52	7.83
	cross-domain BA-LR-v2	0.782	0.818	**12.46**	5.58	2.81	**8.33**	**9.57**	**10.15**	**7.68**

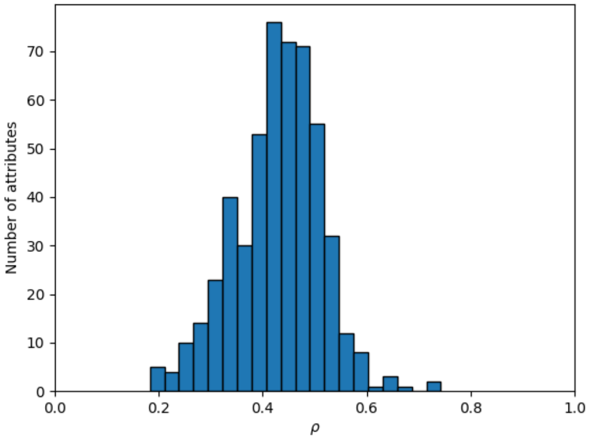

Fig. 5: Histogram of values of the Gaussian copula correlation parameter (VoxCeleb protocol) for the 512 attributes. The parameter models the dependency between probabilities of activation of an attribute on each domain.

similar on matched and unmatched conditions, for instance from 4.031% to 5.815% on H-clean $telephone/telephone$. The BA-LR model trained on $original$ data achieves worse performance than cosine similarity, especially for telephone and cross-domain conditions. The proposed BA-LR-v2 trained on $original$ data outperforms cosine similarity on the unknown telephone condition and on the cross-domain condition. BA-LR-v2 achieves its best performance when its parameters are trained on matched conditions (BA-LR-v2 $original$ and $telephone$).

Finally, the proposed cross-domain BA-LR-v2 method achieves the best performance for cross-domain trials. This improvement is statistically significant only on the *E-clean* trial list. Figure 5 represents the distribution of the estimated value of the correlation parameter ρ. Most of the attributes exhibit a weak to moderate correlation (between 0.2 and 0.6). The cross-domain BA-LR-v2 model weights the contribution of each attribute according to this correlation parameter, as illustrated by Figures 1, 2, 3, 4 for attributes BA386 and BA220.

B. NIST SRE24 experiments

Evaluation of the systems trained on the NIST SRE corpus is reported in Table II. The evaluation corpus is SRE24-eval audio, and we report the official $C_{Primary}$, $minC_{Primary}$ and EER [1]. In addition, we report the EER on specific subsets defined by the enrollment and test channels (CTS or AfV) and the number of enrollment utterances (1 or 3).

The baseline ResNet101 system achieves a $C_{Primary}$ of 0.830 on the challenging SRE24-audio-eval corpus. The binarization of the embeddings produces an important drop in performance, more pronounced in terms of EER than in terms of $C_{Primary}$. BA-LR scoring trained on a corpus containing both CTS and AfV matches cosine similarity for enrollment with a single utterance. The proposed BA-LR-v2 achieves the same overall performance as BA-LR, but with a strong improvement for trials with multiple enrollment utterances. The cross-domain BA-LR-v2 improves discrimination performance for the cross-source trials. Overall, BA-LR systems match the performance of the baseline ResNet101 system in terms of $C_{Primary}$, corresponding to low false alarm operating points, but not in terms of EER.

VI. Discussion

BA-LR is evaluated for the first time on the challenging NIST SRE campaign, demonstrating that explainable systems can be built with a moderate degradation of performance. We reach or improve the performance of cosine similarity scoring on binary attributes with BA-LR scoring, demonstrating that the drop in performance is not due to the explainable scoring mechanism but only to the binary attribute extraction step. Improving this process, particularly to ensure independence of the attributes, could help reduce the performance gap with the baseline system.

The new BA-LR-v2 scoring achieves better performance than the original BA-LR formulation for trials with multiple enrollment utterances. In addition, it leads to a very natural cross-domain scoring that relies on the modeling of the dependence between attribute activations on both conditions. From an explainability point of view, the modeling of the dependence between probabilities of activations on two conditions is crucial because it allows the weight of evidence of each attribute to be balanced with the robustness of the attribute across conditions.

Our first implementation with Gaussian copula leads to a limited performance improvement for cross-domain scoring. A better choice of the model of dependence between the conditions could make this method more efficient, for instance exploring other families of copulae.

VII. Conclusion

We introduce BA-LR-v2, a new formulation of BA-LR, that bridges the gap with classical speaker verification scoring models and improves performance for trials with multiple enrollment utterances. In addition, we model the dependence between the frequencies of activation of an attribute on two conditions, which enables cross-domain speaker verification scoring. Our implementation with a Beta-Bernoulli model and a Gaussian copula gives consistent improvements over the original BA-LR model, both in a controlled experiment on VoxCeleb1 with a simulated channel degradation and on the challenging NIST SRE24 corpus with cross-source trials. This reduces the performance gap between the explainable BA-LR system and state-of-the-art speaker verification systems, while bringing a new dimension of explainability by introducing into the weight of evidence of each attribute its robustness to condition mismatch.

Acknowledgment

This work was performed using HPC resources from GENCI–IDRIS (Grant 2025-AD011014982R1). We thank: Imen Ben-Amor for introducing the BA-LR model, Mickaël Rouvier for training the ResNet101 model, Lukáš Burget and Anna Silnova for the idea of Beta-Bernoulli scoring for BA-LR and Thibaut Vasseur for advice about copula.

References

[1] NIST. , "NIST 2024 Speaker Recognition Evaluation Plan," 2024.

[2] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1, pp. 42–54, Jan. 2000.

[3] D. Reynolds, "Channel robust speaker verification via feature mapping," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, vol. 2, Apr. 2003, pp. II–53.

[4] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[5] L. Ferrer, M. McLaren, and N. Brummer, "A speaker verification backend with robust performance across conditions," *Computer Speech & Language*, vol. 71, p. 101258, Jan. 2022.

[6] H. Zhang, L. Wang, K. A. Lee, M. Liu, J. Dang, and H. Chen, "Meta-learning for cross-channel speaker verification," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 5839–5843.

[7] P.-M. Bousquet and M. Rouvier, "Duration mismatch compensation using four-covariance model and deep neural network for speaker verification," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 1547–1551.

[8] J. Alam, S. Barahona, D. Boboš, L. Burget, S. Cumani, M. Dahmane, J. Han, M. Hlavaček, M. Kodovsky, F. Landini, L. Mošner, P. Palka, T. Pavliček, J. Peng, O. Plchot, G. P. Rajasekhar, J. Rohdin, A. Silnova, T. Stafylakis, and L. Zhang, "ABC system description for NIST SRE 2024."

[9] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, Aug. 2021.

[10] J. Huh, J. S. Chung, A. Nagrani, A. Brown, J.-W. Jung, D. Garcia-Romero, and A. Zisserman, "The VoxCeleb speaker recognition challenge: A retrospective," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3850–3866, 2024.

[11] S. Wang, Z. Chen, B. Han, H. Wang, C. Liang, B. Zhang, X. Xiang, W. Ding, J. Rohdin, A. Silnova, Y. Qian, and H. Li, "Advancing speaker embedding learning: Wespeaker toolkit for research and production," *Speech Communication*, vol. 162, p. 103104, Jul. 2024.

[12] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 5329–5333.

[13] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 95–103, Mar. 2009.

[14] T. Thebaud, G. Hernandez Sierra, S. F. Samson Juan, and M. Tahon, "A Phonetic Analysis of Speaker Verification Systems through Phoneme selection and Integrated Gradients," in *Speaker and Language Recognition Workshop - Odyssey*, Quebec, Canada, Jun. 2024. [Online]. Available: https://hal.science/hal-04578447

[15] Y. Ma, S. Wang, T. Liu, and H. Li, "Expo: Explainable phonetic trait-oriented network for speaker verification," *IEEE Signal Processing Letters*, 2025.

[16] X. Liu, J. Yamagishi, M. Sahidullah, and T. Kinnunen, "Explaining speaker and spoof embeddings via probing," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[17] I. Ben-Amor and J.-F. Bonastre, "BA-LR: Binary-Attribute-based Likelihood Ratio estimation for forensic voice comparison," in *2022 International Workshop on Biometrics and Forensics (IWBF)*. Salzburg, Austria: IEEE, Apr. 2022, pp. 1–6.

[18] I. Ben-Amor, J.-F. Bonastre, and S. Mdhaffar, "Extraction of interpretable and shared speaker-specific speech attributes through binary auto-encoder," in *Interspeech 2024*. ISCA, Sep. 2024, pp. 3230–3234.

[19] I. Ben-Amor, J.-F. Bonastre, B. O'Brien, and P.-M. Bousquet, "Describing the phonetics in the underlying speech attributes for deep and interpretable speaker recognition," in *Interspeech 2023*. ISCA, Aug. 2023, pp. 3207–3211.

[20] I. Ben-Amor, J.-F. Bonastre, and D. V. D. Vloed, "Forensic speaker recognition with BA-LR: Calibration and evaluation on a forensically realistic database," in *The Speaker and Language Recognition Workshop (Odyssey 2024)*. ISCA, Jun. 2024, pp. 9–16.

[21] J. Mishra, M. Chhibber, H.-j. Shim, and T. H. Kinnunen, "Towards explainable spoofed speech attribution and detection: A probabilistic approach for characterizing speech synthesizer components," *Submitted to Computer Speech and Language*, Feb. 2025.

[22] P.-M. Bousquet and M. Rouvier, “Adaptation strategy and clustering from scratch for new domains of speaker recognition,” in *The Speaker and Language Recognition Workshop (Odyssey 2020)*. ISCA, Nov. 2020, pp. 81–87.

[23] J. Alam, P. Kenny, G. Bhattacharya, and M. Kockmann, “Speaker verification under adverse conditions using i-vector adaptation and neural networks,” in *Interspeech 2017*. ISCA, Aug. 2017, pp. 3732–3736.

[24] O. Plchot, P. Matějka, A. Silnova, O. Novotný, M. D. Sánchez, J. Rohdin, O. Glembek, N. Brümmer, A. Swart, J. Jorrín-Prieto, P. García, L. Buera, P. Kenny, J. Alam, and G. Bhattacharya, “Analysis and Description of ABC Submission to NIST SRE 2016,” in *Interspeech 2017*. ISCA, Aug. 2017, pp. 1348–1352.

[25] I. Ben-Amor, “Deep modeling based on voice attributes for explainable speaker recognition: Application in the forensic domain,” Ph.D. dissertation, 2024.

[26] N. Gleiss, “The effect of bandwidth restriction on speech transmission quality in telephony.” in *Proc. 4th Int. Symp. on Human Factors in Telephony (Bad Wiessee, 1968)*. VDE-Verlag, 1970, pp. 1–6.

[27] R. B. Nelsen, *An Introduction to Copulas*, 2nd ed., ser. Springer Series in Statistics. New York: Springer, 2006.

[28] C. Meyer, “The Bivariate Normal Copula,” *Communications in Statistics - Theory and Methods*, vol. 42, no. 13, pp. 2402–2422, Jul. 2013.

[29] S. Cumani, “A copula-based generative score-level fusion model for speaker verification,” in *Interspeech 2025*, pp. 3723–3727.

[30] J. S. Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deep speaker recognition,” in *Interspeech 2018*. ISCA, Sep. 2018, pp. 1086–1090.

[31] A. Nagrani, J. S. Chung, and A. Zisserman, “VoxCeleb: A large-scale speaker identification dataset,” in *Interspeech 2017*. ISCA, Aug. 2017, pp. 2616–2620.

[32] O. Sadjadi, “NIST SRE CTS Superset: A large-scale dataset for telephony speaker recognition,” in *NIST*, 2021.

[33] P.-M. Bousquet and M. Rouvier, “Jeffreys Divergence-Based Regularization of Neural Network Output Distribution Applied to Speaker Recognition,” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, pp. 1–5.

[34] Z.-H. Tan, A. kr. Sarkar, and N. Dehak, “rVAD: An unsupervised segment-based robust voice activity detection method,” *Computer Speech & Language*, vol. 59, pp. 1–21, Jan. 2020.

[35] L. Ferrer and P. Riera, “Confidence Intervals for evaluation in machine learning,” https://github.com/luferrer/ConfidenceIntervals.