POLITECNICO DI TORINO
Repository ISTITUZIONALE

A copula-based generative score-level fusion model for speaker verification

Original
A copula-based generative score-level fusion model for speaker verification / Cumani, S.. - (2025), pp. 3723-3727. (Interspeech 2025 Rotterdam (The Netherlands) 17 - 21 August 2025) [10.21437/Interspeech.2025-147].

Availability:
This version is available at: 11583/3007719 since: 2026-02-17T15:17:07Z

Publisher:
ISCA - International Speech Communication Association

Published
DOI:10.21437/Interspeech.2025-147

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

10 September 2026

Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

A copula-based generative score-level fusion model for speaker verification

Sandro Cumani

Politecnico di Torino, Italy
sandro.cumani@polito.it

Abstract

In this work we present a novel generative approach for the score-level fusion of speaker verification systems. The proposed method employs a copula-based representation of the joint score distribution of multiple speaker recognizers that allows decoupling the dependency structure from the characterization of the marginal densities of the scores of different systems. This allows us to combine complex Variance-Gamma marginals with a simple Gaussian copula to obtain a characterization of the joint target and non-target score distribution that can be effectively employed for the score-level combination of multiple recognizers. Our results on NIST SRE 2019 and SITW datasets show that our approach is competitive with respect to state-of-the-art discriminative score fusion techniques, providing both accurate and well-calibrated scores, with a measured C_{llr} reduction of up to 7% relative with respect to discriminative linear fusion methods.

Index Terms: Speaker verification, score-level fusion, Gaussian copula, Variance-Gamma distribution, score calibration

1. Introduction

The combination of multiple speaker verification systems is often an effective strategy to improve the accuracy of a speaker recognizer. Typical use-cases involve the combination of systems that employ different front-ends (e.g. different speaker embedding extractors) [1, 2, 3, 4, 5, 6, 7], however even the combination of different back-end classifiers sharing a same front-end is often beneficial [8, 9]. Usually the combination is implemented at score level, where a single score is computed from a vector of scores of individual systems [10, 11, 1, 2, 3, 4, 5, 7]. Alternative approaches that operate at earlier stages of the speaker verification pipelines have been proposed in the past, including, for example, embedding-level [6, 12] fusion, where the output of different front-ends are combined and jointly classified by a single back-end. Score-level fusion, however, has become the most preminent approach, thanks to its good performance and its greater flexibility. One of the main advantages of score-level fusion is that it's independent of the system structure, i.e. it can be employed for systems that do not directly produce speaker embeddings, systems that share the same front-end, systems that may require different back-ends for different front-ends, or even multi-modal systems [1, 2]. It also requires the estimation of a small number of parameters, thus reducing the risk of overfitting. The standard approach for score-level fusion is based on a linear score combination, whose weights are estimated by means of discriminative prior-weighted logistic regression [10, 11, 13]. The approach can be seen as an extension of linear logistic regression score calibration [11, 14], and indeed tipically produces scores that are well calibrated [10]. Recently, generative calibration models have been introduced as an alternative to discriminative calibration [15, 16, 17, 18, 19, 20, 21]. These approaches represent the target and non-target score distribution of individual systems in terms of parametric densities, whose parameters are, in some cases, expression of the characteristics of the classification back-end, of the back-end training data and of the calibration population [15, 16]. These methods have proven to be effective, outperforming in several cases linear discriminative calibration models, also thanks to their ability to estimate effective non-linear calibration transformations with a limited number of parameters [15, 19]. In this work we propose an extension of generative calibration suited for score-level fusion of multiple systems. Our approach is based on modeling the joint target and non-target score distribution using a copula-based representation [22, 23] that allows us to decouple modeling the relationships between the different systems from the model of the marginal score distribution of each individual system. This in turn allows us to employ accurate but complex marginal models such as the Variance-Gamma (VΓ) approach of [15] and simple but effective models for the characterization of the scores dependency structure. The joint distribution parameters can be efficiently estimated using inference function for margins [23] approach, allowing us to incrementally estimate the joint model starting from the estimates of generative *calibration* models for the individual systems. As shown in Section 5, the resulting model provides a characterization of the joint target and non target score distributions that allows for effective score-level fusion, as confirmed by our experiments on NIST SRE 2019 [24] and SITW [25] datasets.

The paper is organized as follows. Section 2 briefly recalls generative calibration models. Section 3 introduces the copula-based framework. Section 4 presents our proposed Variance-Gamma Gaussian-copula (VΓ-GC) fusion model. Experimental results are provided in Section 5, and conclusions are given in Section 6.

2. Generative score models

Generative models have recently gained attention as an alternative to dicriminative methods to calibrate speaker verification systems [15, 16, 19, 20, 21]. Generative approaches are based on the characterization of the distribution of same-speaker (target) and different-speaker (non-target) scores, either through an explicit parametric model [15, 19] or through an implicit parametrization that describes well calibrated scores and the form of the calibration transformation [21, 17]. In both cases, given a score s, a well-calibrated score $s_{cal}(s)$ can be computed as the Log-Likelihood Ratio (LLR) between the same-speaker

10.21437/Interspeech.2025-147

$\mathfrak{S}$ and different-speaker $\mathfrak{D}$ hypotheses

$$s_{cal}(s) = \log \frac{f_{S|\mathfrak{S}}(s|\boldsymbol{\theta}_{\mathfrak{S}})}{f_{S|\mathfrak{D}}(s|\boldsymbol{\theta}_{\mathfrak{D}})} , \tag{1}$$

where $f_{S|\mathfrak{S}}(s|\boldsymbol{\theta}_{\mathfrak{S}})$ and $f_{S|\mathfrak{D}}(s|\boldsymbol{\theta}_{\mathfrak{D}})$ are the target and non-target score distribution densities, respectively, with parameters $\boldsymbol{\theta}_{\mathfrak{S}}$ and $\boldsymbol{\theta}_{\mathfrak{D}}$. In [21], the authors propose a generative Constrained Maximum Likelihood Gaussian (CMLG) model where $f_{S|\mathfrak{S}}$ and $f_{S|\mathfrak{D}}$ are Gaussian densities with tied variance:

$$f_{S|\mathfrak{S}}(s) = \mathcal{N}(s|\mu_{\mathfrak{S}}, v) , \quad f_{S|\mathfrak{D}}(s) = \mathcal{N}(s|\mu_{\mathfrak{D}}, v) . \tag{2}$$

The parameters $\mu_{\mathfrak{S}}$, $\mu_{\mathfrak{D}}$ and v can be estimated by Maximum Likelihood (ML) over a calibration set. Unsupervised variations of CMLG have been presented in [20], and generative approaches were further extended in [19], where the authors consider different parametric distribution families. More recently, an analysis of the theoretical score distribution of Gaussian-distributed speaker embeddings has been introduced in [17, 18], and further refined in [15, 16], where the authors have analyzed the effects of embedding population mismatch on the target and non-target score distribution of Probabilistic Linear Discriminant Analysis (PLDA) and PLDA-derived classifiers. These works have shown that properly tied Variance-Gamma (VΓ) densities provide powerful and accurate score models for several practical use-cases. In this work we employ the VΓ model of [15, eq. (42)] to characterize the marginal densities of individual speaker verification systems. The model describes the target and non-target scores in terms of Variance-Gamma [26] densities

$$\begin{aligned} f_{S|\mathfrak{S}}(s) &= f_{\mathrm{V}\Gamma}\left(s|\lambda, \mu_{\mathfrak{S}}, \frac{\alpha_{\mathfrak{S}}}{a_{\mathfrak{S}}}, \frac{\beta_{\mathfrak{S}}}{a_{\mathfrak{S}}}\right) , \\ f_{S|\mathfrak{D}}(s) &= f_{\mathrm{V}\Gamma}\left(s|\lambda, \mu_{\mathfrak{D}}, \alpha_{\mathfrak{D}}, \beta_{\mathfrak{D}}\right) , \end{aligned} \tag{3}$$

where $f_{\mathrm{V}\Gamma}$ is the Variance-Gamma density

$$f_{\mathrm{V}\Gamma}(s|\lambda, \alpha, \beta, \mu) = \frac{\gamma^{2\lambda}|x-\mu|^{\lambda-\frac{1}{2}} K_{\lambda-\frac{1}{2}}(\alpha|x-\mu|)}{\sqrt{\pi}\Gamma(\lambda)(2\alpha)^{\lambda-\frac{1}{2}}} e^{\beta(x-\mu)}. \tag{4}$$

The model parameters $\boldsymbol{\theta}_{\mathfrak{S}} = (\lambda, \mu_{\mathfrak{S}}, \alpha_{\mathfrak{S}}, \beta_{\mathfrak{S}}, a_{\mathfrak{S}})$ and $\boldsymbol{\theta}_{\mathfrak{D}} = (\lambda, \mu_{\mathfrak{D}}, \alpha_{\mathfrak{D}}, \beta_{\mathfrak{D}})$ are tied, and depend on a set of "effective" variance parameters $b_{\mathcal{M}}, w_{\mathcal{M}}, b_{\mathcal{C}}, w_{\mathcal{C}}$ and free parameters $\lambda, \mu_{\mathfrak{D}}, \mu_{\mathfrak{S}}, a_{\mathfrak{S}}$ as

$$\begin{aligned} t_{\mathcal{M}} &= b_{\mathcal{M}} + w_{\mathcal{M}} , & t_{\mathcal{C}} &= b_{\mathcal{C}} + w_{\mathcal{C}} , \\ \boldsymbol{\Sigma}_{\mathcal{M},\mathfrak{S}} &= \begin{bmatrix} t_{\mathcal{M}} & b_{\mathcal{M}} \\ b_{\mathcal{M}} & t_{\mathcal{M}} \end{bmatrix} , & \boldsymbol{\Sigma}_{\mathcal{M},\mathfrak{D}} &= \begin{bmatrix} t_{\mathcal{M}} & 0 \\ 0 & t_{\mathcal{M}} \end{bmatrix} , \\ \mathbf{A} &= \boldsymbol{\Sigma}_{\mathcal{M},\mathfrak{D}}^{-1} - \boldsymbol{\Sigma}_{\mathcal{M},\mathfrak{S}}^{-1} , \\ \boldsymbol{\Sigma}_{\mathfrak{S}} &= \begin{bmatrix} t_{\mathcal{C}} & b_{\mathcal{C}} \\ b_{\mathcal{C}} & t_{\mathcal{T}} \end{bmatrix} , & \boldsymbol{\Sigma}_{\mathfrak{D}} &= \begin{bmatrix} t_{\mathcal{C}} & 0 \\ 0 & t_{\mathcal{T}} \end{bmatrix} , \\ \beta_{\mathfrak{S}} &= -\frac{1}{2}\frac{\mathrm{Tr}(\mathbf{A}\boldsymbol{\Sigma}_{\mathfrak{S}})}{\det(\mathbf{A}\boldsymbol{\Sigma}_{\mathfrak{S}})} , & \beta_{\mathfrak{D}} &= -\frac{1}{2}\frac{\mathrm{Tr}(\mathbf{A}\boldsymbol{\Sigma}_{\mathfrak{D}})}{\det(\mathbf{A}\boldsymbol{\Sigma}_{\mathfrak{D}})} , \\ \gamma_{\mathfrak{S}}^2 &= -\frac{1}{\det(\mathbf{A}\boldsymbol{\Sigma}_{\mathfrak{S}})} , & \gamma_{\mathfrak{D}}^2 &= -\frac{1}{\det(\mathbf{A}\boldsymbol{\Sigma}_{\mathfrak{D}})} , \\ \alpha_{\mathfrak{S}}^2 &= \gamma_{\mathfrak{S}}^2 + \beta_{\mathfrak{S}}^2 , & \alpha_{\mathfrak{D}}^2 &= \gamma_{\mathfrak{D}}^2 + \beta_{\mathfrak{D}}^2 . \end{aligned} \tag{5}$$

3. Joint score models

A straightforward extension of generative calibration to score-level fusion consists in replacing univariate densities with multivariate models that characterize the joint distribution of the target and non-target scores. Let $\mathbf{S} = (S_1, \ldots, S_d)$ be a Random Vector that represents the scores of d speaker verification systems, and let $f_{\mathbf{S}|\mathfrak{S}}(\mathbf{s}|\boldsymbol{\theta}_{\mathfrak{S}})$ and $f_{\mathbf{S}|\mathfrak{D}}(\mathbf{s}|\boldsymbol{\theta}_{\mathfrak{D}})$ denote joint conditional densities for $\mathbf{S}$, with parameters $\boldsymbol{\theta}_{\mathfrak{S}}$ and $\boldsymbol{\theta}_{\mathfrak{D}}$, respectively. A fusion score s_f for the score vector $\mathbf{s}$ can be computed as the LLR

$$s_f(\mathbf{s}) = \log \frac{f_{\mathbf{S}|\mathfrak{S}}(\mathbf{s}|\boldsymbol{\theta}_{\mathfrak{S}})}{f_{\mathbf{S}|\mathfrak{D}}(\mathbf{s}|\boldsymbol{\theta}_{\mathfrak{D}})} . \tag{6}$$

To model the target and non-target conditional densities we start considering a simple multivariate Gaussian model

$$f_{\mathbf{S}|\mathfrak{S}}(\mathbf{s}) = \mathcal{N}(\mathbf{s}|\boldsymbol{\mu}_{\mathfrak{S}}, \boldsymbol{\Sigma}_{\mathfrak{S}}) , \quad f_{\mathbf{S}|\mathfrak{D}}(\mathbf{s}) = \mathcal{N}(\mathbf{s}|\boldsymbol{\mu}_{\mathfrak{D}}, \boldsymbol{\Sigma}_{\mathfrak{D}}) , \tag{7}$$

where $\boldsymbol{\mu}_{\mathfrak{S}}, \boldsymbol{\Sigma}_{\mathfrak{S}}, \boldsymbol{\mu}_{\mathfrak{D}}, \boldsymbol{\Sigma}_{\mathfrak{D}}$ are parameters that can be estimated by ML on a development set. The joint model induces a form for the marginal densities that corresponds to a calibration transformation for the individual systems. Ideally we would like the marginal densities to preserve the characteristics of generative calibration models. For example, CMLG marginals can be recovered by imposing that the covariance matrices $\boldsymbol{\Sigma}_{\mathfrak{S}}$ and $\boldsymbol{\Sigma}_{\mathfrak{D}}$ share the same diagonal:

$$\boldsymbol{\Sigma}_{\mathfrak{h}} = \begin{bmatrix} v_1 & r_{\mathfrak{h},12} & \ldots & r_{\mathfrak{h},1d} \\ r_{\mathfrak{h},12} & v_2 & \ldots & r_{\mathfrak{h},2d} \\ \vdots & \ldots & \ddots & \vdots \\ r_{\mathfrak{h},1d} & r_{\mathfrak{h},2d} & \ldots & v_d \end{bmatrix} , \quad \mathfrak{h} \in \{\mathfrak{S}, \mathfrak{D}\} . \tag{8}$$

As we show in the experimental section, model (7), (8) does not prove accurate enough for score-level fusion. The Gaussian marginal assumption is indeed often quite crude, and the inaccuracies in modeling the marginal densities are inherited by the joint density model. In the context of calibration VΓ densities were introduced as a more flexible and powerful alternative to Gaussian models. The rationale behind VΓ densities derives from an analysis of the score distribution induced by Gaussian-distributed embedding for PLDA-like classifiers. Although a similar analysis may eventually lead to the definition of a robust joint score model, such a model may easily become intractable, due to the complexity in characterizing the dependency sources of different setups (e.g. same embedding with different back-ends, same back-end used for different front-ends, and so on) and the lack of closed form expressions for the complex distribution densities that would arise[1]. To keep the model tractable, we adopt a different approach and, following Sklar's theorem [27], we employ a copula-based representation [22, 23] of the joint densities that decouples modeling of the scores marginals from the characterization of the scores dependencies. As shown in Section 4, this allows us to keep the benefits of the accurate Variance-Gamma characterization of target and non target scores of individual systems. We consider models that can be expressed as

$$\begin{aligned} f_{\mathbf{S}|\mathfrak{h}}(\mathbf{s}|\boldsymbol{\theta}_{\mathfrak{h}}, \boldsymbol{\rho}_{\mathfrak{h}}) = c_{\mathfrak{h}}\left(F_{S_1|\mathfrak{h}}(s_1|\boldsymbol{\theta}_{\mathfrak{h},1}) \ldots F_{S_d|\mathfrak{h}}(\mathbf{s}_d|\boldsymbol{\theta}_{\mathfrak{h},d})|\boldsymbol{\rho}_{\mathfrak{h}}\right) \\ \cdot \prod_{i=1}^{d} f_{S_i|\mathfrak{h}}(s_i|\boldsymbol{\theta}_{\mathfrak{h},i}) , \quad \mathfrak{h} \in \{\mathfrak{S}, \mathfrak{D}\} , \end{aligned} \tag{9}$$

where $f_{S_i|\mathfrak{h}}(s_i|\boldsymbol{\theta}_{\mathfrak{S},i})$ and $f_{S_i|\mathfrak{D}}(s_i|\boldsymbol{\theta}_{\mathfrak{D},i})$ are the marginal target and non-target score distribution densities for the i-th individual system (component s_i of the score vector $\mathbf{s}$), with parameters $\boldsymbol{\theta}_{\mathfrak{S},i}$ and $\boldsymbol{\theta}_{\mathfrak{D},i}$, respectively. $F_{S_i|\mathfrak{S}}$ and $F_{S_i|\mathfrak{D}}$ are

[1]The VΓ calibration approach of [15] already requires simplifying assumptions to keep the derivations tractable.

the corresponding cumulative distribution functions (CDF), and $c_{\mathfrak{S}}(u_1 \dots u_n|\boldsymbol{\rho}_{\mathfrak{S}})$ and $c_{\mathfrak{D}}(u_1 \dots u_n|\boldsymbol{\rho}_{\mathfrak{D}})$ are parametric copula density functions, with parameters $\boldsymbol{\rho}_{\mathfrak{S}}$ and $\boldsymbol{\rho}_{\mathfrak{D}}$, that capture the relationships among the different scores. Given a score vector $\mathbf{s}$, the fused score $s_f(\mathbf{s})$ is can be computed as

$$\begin{aligned} s_f(\mathbf{s}) &= \log \frac{f_{\mathbf{S}|\mathfrak{S}}(\mathbf{s}|\boldsymbol{\theta}_{\mathfrak{S}}, \boldsymbol{\rho}_{\mathfrak{S}})}{f_{\mathbf{S}|\mathfrak{D}}(\mathbf{s}|\boldsymbol{\theta}_{\mathfrak{D}}, \boldsymbol{\rho}_{\mathfrak{D}})} \\ &= \sum_{i=1}^{d} \log \frac{f_{S_i|\mathfrak{S}}(s_i|\boldsymbol{\theta}_{\mathfrak{S},i})}{f_{S_i|\mathfrak{D}}(s_i|\boldsymbol{\theta}_{\mathfrak{D},i})} \\ &\quad + \log \frac{c_{\mathfrak{S}}\left(F_{S_1|\mathfrak{S}}(s_1|\boldsymbol{\theta}_{\mathfrak{S},1}) \dots F_{S_d|\mathfrak{S}}(s_d|\boldsymbol{\theta}_{\mathfrak{S},d})|\boldsymbol{\rho}_{\mathfrak{S}}\right)}{c_{\mathfrak{D}}\left(F_{S_1|\mathfrak{D}}(s_1|\boldsymbol{\theta}_{\mathfrak{D},1}) \dots F_{S_d|\mathfrak{D}}(s_d|\boldsymbol{\theta}_{\mathfrak{D},d})|\boldsymbol{\rho}_{\mathfrak{D}}\right)}, \end{aligned}$$

i.e., the sum of the *recalibrated scores of the individual systems* and an additional term that accounts for the dependency structure induced by the copula functions [22].

4. Gaussian copula models

From Sklar's theorem [27], we can freely combine marginal and copula densities and obtain valid joint density models. The Gaussian model (7), (8) can be represented as the combination of CMLG marginals with a Gaussian copula, by letting

$$f_{S_i|\mathfrak{h}}(s_i) = \mathcal{N}(s_i|\mu_{\mathfrak{h},i}, v_i)\ , \quad \mathfrak{h} \in \{\mathfrak{S}, \mathfrak{D}\}\ , \tag{10}$$

combined with the Gaussian copula

$$C_{\mathfrak{h}}(u_1 \dots u_d|\mathbf{R}_{\mathfrak{h}}) = \boldsymbol{\Phi}\left(\Phi^{-1}(u_1) \dots \Phi^{-1}(u_d)|\mathbf{R}_{\mathfrak{h}}\right)\ , \tag{11}$$

whose density is given by

$$c_{\mathfrak{h}}(u_1 \dots u_d|\mathbf{R}_{\mathfrak{h}}) = \frac{\mathcal{N}\left(\left[\Phi^{-1}(u_1) \dots \Phi^{-1}(u_d)\right]^T \middle| \mathbf{0}, \mathbf{R}_{\mathfrak{h}}\right)}{\prod_{i=1}^{d} \mathcal{N}(\Phi^{-1}(u_i)|0,1)}\ . \tag{12}$$

$\boldsymbol{\Phi}(\cdot|\mathbf{R})$ is the CDF of a zero-mean multivariate normal distribution with covariance matrix $\mathbf{R}$ and $\Phi^{-1}(\cdot)$ is the inverse CDF of a standard normal distribution. The copula parameters $\mathbf{R}_{\mathfrak{S}}$ and $\mathbf{R}_{\mathfrak{D}}$ are correlation matrices

$$\mathbf{R}_{\mathfrak{h}} = \begin{bmatrix} 1 & \rho_{\mathfrak{h},12} & \dots & \rho_{\mathfrak{h},1d} \\ \rho_{\mathfrak{h},12} & 1 & \dots & \rho_{\mathfrak{h},2d} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{\mathfrak{h},1d} & \rho_{\mathfrak{h},2d} & \dots & 1 \end{bmatrix}\ , \quad \mathfrak{h} \in \{\mathfrak{S}, \mathfrak{D}\}\ , \tag{13}$$

and fully specify the dependency structure of the model. The multivariate CMLG extension in (7), (8) can be obtained from (10), (12) and (13) setting $r_{\mathfrak{h},ij} = \rho_{\mathfrak{h},ij}\sqrt{v_i v_j}$.

To address the limitations of CMLG-based fusion, in this work we propose to preserve the Gaussian copula dependency structure (12), but we replace the Gaussian marginals with the more accurate VΓ model (3). The joint target and non-target densities are given by

$$\begin{aligned} f_{\mathbf{S}|\mathfrak{h}}(\mathbf{s}|\boldsymbol{\theta}_{\mathfrak{h}}, \mathbf{R}_{\mathfrak{h}}) = \ & c_g(F_{\mathrm{V\Gamma}}(s_1|\boldsymbol{\theta}_{\mathfrak{h},i}) \dots F_{\mathrm{V\Gamma}}(s_d|\boldsymbol{\theta}_{\mathfrak{h},i})|\mathbf{R}_{\mathfrak{h}}) \\ & \cdot \prod_{i=1}^{d} f_{\mathrm{V\Gamma}}(s|\boldsymbol{\theta}_{\mathfrak{h},i})\ , \quad \mathfrak{h} \in \{\mathfrak{S}, \mathfrak{D}\}\ , \end{aligned} \tag{14}$$

where $\boldsymbol{\theta}_{\mathfrak{S}} = (\boldsymbol{\theta}_{\mathfrak{S},1} \dots \boldsymbol{\theta}_{\mathfrak{S},d})$ and $\boldsymbol{\theta}_{\mathfrak{D}} = (\boldsymbol{\theta}_{\mathfrak{D},1} \dots \boldsymbol{\theta}_{\mathfrak{D},d})$ are the parameters of the marginal target and non-target score distribution of the individual systems, and $F_{\mathrm{V\Gamma}}$ denotes the VΓ CDF. The model parameters can be extimated by maximizing a weighted log-likelihood

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}_{\mathfrak{S}}, \boldsymbol{\theta}_{\mathfrak{D}}, \mathbf{R}_{\mathfrak{S}}, \mathbf{R}_{\mathfrak{D}}) = \ & \frac{\zeta}{|\mathcal{S}_{\mathfrak{S}}|} \sum_{\mathbf{s} \in \mathcal{S}_{\mathfrak{S}}} \log f_{\mathbf{S}|\mathfrak{S}}\left(\mathbf{s}|\boldsymbol{\theta}_{\mathfrak{S}}, \mathbf{R}_{\mathfrak{S}}\right) \\ & + \frac{1-\zeta}{|\mathcal{S}_{\mathfrak{D}}|} \sum_{\mathbf{s} \in \mathcal{S}_{\mathfrak{D}}} \log f_{\mathbf{S}|\mathfrak{D}}\left(s|\boldsymbol{\theta}_{\mathfrak{D}}, \mathbf{R}_{\mathfrak{D}}\right)\ , \end{aligned} \tag{15}$$

where $\mathcal{S}_{\mathfrak{S}}$ and $\mathcal{S}_{\mathfrak{D}}$ are the sets of target and non-target score vectors, and $\zeta \in (0,1)$ is a tunable weight.

Direct optimization of the likelihood is complex due to the presence of the VΓ CDF, which cannot be expressed in closed form, and depends on the optimization parameters. We therefore employ a two-step approach, known as inference function for margins [23], that consists in the independent ML optimization of the marginal densities, i.e., the estimation of the *individual calibration model* parameters $\boldsymbol{\theta}_{\mathfrak{S}}, \boldsymbol{\theta}_{\mathfrak{D}}$, followed by ML estimation of the copula parameters $\mathbf{R}_{\mathfrak{S}}$ and $\mathbf{R}_{\mathfrak{D}}$. We can estimate the marginal density parameters $\boldsymbol{\theta}_{\mathfrak{S},i}$, $\boldsymbol{\theta}_{\mathfrak{D},i}$ by maximizing, as in [15], the marginal weighted log-likelihood

$$\begin{aligned} \mathcal{L}_m(\boldsymbol{\theta}_{\mathfrak{S},i}, \boldsymbol{\theta}_{\mathfrak{D},i}) = \ & \frac{\zeta}{|\mathcal{S}_{\mathfrak{S}}|} \sum_{s \in \mathcal{S}_{\mathfrak{S},i}} \log f_{\mathrm{V\Gamma}}(s|\boldsymbol{\theta}_{\mathfrak{S},i}) \\ & + \frac{1-\zeta}{|\mathcal{S}_{\mathfrak{D}}|} \sum_{s \in \mathcal{S}_{\mathfrak{D},i}} \log f_{\mathrm{V\Gamma}}(s|\boldsymbol{\theta}_{\mathfrak{D},i})\ , \end{aligned} \tag{16}$$

where $\mathcal{S}_{\mathfrak{S},i}$ and $\mathcal{S}_{\mathfrak{D},i}$ are the sets of target and non-target scores of the i–th system. Once we have obtained the marginal parameters, we maximize the joint log-likelihood $\mathcal{L}$ in (15) with respect to the copula parameters $\mathbf{R}_{\mathfrak{S}}$ and $\mathbf{R}_{\mathfrak{D}}$. We observe that, since the marginals do not depend on $\mathbf{R}_{\mathfrak{S}}$ and $\mathbf{R}_{\mathfrak{D}}$, we can replace maximization of $\mathcal{L}$ with the maximization of

$$\begin{aligned} \mathcal{L}'(\mathbf{R}_{\mathfrak{S}}, \mathbf{R}_{\mathfrak{D}}) = \ & \frac{\zeta}{|\mathcal{S}_{\mathfrak{S}}|} \sum_{\mathbf{s} \in S_{\mathfrak{S}}} \ell(\mathbf{g}(\mathbf{s}, \boldsymbol{\theta}_{\mathfrak{S}})|\mathbf{R}_{\mathfrak{S}}) \\ & + \frac{1-\zeta}{|\mathcal{S}_{\mathfrak{D}}|} \sum_{\mathbf{s} \in S_{\mathfrak{D}}} \ell(\mathbf{g}(\mathbf{s}, \boldsymbol{\theta}_{\mathfrak{D}})|\mathbf{R}_{\mathfrak{D}})\ , \end{aligned} \tag{17}$$

where function $\mathbf{g}$ transforms a d-dimensional score vector $\mathbf{s} = [s_1 \dots s_d]^T$ into a d-dimensional vector

$$\mathbf{g}(\mathbf{s}, \boldsymbol{\theta}) = \left[\Phi^{-1}(F_{\mathrm{V\Gamma}}(s_i|\boldsymbol{\theta}_i) \dots \Phi^{-1}(F_{\mathrm{V\Gamma}}(s_i|\boldsymbol{\theta}_i)\right]^T\ , \tag{18}$$

i.e., the d scores of the score vector $\mathbf{s}$ are transformed through the CDF of the corresponding VΓ marginal and through the inverse CDF of a standard normal distribution. Function ℓ is given by

$$\ell(\mathbf{x}) = \log \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{R})\ . \tag{19}$$

We can observe that (17) requires optimizing a normal log-likelihood over the transformed score vectors $\mathbf{g}(\mathbf{s}, \boldsymbol{\theta}_{\mathfrak{D}})$ and $\mathbf{g}(\mathbf{s}, \boldsymbol{\theta}_{\mathfrak{S}})$, with the constraint that $\mathbf{R}_{\mathfrak{S}}$ and $\mathbf{R}_{\mathfrak{D}}$ represent correlation matrices (i.e., their diagonal should be 1). Since the VΓ CDF cannot be expressed in closed form, we employ numerical integration to compute $F_{\mathrm{V\Gamma}}$.

5. Experimental results

To assess the performance of our approach we compare our VΓ marginal Gaussian Copula (VΓ-GC) model with the state-of-the-art linear Logistic Regression (Log-Reg) method [10, 11,

Table 1: *C_{llr}, EER % and C_{prim} of different calibration and fusion models on the SRE 2019 Evaluation set. Rows labeled* min. cost *show the minimum costs for the non-calibrated scores.*

	C_{llr}	EER %	C_{prim}	C_{llr}	EER %	C_{prim}	C_{llr}	EER %	C_{prim}	C_{llr}	EER %	C_{prim}
	S1 ECAPA - PLDA			S2 ECAPA - PSVM			S3 FTDNN - PLDA			S4 FTDNN - PSVM		
min. cost	*0.176*	4.7	*0.365*	*0.135*	3.5	*0.330*	*0.163*	4.3	*0.326*	*0.145*	3.7	*0.350*
Log-Reg ($\pi = 0.1$)	0.191	4.7	0.376	0.145	3.5	0.337	0.176	4.3	**0.332**	0.154	3.7	0.361
Log-Reg ($\pi = 0.5$)	0.187	4.7	0.419	0.143	3.5	0.370	0.172	4.3	0.371	0.152	3.7	0.397
CMLG	0.187	4.7	0.455	0.147	3.5	0.429	0.172	4.3	0.432	0.155	3.7	0.467
VΓ	**0.178**	4.7	**0.366**	**0.137**	3.5	**0.332**	**0.165**	4.3	0.333	**0.147**	3.7	**0.352**
	S1 + S2 (ECAPA)			S2 + S4 (PSVM)			S2 + S3			S1 + S2 + S3 + S4		
Log-Reg ($\pi = 0.1$)	0.142	**3.4**	0.311	0.123	**2.8**	**0.268**	0.129	3.0	**0.266**	0.120	**2.7**	**0.249**
Log-Reg ($\pi = 0.5$)	0.140	**3.4**	0.352	0.121	**2.8**	0.300	0.126	3.0	0.296	0.118	**2.7**	0.281
CMLG-GC	0.167	3.9	0.349	0.129	2.9	0.342	0.141	3.3	0.319	0.168	3.7	0.279
VΓ-GC	**0.133**	**3.4**	**0.309**	**0.115**	**2.8**	0.270	**0.120**	**2.9**	0.268	**0.114**	2.8	0.262

13]. A first set of experiments on the SRE 2019 Evaluation set is presented in Table 1. We consider two different embedding extractors: a CNN-ECAPA architecture that combines an ECAPA network [28] with CNN input blocks [29], and a Factorized Time-Delay (FTDNN) architecture, implemented as in [30]. The front-ends have been trained on a common list including VoxCeleb1 and VoxCeleb2 [31], Mixer 4,5 and 6 [32, 33] and Switchboard [34, 35, 36, 37, 38, 39] data. The MUSAN [40] and the AIR [41] datasets were used for data augmentation. We also consider two different front-ends: Probabilistic Linear Discriminant Analysis (PLDA) [42, 43, 44] and Pairwise Support Vector Machine (PSVM) [45, 46, 47, 48, 49]. We consider three metrics: the calibration-sensitive C_{llr} [10], the actual primary cost C_{prim}, as defined by NIST for SRE 2019 [24], and the calibration-insensitive Equal Error Rate (EER). The calibration models employed as marginal densities for our fusion models and the copula density parameters have been estimated on a subset of the SRE 2019 Progress set. Since the quality of the marginal density models affects the generative fusion, we also report the calibration results for individual systems. For calibration, we consider prior-weighted logistic regression models [11, 14], the CMLG approach [21] and the VΓ method of [15], corresponding to the marginal models employed by VΓ-GC. For fusion, in addition the Log-Reg baseline we consider the Gaussian-copula extension of the CMLG (CMLG-GC) presented in Section 3. For discriminative models we show results with target prior π set to 0.1 and to 0.5, respectively, with the latter models providing minor C_{llr} improvements but at the cost of significant degradation of C_{prim}. For generative models we employ a weight $\zeta = 0.5$. The first five rows of the table show minimum costs and calibration performance of different models for different front-end / back-end combinations. The last four rows compare our VΓ-GC with our baselines for different systems combinations. We can observe that fusion and calibration results are consistent. CMLG calibration models are able to provide acceptable C_{llr}, but present significant degradation in terms of C_{prim}, due to their weaker characterization of the score distribution [17]. The issue is amplified in CMLG-GC models, that behave significantly worse than other approaches. VΓ calibration models are able to better capture the scores characteristics, achieving the best C_{llr}, and optimal or close-to-optimal C_{prim}, despite not being trained for a specific working point. The VΓ-GC fusion benefits from the more accurate VΓ marginals, and achieves again optimal C_{llr} and optimal or close-to-optimal C_{prim}.

A second set of experiments was conducted on the SITW [25] Evaluation dataset. In this case, we consider two systems, an ECAPA - PSVM and a FTDNN - PSVM. The results are shown in Table 2. The calibration and fusion models were trained on the SITW Development set. As for SRE 2019, we can observe that VΓ-GC provides competitive results with respect to state-of-the-art approaches, although the gains of all the methods are smaller than in the SRE19 scenario.

Table 2: *C_{llr}, EER % and C_{prim} of different calibration and fusion models on the SITW Evaluation set. Rows labeled* min. cost *show the minimum costs for the non-calibrated scores.*

		C_{llr}	EER %	C_{prim}
ECAPA	*min. cost*	*0.07*	*1.9*	*0.22*
	Log-Reg ($\pi = 0.1$)	0.08	1.9	0.23
	CMLG	0.08	1.9	0.22
	VΓ	0.08	1.9	0.22
FTDNN	*min. cost*	*0.11*	*3.1*	*0.30*
	Log-Reg ($\pi = 0.1$)	0.11	3.1	0.36
	CMLG	0.12	3.1	0.38
	VΓ	0.11	3.1	0.31
FUSION	Log-Reg ($\pi = 0.1$)	0.07	1.7	0.20
	CMLG-GC	0.08	1.9	0.21
	VΓ-GC	0.07	1.7	0.19

6. Conclusions

We have presented a novel generative approach for the score-level fusion of speaker verification systems. The method is based on the combination of a robust Variance-Gamma marginal model, able to well characterize the score distribution of individual systems, and a Gaussian copula that captures the relationships between scores of different systems. The resulting model provides competitive results with respect to state-of-the-art discriminative methods, successfully extending recent generative calibration models to multi-system score level fusion. Future work will analyze more complex copula functions, with the aim of further improving the fusion performance.

7. References

[1] M. J. Alam *et al.*, "Development of ABC systems for the 2021 edition of NIST speaker recognition evaluation," in *Proc. Odyssey 2022*, 06 2022, pp. 346–353.

[2] J. Villalba *et al.*, "Advances in cross-lingual and cross-source audio-visual speaker recognition: The JHU-MIT system for NIST SRE21," in *Proc. Odyssey 2022*, 06 2022, pp. 213–220.

[3] K. A. Lee *et al.*, "I4U system description for NIST SRE-20 CTS challenge," in *SRE 2021, NIST Speaker Recognition Evaluation Workshop*, NIST, Ed., 2021.

[4] M. J. Alam *et al.*, "Analysis of ABC submission to NIST SRE 2019 CMN and VAST challenge," in *Proc. Odyssey 2020*, 11 2020, pp. 289–295.

[5] D. Colibro *et al.*, "Nuance - Politecnico di Torino's 2016 NIST Speaker Recognition Evaluation system," in *Proc. Interspeech 2017*, 2017, pp. 1338–1342.

[6] L. Ferrer *et al.*, "A noise-robust system for NIST 2012 speaker recognition evaluation," in *Proc. Interspeech 2013*, 2013.

[7] D. Colibro and al., "Nuance–Politecnico di Torino 2012 NIST Speaker Recognition Evaluation system," in *Proc. Interspeech 2013*, 2013, pp. 1996–2000.

[8] S. Cumani and P. Laface, "Joint estimation of PLDA and non–linear transformations of speaker vectors," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1890–1900, 2017.

[9] S. Cumani *et al.*, "Analysis of the ABC classification backends for NIST SRE24," in *Proc. of Interspeech 2025*, 2025.

[10] N. Brümmer, "Measuring, refining and calibrating speaker and language information extracted from speech," Ph.D. dissertation, Stellenbosch University, South Africa, 2010.

[11] N. Brummer *et al.*, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST Speaker Recognition Evaluation 2006," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[12] M. Kockmann *et al.*, "Ivector fusion of prosodic and cepstral features for speaker verification," in *Proc. Interspeech 2011*, 2011.

[13] N. Brümmer, "Focal toolkit," Available at http://sites.google.com/site/nikobrummer/focal.

[14] N. Brümmer and G. R. Doddington, "Likelihood-ratio calibration using prior-weighted proper scoring rules," in *Proc. Interspeech 2013*, 2013.

[15] S. Cumani and S. Sarni, "The distributions of uncalibrated speaker verification scores: a generative model for domain mismatch and trial-dependent calibration," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, pp. 1–16, 01 2023.

[16] S. Cumani and S. Sarni, "A generative model for duration-dependent score calibration," in *Proc. Interspeech 2021*, 2021, pp. 4598–4602.

[17] S. Cumani, "On the distribution of speaker verification scores: Generative models for unsupervised calibration," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 547–562, 2021.

[18] S. Cumani, "Normal variance-mean mixtures for unsupervised score calibration," in *Interspeech 2019*, 2019, pp. 401–405.

[19] N. Brümmer, A. Swart, and D. van Leeuwen, "A comparison of linear and nonlinear calibrations for speaker recognition," in *Proc. Odyssey 2014*, 2014, pp. 14–18.

[20] N. Brümmer and D. Garcia-Romero, "Generative modelling for unsupervised score calibration," in *Proc. ICASSP 2014*, 2014, pp. 1680–1684.

[21] D. van Leeuwen and N. Brümmer, "The distribution of calibrated likelihood-ratios in speaker recognition," in *Proc. Interspeech 2013*, 2013, pp. 1619–1623.

[22] S. C. Dass, K. Nandakumar, and A. K. Jain, "A principled approach to score level fusion in multimodal biometric systems," in *Audio- and Video-Based Biometric Person Authentication*. Springer, 2005, pp. 1049–1058.

[23] H. Joe and J. J. Xu, "The estimation method of inference functions for margins for multivariate models," Oct 1996. [Online]. Available: https://open.library.ubc.ca/collections/facultyresearchandpublications/52383/items/1.0225985

[24] S. O. Sadjadi *et al.*, "The 2019 NIST Speaker Recognition Evaluation CTS challenge," in *Proc. Odyessy 2020*, 2020.

[25] M. McLaren *et al.*, "The Speakers in the Wild (SITW) Speaker Recognition Database," in *Proc. Interspeech 2016*, 2016.

[26] D. Madan, P. Carr, and E. Chang, "The Variance Gamma process and option pricing," *Eurepean Finance Review*, vol. 2, pp. 79–105, 1998.

[27] M. Sklar, "Fonctions de répartition à n dimensions et leurs marges," *Annales de l'ISUP*, 1959.

[28] B. Desplanques *et al.*, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. Interspeech 2020*, 2020.

[29] J. Thienpondt *et al.*, "Integrating frequency translational invariance in tdnns and frequency positional information in 2d resnets to enhance speaker verification," in *Proc. Interspeech 2021*, 2021.

[30] J. Villalba *et al.*, "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations," *Computer Speech & Language*, vol. 60, 2020.

[31] A. Nagrani *et al.*, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[32] C. Cieri, L. Corson, D. Graff, , and K. Walker, "Resources for new research directions in speaker recognition: the Mixer 3, 4 and 5 corpora," in *Proc. Interspeech 2007*, 2007.

[33] L. Brandschain *et al.*, "The Mixer 6 corpus: resources for cross-channel and text independent speaker recognition," in *Proc. LREC 2010*, 2010.

[34] J. Godfrey and E.Holliman, "Switchboard-1 release 2," 1993.

[35] D. Graff *et al.*, "Switchboard-2 phase I," 1998.

[36] ——, "Switchboard-2 phase II," 1999.

[37] ——, "Switchboard-2 phase III," 2002.

[38] ——, "Switchboard cellular part 1 audio," 2001.

[39] ——, "Switchboard cellular part 2 audio," 2004.

[40] D. Snyder, G. Chen, and D. Povey, "MUSAN: a music, speech, and noise corpus," 2015, arXiv:1510.08484v1.

[41] T. Ko *et al.*, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP 2017*, 2017.

[42] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proc. ECCV'06*, 2006.

[43] P. Kenny, "Bayesian speaker verification with Heavy-Tailed Priors," in *Keynote presentation, Proc. Odyssey 2010*, 2010.

[44] N. Brümmer *et al.*, "Gaussian meta-embeddings for efficient scoring of a heavy-tailed plda model," in *Proc. Odyssey 2018*, 2018.

[45] S. Cumani and P. Laface, "Large scale training of Pairwise Support Vector Machines for speaker recognition," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 11, pp. 1590–1600, 2014.

[46] S. Cumani and P. Laface, "Training pairwise support vector machines with large scale datasets," in *Proc. ICASSP 2014*, 2014.

[47] S. Cumani, N. Brümmer, L. Burget, P. Laface, O. Plchot, and V. Vasilakakis, "Pairwise discriminative speaker verification in the i-vector space," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1217–1227, 2013.

[48] S. Cumani *et al.*, "Gender independent discriminative speaker recognition in i-vector space," in *Proc. of ICASSP 2012*, 2012.

[49] S. Cumani and P. Laface, "Generative pairwise models for speaker recognition," in *Proc. of Odyssey 2014*, 2014.