Spherical-Gaussian TPSDA: combining PLDA, T-PSDA and duration models for speaker verification

Odyssey 2026
23-26 June 2026, Lisbon, Portugal

Spherical-Gaussian TPSDA: combining PLDA, T-PSDA and duration models for speaker verification

Sandro Cumani[1]

[1]Politecnico di Torino, Italy
sandro.cumani@polito.it

Abstract

Despite the effectiveness of cosine scoring compared to classical Probabilistic Linear Discriminant Analysis (PLDA), structured generative models such as Toroidal Probabilistic Spherical Discriminant Analysis (T-PSDA) can still provide superior performance for DNN-based speaker verification. In this work we investigate the relationship between T-PSDA and an isotropic simplified PLDA (ISO-PLDA) model, and propose a novel approach that combines the ISO-PLDA Gaussian likelihood with the structured directional T-PSDA speaker characterization to obtain a Spherical-Gaussian (SG-TPSDA) model that is able to match and, in some cases, improve T-PSDA performance. Furthermore, we show how to extend both ISO-PLDA and SG-TPSDA to explicitly account for utterance duration. Our results on NIST and CN-Celeb data show that the proposed approach is effective and capable of leveraging duration information to further improve speaker verification accuracy.

Index Terms: Speaker verification, Probabilistic Linear Discriminant Analysis, Spherical Discriminant Analysis, duration models

1. Introduction

Recent years have seen the rise of speaker verification systems based on DNN embeddings. The success of DNN embedding extractors in computing information rich utterance representations, and in particular of models trained with angular distances, has been accompanied by a decline of interest in back-end modeling. This is in part due to the good performance of simple back-ends based on cosine scoring [1, 2, 3], which have shown the potential to match or even outperform more complex generative approaches such as Probabilistic Linear Discriminant Analysis (PLDA) [4, 5, 6]. Recent works, however, have shown that robust back-ends can still provide superior performance. In [7] the authors have successfully employed domain-adapted PLDA with DNN embeddings to address the NIST SRE21 challenge[1]. More recently, the authors of [8] have shown that the relative performance of cosine and PLDA can significantly change depending on the front-end and the evaluation metric, and that robust back-ends such as Pairwise Support Vector Machine [9, 10], originally developed for the i-vector [11] framework, consistently outperform both. On the generative side, the authors of [1] have shown that a diagonal variant of the two-covariance model [12] was able to outperform classical PLDA and achieve competitive results with respect to cosine scoring. The diagonal PLDA was further constrained to a spherical model in [13] to act as a probabilistic version of cosine. A different approach specifically tailored towards unit-norm embeddings has been proposed in [2] and further extended in [14]. The Toroidal Probabilistic Spherical Discriminant Analysis (T-PSDA) [14] models directional data with a significant degree of flexibility, allowing for explicit inclusion of multiple speaker and channel subspaces, similarly to classical PLDA. The presence of reduced-dimensional speaker subspaces, which are absent in the Spherical PLDA formulation of [13], proves crucial, allowing T-PSDA to significantly outperform cosine scoring in several cases.

The shift from i-vectors to DNN embeddings has also seen reduced interest in modeling nuisance sources at back-end level, and compensation of utterance duration effects are usually left to the front-end and to the length-normalization that is implicit in cosine scoring, or to score normalization and score calibration [15, 16, 17, 18]. The i-vector framework provided a natural way to incorporate duration-dependent uncertainty, encoded in the i-vector posterior covariance, into PLDA [19, 20, 21, 22, 23]. On the other hand, DNN embeddings typically do not provide explicit uncertainty measures, and although architectural solutions have been proposed that try to address this limitation [24, 25] and incorporate uncertainty in cosine scoring [26], they require modification of the embedding extractor and are often not available for pre-trained models.

The goal of this work is to re-analyze the connections between PLDA and T-PSDA to propose a novel approach that combines an i-vector/PLDA inspired duration model with the effective structured speaker characterization of T-PSDA. We start introducing an extension of the Spherical PLDA model of [13] that employs isotropic covariance terms but, similarly to classical PLDA, allows for a reduced dimensional speaker subspace. We refer to this model as Isotropic PLDA, or ISO-PLDA for short. We show that ISO-PLDA can significantly improve performance with respect to cosine scoring, despite being, in most of our tests, slightly worse than T-PSDA. We proceed showing how to extend ISO-PLDA to incorporate duration information without relying on an external measure of embedding uncertainty. Given the superiority of T-PSDA compared to ISO-PLDA, we then focus on the relationship between the two approaches, and propose a novel model that combines the ISO-PLDA likelihood with the T-PSDA speaker prior. The model, named Spherical-Gaussian TPSDA (SG-TPSDA), achieves similar performance as T-PSDA on unit-norm embeddings, but, in contrast with T-PSDA, does not require inputs to be normalized, leading to a slight increase of performance in some use cases when length normalization is not applied. Finally, the Gaussian likelihood of the SG-TPSDA allows us to incorporate in SG-TPSDA our ISO-PLDA duration model, achieving a duration-aware back-end able to outperform both ISO-PLDA and T-PSDA.

[1]https://www.nist.gov/itl/iad/mig/nist-2021-speaker-recognition-evaluation-sre21

10.21437/Odyssey.2026-42

The rest of the paper is organized as follows. Section 2 introduces our ISO-PLDA and our duration models. Section 3 presents our SG-TPSDA approach. Experimental results are shown in Section 4, and conclusions are drawn in Section 5.

2. Isotropic PLDA

2.1. Simplified PLDA, two-covariance model and cosine scoring

The simplified PLDA model [6] assumes that a D-dimensional speaker embedding $\boldsymbol{\phi}_{s,i}$ can be represented in terms of a low-dimensional speaker factor as

$$\boldsymbol{\phi}_{s,i} = \mathbf{m} + \mathbf{U}\mathbf{y}_s + \boldsymbol{\varepsilon}_{s,i} \,, \tag{1}$$

where s denotes the speaker, i is an utterance index, $\mathbf{m}$ is a mean vector, $\mathbf{y}_s$ is a d-dimensional, with $d \leq D$, speaker factor constrained in the subspace identified by the $D \times d$ factor loading matrix $\mathbf{U}$, and $\boldsymbol{\varepsilon}_{s,i}$ is a residual noise term. PLDA assumes Gaussian priors for both speaker and noise terms, corresponding to Gaussian speaker prior and conditional likelihood

$$\begin{aligned} P(\mathbf{y}_s) &= \mathcal{N}(\mathbf{y}_s|\mathbf{0},\mathbf{I}) \,, \\ P(\boldsymbol{\phi}_{s,i}|\mathbf{y}_s) &= \mathcal{N}\left(\boldsymbol{\phi}_{s,i}|\mathbf{m}+\mathbf{U}\mathbf{y}_s,\mathbf{W}\right) \,, \end{aligned} \tag{2}$$

where $\mathbf{W}$ is the residual noise covariance matrix. Assuming a D-dimensional speaker subspace (i.e. $d = D$) leads to the two-covariance model of [12], which can be expressed in terms of full-rank between class and within-class covariance matrices $\mathbf{B} = \mathbf{U}\mathbf{U}^T$ and $\mathbf{W}$ as

$$\boldsymbol{\phi}_{s,i} = \mathbf{y}_s + \boldsymbol{\varepsilon}_{s,i} \,, \tag{3}$$

with

$$\begin{aligned} P(\mathbf{y}_s) &= \mathcal{N}(\mathbf{y}_s|\mathbf{m},\mathbf{B}) \,, \\ P(\boldsymbol{\phi}_{s,i}|\mathbf{y}_s) &= \mathcal{N}\left(\boldsymbol{\phi}_{s,i}|\mathbf{y}_s,\mathbf{W}\right) \,. \end{aligned} \tag{4}$$

Given a pair (trial) of embeddings $(\boldsymbol{\phi}_e, \boldsymbol{\phi}_t)$ scoring entails computing the log-likelihood ratio (LLR) l between the same-speaker and different speaker hypotheses:

$$l(\boldsymbol{\phi}_e, \boldsymbol{\phi}_t) = \log \frac{\int P(\boldsymbol{\phi}_e|\mathbf{y})P(\boldsymbol{\phi}_t|\mathbf{y})P(\mathbf{y})d\mathbf{y}}{\int P(\boldsymbol{\phi}_e|\mathbf{y})P(\mathbf{y})d\mathbf{y} \int P(\boldsymbol{\phi}_t|\mathbf{y})P(\mathbf{y})d\mathbf{y}} \,. \tag{5}$$

The LLR of the two-covariance model (4) corresponds to a quadratic form of the trial [9, 27]

$$l(\boldsymbol{\phi}_e, \boldsymbol{\phi}_t) = 2\boldsymbol{\phi}_e^T\boldsymbol{\Lambda}\boldsymbol{\phi}_t + \boldsymbol{\phi}_e^T\boldsymbol{\Gamma}\boldsymbol{\phi}_e + \boldsymbol{\phi}_t^T\boldsymbol{\Gamma}\boldsymbol{\phi}_t + \mathbf{c}^T\boldsymbol{\phi}_e + \mathbf{c}^T\boldsymbol{\phi}_t + k \,, \tag{6}$$

where $\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}$ are symmetric matrices, $\mathbf{c}$ is a vector and k a scalar, whose values depend on $\mathbf{B}$, $\mathbf{W}$ and $\mathbf{m}$:[2]

$$\begin{aligned} \boldsymbol{\Lambda} &= \frac{1}{2}\mathbf{W}^{-T}\tilde{\boldsymbol{\Lambda}}\mathbf{W}^{-1} \,, \quad \boldsymbol{\Gamma} = \frac{1}{2}\mathbf{W}^{-T}\left(\tilde{\boldsymbol{\Lambda}} - \tilde{\boldsymbol{\Gamma}}\right)\mathbf{W}^{-1} \,, \\ k &= \frac{1}{2}\tilde{k} + \frac{1}{2}\left(\mathbf{B}^{-1}\mathbf{m}\right)^T\left(\tilde{\boldsymbol{\Lambda}} - 2\tilde{\boldsymbol{\Gamma}}\right)\mathbf{B}^{-1}\mathbf{m} \,, \\ \mathbf{c} &= \mathbf{W}^{-T}\left(\tilde{\boldsymbol{\Lambda}} - \tilde{\boldsymbol{\Gamma}}\right)\mathbf{B}^{-1}\mathbf{m} \,, \end{aligned} \tag{7}$$

with

$$\begin{aligned} \tilde{\boldsymbol{\Lambda}} &= \left(\mathbf{B}^{-1} + 2\mathbf{W}^{-1}\right)^{-1} \,, \quad \tilde{\boldsymbol{\Gamma}} = \left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\right)^{-1} \,, \\ \tilde{k} &= -2\log|\tilde{\boldsymbol{\Gamma}}| + \log|\mathbf{B}| + \log|\tilde{\boldsymbol{\Lambda}}| + \mathbf{m}^T\mathbf{B}^{-1}\mathbf{m} \,. \end{aligned} \tag{8}$$

[2] Our work [9] contains an error in the expressions for k and $\tilde{k}$, that was fixed in [28]. Please notice that in [9] and [28] $\mathbf{B}$ and $\mathbf{W}$ denote precision matrices, while here they denote covariance matrices.

Assuming zero-mean embeddings, $\mathbf{m} = 0$, gives

$$l(\boldsymbol{\phi}_e, \boldsymbol{\phi}_t) = 2\boldsymbol{\phi}_e^T\boldsymbol{\Lambda}\boldsymbol{\phi}_t + \boldsymbol{\phi}_e^T\boldsymbol{\Gamma}\boldsymbol{\phi}_e + \boldsymbol{\phi}_t^T\boldsymbol{\Gamma}\boldsymbol{\phi}_t + k' \,, \tag{9}$$

for some (irrelevant) constant k'. As long as embeddings have unit norm, cosine scoring can be expressed as the r.h.s. of (9) by setting $\boldsymbol{\Gamma} = \mathbf{0}, \boldsymbol{\Lambda} = \mathbf{I}, k' = 0$. Although this solution does not correspond directly to a well-formed PLDA model, it's easy to verify that an isotropic covariance, or spherical [26], model

$$\begin{aligned} P(\mathbf{y}_s) &= \mathcal{N}(\mathbf{y}_s|\mathbf{m},\boldsymbol{\gamma}^2\mathbf{I}) \,, \\ P(\boldsymbol{\phi}_{s,i}|\mathbf{y}_s) &= \mathcal{N}(\boldsymbol{\phi}_{s,i}|\mathbf{y}_s,\lambda^{-1}\mathbf{I}) \,, \end{aligned} \tag{10}$$

where γ and λ are scalars, would provide, for unit-norm embeddings, a LLR equivalent up to an affine transformation to cosine scoring [26]. Spherical PLDA had been already employed in [29, 18] as a simplification that allows deriving a Variance-Gamma-based calibration method. In [26] it was employed as a probabilistic alternative to cosine.

2.2. ISO-PLDA

The lack of a speaker subspace results in Spherical PLDA and cosine scores differing only by an affine transformation. Their performance in terms of calibration-insensitive metrics such as minimum Detection Cost Functions or Equal Error Rate is, therefore, the same. T-PSDA has shown the potential to outperform cosine thanks to a complex structure, that, similarly to classical PLDA, can include both speaker and channel subspaces. In this section we introduce an isotropic covariance variant of the simplified PLDA in (2) that also includes an explicit speaker subspace. To avoid confusion with the two-covariance-derived Spherical PLDA of [26], we call this approach Isotropic PLDA (ISO-PLDA). ISO-PLDA can be formulated as

$$\boldsymbol{\phi}_{s,i} = \mathbf{m} + \mathbf{U}\mathbf{y}_s + \boldsymbol{\varepsilon}_{s,i} \tag{11}$$

with isotropic covariance Gaussian prior and likelihood:

$$\begin{aligned} P(\mathbf{y}_s) &= \mathcal{N}(\mathbf{y}_s|\mathbf{0},\boldsymbol{\gamma}^2\mathbf{I}) \,, \\ P(\boldsymbol{\phi}_{s,i}|\mathbf{y}_s) &= \mathcal{N}(\boldsymbol{\phi}_{s,i}|\mathbf{m}+\mathbf{U}\mathbf{y}_s,\boldsymbol{\lambda}^{-1}\mathbf{I}) \,. \end{aligned} \tag{12}$$

where $\mathbf{m}$ is a mean vector, γ and λ are scalars, and $\mathbf{U}$ is a $D \times d$ matrix, constrained to have orthonormal columns $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. The constraints on $\mathbf{U}$ allows us to preserve the isotropic characteristics both in the speaker subspace and in its complement. The ISO-PLDA model can be expressed in an equivalent form that will be useful in the following as

$$\boldsymbol{\phi}_{s,i} = \mathbf{m} + \gamma\mathbf{U}\mathbf{y}_s + \boldsymbol{\varepsilon}_{s,i} \,, \tag{13}$$

with

$$\begin{aligned} P(\mathbf{y}_s) &= \mathcal{N}(\mathbf{y}_s|\mathbf{0},\mathbf{I}) \,, \\ P(\boldsymbol{\phi}_{s,i}|\mathbf{y}_s) &= \mathcal{N}(\boldsymbol{\phi}_{s,i}|\boldsymbol{\mu}+\gamma\mathbf{U}\mathbf{y}_s,\boldsymbol{\lambda}^{-1}\mathbf{I}) \end{aligned} \tag{14}$$

The EM algorithm can be employed to estimate the model parameters, with the constraints on $\mathbf{U}$ addressed similarly to those of the SG-TPSDA model described in Section 3.2.

2.3. Duration modeling for ISO-PLDA

PLDA has been extended in [19, 22] and in [20, 21] to explicitly model i-vector uncertainty, which is mainly affected by utterance duration, through an additional, utterance-dependent, additive factor. Unfortunately, most speaker embedding extractors

do not provide for an explicit model of the embedding uncertainty. In [18] the authors propose a calibration approach derived from an analysis of PLDA LLRs that allows for explicit modeling of effective between and within class variability at score level. In particular, the authors show that the scores of a spherical PLDA model can be represented in terms of Variance-Gamma (VΓ) densities whose parameters represent between and within class variances of the training and evaluation populations. This allows the authors to incorporate at calibration level an utterance-dependent term that represents both residual noise variance and variability due to utterance duration through an "equivalent", utterance dependent, effective variance

$$\lambda_{s,i}^{-1} = \lambda^{-1} + \frac{\psi}{d_{s,i} + \eta} \tag{15}$$

where λ^{-1} is an effective residual noise variance, ψ and η are model parameters, and $d_{s,i}$ is the duration of the utterance from which embedding $\boldsymbol{\phi}_{s,i}$ was extracted. Given the effectiveness of the duration-aware VΓ model in improving both calibration insensitive and calibration sensitive metrics [18], we propose to incorporate a similar model in ISO-PLDA, through an additional, isotropic covariance term:

$$\boldsymbol{\phi}_{s,i} = \mathbf{m} + \gamma \mathbf{U}\mathbf{y}_s + \boldsymbol{\delta}_{s,i} + \boldsymbol{\varepsilon}_{s,i} \ , \tag{16}$$

where

$$P(\mathbf{y}_s) = \mathcal{N}(\mathbf{y}_s|\mathbf{0}, \mathbf{I}) \ , \quad P(\boldsymbol{\delta}_{s,i}) = \mathcal{N}\left(\boldsymbol{\delta}_{s,i}|\mathbf{0}, \frac{\psi}{d_{s,i} + \eta}\mathbf{I}\right) \ ,$$
$$P(\boldsymbol{\phi}_{s,i}|\mathbf{y}_s, \boldsymbol{\delta}_s) = \mathcal{N}(\boldsymbol{\phi}_{s,i}|\mathbf{m} + \mathbf{U}\mathbf{y}_s + \boldsymbol{\delta}_{s,i}, \lambda^{-1}\mathbf{I}) \ . \tag{17}$$

The duration dependent term $\boldsymbol{\delta}_{s,i}$ and the noise term $\boldsymbol{\varepsilon}_{s,i}$ can be unified in a single, isotropic covariance, duration dependent factor, leading to the model

$$\boldsymbol{\phi}_{s,i} = \mathbf{m} + \gamma \mathbf{U}\mathbf{y}_s + \boldsymbol{\varepsilon}_{s,i} \ , \tag{18}$$

with

$$P(\mathbf{y}_s) = \mathcal{N}(\mathbf{y}_s|\mathbf{0}, \mathbf{I}) \ ,$$
$$P(\boldsymbol{\phi}_{s,i}|\mathbf{y}_s) = \mathcal{N}\left(\boldsymbol{\phi}_{s,i}|\mathbf{m} + \gamma \mathbf{U}\mathbf{y}_s, \lambda_{s,i}^{-1}\mathbf{I}\right) \ . \tag{19}$$

3. Spherical-Gaussian T-PSDA

As we show in Section 4, ISO-PLDA can significantly outperform cosine scoring, and thus Spherical PLDA. However, the T-PSDA model proves, in most cases, more effective. Since duration modeling allows improving the performance of ISO-PLDA, in the following section we show a way to combine T-PSDA with the ISO-PLDA duration model.

3.1. T-PSDA and ISO-PLDA

T-PSDA has been originally introduced in [14] as a probabilistic, subspace-constrained alternative to cosine scoring for modeling directional data (i.e. unit-norm embeddings). T-PSDA extends a previous model by the same authors, the Probabilistic Spherical Discriminant Analysis (PSDA) approach [2]. PSDA can be interpreted as a probabilistic alternative to cosine scoring that achieves very similar results. In contrast with Spherical PLDA, however, PSDA is not necessarily equivalent to cosine, even for calibration-insensitive metrics. Nevertheless, Spherical PLDA and PSDA are similar, as they both lack a speaker subspace and achieve similar performance. T-PSDA extends PSDA by re-introducing speaker and channel subspaces, employed in classical PLDA to constrain the speaker and channel factors, in the directional distribution framework of PSDA. The presence of explicit speaker subspaces proves effective, allowing T-PSDA to significantly outperform cosine scoring. T-PSDA shares many similarities with our ISO-PLDA model. To better relate the two approaches we consider a simplified T-PSDA that includes a single speaker subspace and no explicit channel subspaces. This model characterizes a unit-norm embedding in terms of von Mises-Fisher distributions as

$$P(\mathbf{y}_s) = \mathcal{V}(\mathbf{y}_s|\boldsymbol{\mu}) \ ,$$
$$P(\boldsymbol{\phi}_{s,i}|\mathbf{y}_s) = \mathcal{V}(\boldsymbol{\phi}_{s,i}|\lambda \mathbf{U}\mathbf{y}_s) \ , \tag{20}$$

where $\mathcal{V}$ denotes the von Mises-Fisher density

$$\mathcal{V}(\mathbf{u}|\mathbf{a}) = C(\|\mathbf{a}\|)e^{\mathbf{a}^T\mathbf{u}} \ , \quad C(k) = \frac{k^{\frac{p}{2}-1}}{(2\pi)^{\frac{p}{2}} I_{\frac{p}{2}-1}(k)} \tag{21}$$

defined over the $p-1$ dimensional hypersphere $\mathbb{S}^{p-1} = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\| = 1\}$, where p is the dimension of $\mathbf{x}$. $\mathbf{I}_\nu(x)$ denotes the modified Bessel function of the first kind of order $\boldsymbol{\nu}$, while $\mathbf{a} \in \mathbb{R}^p$ is a parameter. In (20) $\boldsymbol{\mu}$ is a parameter that controls the speaker prior, and $\mathbf{U}$ is a $D \times d$ speaker factor loading matrix constrained to have orthonormal columns, $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, as to keep the prior conjugate with the likelihood. We can observe that the simplified T-PSDA shares many similarities with ISO-PLDA. If we let the concentration parameter be $k = \|\boldsymbol{\mu}\| = 0$, both models represent speaker effects in terms of a subspace-constrained, isotropic speaker factor $\mathbf{y}_s$. Furthermore, as the authors of [13] observed for PSDA and Spherical PLDA, T-PSDA can be interpreted as an ISO-PLDA model where the speaker factors and the embeddings are conditioned on being unit norm. Indeed, the von Mises-Fisher density $\mathcal{V}(\lambda\boldsymbol{\mu})$ can be interpreted as the density of samples drawn from a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \lambda^{-1}\mathbf{I})$, conditioned on the samples having unit norm. If we start from a slightly modified ISO-PLDA model

$$\mathbf{y}_s \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}) \ ,$$
$$\boldsymbol{\phi}_{s,i}|\mathbf{y}_s \sim \mathcal{N}(\gamma \mathbf{U}\mathbf{y}_s, \lambda^{-1}\mathbf{I}) \ , \tag{22}$$

where we allowed for a non-zero speaker prior mean $\boldsymbol{\mu}$ (which would be irrelevant for ISO-PLDA, as it would be re-absorbed in the term $\mathbf{m}$), and further condition $\mathbf{y}_s$ and $\boldsymbol{\phi}_{s,i}$ to have unit norm, we recover

$$\mathbf{y}_s | \, \|\mathbf{y}_s\| = 1 \sim \mathcal{V}(\boldsymbol{\mu}) \tag{23}$$
$$\boldsymbol{\phi}_{s,i}|(\mathbf{y}_s, \left\|\boldsymbol{\phi}_{s,i}\right\| = 1) \sim \mathcal{V}(\lambda\gamma \mathbf{U}\mathbf{y}_s) \tag{24}$$

i.e., a simplified T-PSDA model where the likelihood concentration parameter is given by $\lambda\gamma$. It should be noted that, although similar, the two models are not equivalent even for unit-norm embeddings. Furthermore, the VMF prior corresponds to conditioning on the embeddings having unit norm, whereas, if we assume the ISO-PLDA model for non-normalized embeddings, length-normalized embeddings would be modeled by a projected Normal, rather than a VMF density. Experimental results show that ISO-PLDA is competitive with, although slightly worse than, T-PSDA.

3.2. Spherical-Gaussian TPSDA

In this Section we propose a new model that combines the flexibility of the ISO-PLDA model, which does not require embeddings to have unit norm and can easily incorporate

utterance-dependent nuisance terms, with the superior performance shown by T-PSDA. Similarly to ISO-PLDA, the proposed Spherical-Gaussian T-PSDA (SG-TPSDA) represents an embedding as a linear combination of different terms, but employs a VMF prior for the speaker factors. The model can be formulated as

$$\boldsymbol{\phi}_{s,i} = \mathbf{m} + \gamma \mathbf{U} \mathbf{y}_s + \varepsilon_{s,i} \,, \tag{25}$$

where $\mathbf{m}$ is a vector, γ a scalar and $\mathbf{U}$ a $D \times d$ factor loading matrix constrained to have orthonormal columns, $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. The speaker factor $\mathbf{y}_s$ is assumed to follow a VMF prior distribution, while the conditional likelihood for an embedding is an isotropic covariance Gaussian:

$$\begin{aligned} P(\mathbf{y}_s) &= \mathcal{V}(\mathbf{y}_s|\boldsymbol{\mu}) \,, \\ P(\boldsymbol{\phi}_{s,i}|\mathbf{y}_s) &= \mathcal{N}(\boldsymbol{\phi}_{s,i}|\mathbf{m} + \gamma\mathbf{U}\mathbf{y}_s, \lambda^{-1}\mathbf{I}) \,. \end{aligned} \tag{26}$$

We can observe that the conditional likelihood takes the same expression as our second formulation of the ISO-PLDA model in (14), whereas the speaker factor $\mathbf{y}_s$ has the same prior distribution as in T-PSDA. Given a set of n_s embeddings $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_{s,i}\}_{i=1}^{n_s}$ from speaker s we can express the joint density of the embeddings and the speaker factor $\mathbf{y}_s$ as

$$\begin{aligned} \log P(\boldsymbol{\Phi}, \mathbf{y}_s) &= \log P(\mathbf{y}_s) + \sum_{i=1}^{n_s} \log P(\boldsymbol{\phi}_{s,i}|\mathbf{y}_s) \\ &= \boldsymbol{\mu}^T\mathbf{y}_s - \frac{1}{2}\sum_{i=1}^{n_s} \lambda \left\|\boldsymbol{\phi}_{s,i} - \mathbf{m} - \gamma\mathbf{U}\mathbf{y}_s\right\|^2 + c \\ &= -\frac{1}{2}\sum_{i=1}^{n_s}\left[\lambda\left\|\boldsymbol{\phi}_{s,i} - \mathbf{m}\right\|^2\right] - \frac{1}{2}\sum_{i=1}^{n_s}\lambda\gamma^2 \\ &\quad + \sum_{i=1}^{n_s}\left[\lambda\left(\boldsymbol{\phi}_{s,i} - \mathbf{m}\right)^T\gamma\mathbf{U}\mathbf{y}_s\right] + \boldsymbol{\mu}^T\mathbf{y}_s + c \end{aligned} \tag{27}$$

where c is a constant, and we employed the fact that

$$(\gamma\mathbf{U}\mathbf{y}_s)^T(\gamma\mathbf{U}\mathbf{y}_s) = \gamma^2\mathbf{y}_s^T\mathbf{U}^T\mathbf{U}\mathbf{y}_s = \gamma^2\mathbf{y}_s^T\mathbf{y}_s = \gamma^2 \,. \tag{28}$$

The constraints on $\mathbf{U}$ allowed ISO-PLDA to present isotropic characteristics in both the speaker subspace and its complement. In SG-TPSDA the same constraints allow preserving the prior-likelihood conjugancy when the ISO-PLDA Gaussian speaker prior is replaced by the T-PSDA VMF prior. This allows us to easily derive the posterior for $\mathbf{y}_s|\boldsymbol{\Phi}$, which is also VMF:

$$P(\mathbf{y}_s|\boldsymbol{\Phi}) = \mathcal{V}\left(\mathbf{y}_s\middle|\boldsymbol{\mu} + \lambda\gamma\mathbf{U}^T\sum_{i=1}^{n_s}\left[\boldsymbol{\phi}_{s,i} - \mathbf{m}\right]\right) \tag{29}$$

We can compare the posterior (29) to the T-PSDA posterior that we would obtain with the ISO-PLDA induced parametrization of T-PSDA in (23):

$$P_{\text{T-PSDA}}(\mathbf{y}_s|\boldsymbol{\Phi}) = \mathcal{V}\left(\mathbf{y}_s\middle|\boldsymbol{\mu} + \lambda\gamma\mathbf{U}^T\sum_{i=1}^{n_s}\left[\boldsymbol{\phi}_{s,i}\right]\right) \,. \tag{30}$$

It can be noticed that the two posteriors have the same expression, as long as we assume zero-mean embeddings ($\mathbf{m} = 0$) for SG-TPSDA. On the other hand, SG-TPSDA and T-PSDA differ in the likelihood term: SG-TPSDA differentiates the between and within-speaker variability through two coefficients that act as "effective" variance γ^2 and λ^{-1}, whereas T-PSDA models them jointly through the product $\lambda\gamma$. Nevertheless, even if the ML solution for the two models may differ, our experiments show that the two models provide almost identical results for unit-norm embeddings.

As for PLDA, ISO-PLDA and T-PSDA, the EM algorithm can be employed to estimate the SG-TPSDA parameters. Given a set of n speakers with associated embedding sets $\boldsymbol{\Phi}_1 \dots \boldsymbol{\Phi}_n$, the EM auxiliary function is given by

$$\begin{aligned} &Q(\boldsymbol{\Pi}, \boldsymbol{\Pi}') = \\ &= \mathbb{E}_{P(\mathbf{y}_1 \dots \mathbf{y}_n|\boldsymbol{\Phi}_1 \dots \boldsymbol{\Phi}_n, \boldsymbol{\Pi}')}\left[\log P(\boldsymbol{\Phi}_1 \dots \boldsymbol{\Phi}_n, \mathbf{y}_1, \dots \mathbf{y}_n|\boldsymbol{\Pi})\right] \\ &= \sum_{s=1}^{n} \mathbb{E}_{P(\mathbf{y}_s|\boldsymbol{\Phi}_s, \boldsymbol{\Pi}')}\left[\log P(\boldsymbol{\Phi}_s|\mathbf{y}_s, \boldsymbol{\Pi}) \log P(\mathbf{y}_s|\boldsymbol{\Pi})\right] \\ &= \frac{ND}{2}\log\lambda - \frac{1}{2}\sum_{s=1}^{n}\sum_{i=1}^{n_s}\left[\lambda\left\|\boldsymbol{\phi}_{s,i} - \mathbf{m}\right\|^2\right] - \frac{1}{2}\lambda\gamma^2 N \\ &\quad + \lambda\gamma\,\mathrm{Tr}\left[\mathbf{U}\sum_{s=1}^{n}\mathbb{E}\left[\mathbf{y}_s\right]\sum_{i=1}^{n_s}\left(\boldsymbol{\phi}_{s,i} - \mathbf{m}\right)^T\right] \,, \end{aligned} \tag{31}$$

where $\boldsymbol{\Pi}' = (\mathbf{m}', \mathbf{U}', \lambda', \gamma', \boldsymbol{\mu}')$ are the current estimates of the model parameters, $\boldsymbol{\Pi} = (\mathbf{m}, \mathbf{U}, \lambda, \gamma, \boldsymbol{\mu})$ are the parameters we are maximizing Q with respect to, and N is the total number of utterances $N = \sum_{s=1}^{n} n_s$. As for T-PSDA [14] the maximization of Q can be performed in steps with a coordinate ascent approach. The maximum with respect to $\mathbf{m}$ is given by

$$\mathbf{m} = \frac{1}{N}\sum_{s=1}^{n}\sum_{i=1}^{n_s}\left(\boldsymbol{\phi}_{s,i} - \gamma\mathbf{U}\mathbb{E}\left[\mathbf{y}_s\right]\right) \,, \tag{32}$$

although in practice, if embeddings have been centered, assuming $\mathbf{m} = 0$ does not significantly affect results. The maximizer of Q with respect to $\mathbf{U}$ corresponds to the maximizer of

$$\mathrm{Tr}\left[\mathbf{U}\mathbf{A}\right], \; \mathbf{A} = \sum_{s=1}^{n}\mathbb{E}\left[\mathbf{y}_s\right]\sum_{i=1}^{n_s}\left(\boldsymbol{\phi}_{s,i} - \mathbf{m}\right)^T, \; s.t.\; \mathbf{U}^T\mathbf{U} = \mathbf{I} \,. \tag{33}$$

From von Neumann trace inequality the maximum is obtained by $\mathbf{U} = \mathbf{V_A}\mathbf{U_A^T}$, where $\mathbf{U}_A$ and $\mathbf{V}_A$ are given by the truncated Singular Value Decomposition (SVD) of $\mathbf{A} = \mathbf{U_A}\boldsymbol{\Sigma_A}\mathbf{V_A^T}$. The maximum with respect to λ and γ is obtained by

$$\begin{aligned} \gamma &= \frac{1}{N}\sum_{s=1}^{n}\left[\sum_{i=1}^{n_s}\left(\boldsymbol{\phi}_{s,i} - \mathbf{m}\right)^T\right]\mathbf{U}\mathbb{E}\left[\mathbf{y}_s\right] \,, \\ \lambda &= \frac{1}{ND}\left[\sum_{s=1}^{n}\sum_{i=1}^{n_s}\left\|\boldsymbol{\phi}_{s,i} - \mathbf{m}\right\|^2 - N\gamma^2\right] \,. \end{aligned} \tag{34}$$

Regarding $\boldsymbol{\mu}$, we express the parameter as

$$\boldsymbol{\mu} = k\widehat{\boldsymbol{\mu}} \,, \quad \|\widehat{\boldsymbol{\mu}}\| = 1 \,. \tag{35}$$

The optimal solution for $\widehat{\boldsymbol{\mu}}$ corresponds to

$$\widehat{\boldsymbol{\mu}} = \frac{\boldsymbol{\eta}}{\|\boldsymbol{\eta}\|} \,, \quad \boldsymbol{\eta} = \sum_{s=1}^{n}\mathbb{E}\left[\mathbf{y}_s\right] \,, \tag{36}$$

whereas the solution for k can be obtained through numerical maximization of

$$\mathcal{L}(k) = \sum_{s=1}^{n}\left[\left(\frac{d}{2} - 1\right)\log k - \log I_{\frac{d}{2}-1}(k) + k\widehat{\boldsymbol{\mu}}^T\mathbb{E}\left[\mathbf{y}_s\right]\right] \,. \tag{37}$$

The speaker verification LLR for a pair of embeddings $(\boldsymbol{\phi}_e, \boldsymbol{\phi}_t)$ can be expressed as [30, 14]

$$l(\boldsymbol{\phi}_e, \boldsymbol{\phi}_t) = \log \frac{P(\mathbf{y}|\boldsymbol{\phi}_e)P(\mathbf{y}|\boldsymbol{\phi}_t)}{P(\mathbf{y}|\boldsymbol{\phi}_e, \boldsymbol{\phi}_t)P(\mathbf{y})} = \log \frac{\mathcal{V}(\mathbf{y}|\boldsymbol{\mu}_e)\mathcal{V}(\mathbf{y}|\boldsymbol{\mu}_t)}{\mathcal{V}(\mathbf{y}|\boldsymbol{\mu}_{e,t})\mathcal{V}(\mathbf{y}|\boldsymbol{\mu})} \tag{38}$$

for any unit vector $\mathbf{y}$, where $\boldsymbol{\mu}_e$, $\boldsymbol{\mu}_t$ and $\boldsymbol{\mu}_{e,t} = \boldsymbol{\mu}_e + \boldsymbol{\mu}_t - \boldsymbol{\mu}$ are the parameters of the speaker factor posterior (29) given the enrollment, the test, and both embeddings, respectively. Choosing $\mathbf{y}$ orthogonal to all the prior and posterior parameter vectors in (38) we obtain

$$l(\boldsymbol{\phi}_e, \boldsymbol{\phi}_t) = \log \frac{C(\|\boldsymbol{\mu}_e\|)C(\|\boldsymbol{\mu}_t\|)}{C(\|\boldsymbol{\mu}_e + \boldsymbol{\mu}_t - \boldsymbol{\mu}\|)C(\|\boldsymbol{\mu}\|)} \tag{39}$$

i.e., as expected from the form of the posterior, the same expression as the T-PSDA LLR [14].

3.3. SG-TPSDA and duration modeling

We now have all the building blocks to derive our duration-aware SG-TPSDA model. As for SG-TPSDA, we combine the T-PSDA prior with the duration-aware ISO-PLDA likelihood:

$$\boldsymbol{\phi}_{s,i} = \mathbf{m} + \gamma \mathbf{U}\mathbf{y}_s + \boldsymbol{\varepsilon}_{s,i} \ , \tag{40}$$

where $\mathbf{y}_s$ follows a VMF prior distribution, while the residual noise $\boldsymbol{\varepsilon}_{s,i}$ has a duration-dependent, isotropic covariance, Gaussian distribution as in (19):

$$\begin{aligned} P(\mathbf{y}_s) &= \mathcal{V}(\boldsymbol{\mu}) \ , \\ P(\boldsymbol{\phi}_{s,i}|\mathbf{y}_s) &= \mathcal{N}\left(\boldsymbol{\phi}_{s,i}\middle|\mathbf{m} + \gamma\mathbf{U}\mathbf{y}_s, \lambda_{s,i}^{-1}\mathbf{I}\right) \ , \end{aligned} \tag{41}$$

with $\lambda_{s,i}$ defined as in (15):

$$\lambda_{s,i}^{-1} = \lambda^{-1} + \frac{\psi}{d_{s,i} + \eta} \tag{42}$$

The posterior for $\mathbf{y}_s$ takes a similar form as in SG-TPSDA, with the precision term λ replaced by the effective, utterance dependent precision term $\lambda_{s,i}$:

$$P(\mathbf{y}_s|\boldsymbol{\Phi}) = \mathcal{V}\left(\mathbf{y}_s\middle|\boldsymbol{\mu} + \gamma\mathbf{U}^T\sum_{i=1}^{n_s}\lambda_{s,i}\left[\boldsymbol{\phi}_{s,i} - \mathbf{m}\right]\right) \tag{43}$$

The model parameters $\mathbf{U}, \mathbf{m}, \gamma, \boldsymbol{\mu}$ can be estimated similarly as in the SG-TSPDA model. Replacing the terms λ with $\lambda_{s,i}$ in the likelihood (27) we can derive the EM auxiliary function and obtain the update rule for $\mathbf{m}$ as

$$\mathbf{m} = \frac{\sum_{s=1}^{n}\sum_{i=1}^{n_s}\lambda_{s,i}\left(\boldsymbol{\phi}_{s,i} - \gamma\mathbf{U}\mathbb{E}\left[\mathbf{y}_s\right]\right)}{\sum_{s=1}^{n}\sum_{i=1}^{n_s}\lambda_{s,i}} \ . \tag{44}$$

$\mathbf{U}$ can be computed from the truncated SVD

$$\mathbf{A} = \sum_{s=1}^{n}\mathbb{E}\left[\mathbf{y}_s\right]\sum_{i=1}^{n_s}\lambda_{s,i}\left(\boldsymbol{\phi}_{s,i} - \mathbf{m}\right)^T \tag{45}$$

as $\mathbf{U} = \mathbf{V}_\mathbf{A}\mathbf{U}_\mathbf{A}^T$. The update for γ is given by

$$\gamma = \frac{\sum_{s=1}^{n}\left[\sum_{i=1}^{n_s}\lambda_{s,i}\left(\boldsymbol{\phi}_{s,i} - \mathbf{m}\right)^T\right]\mathbf{U}\mathbb{E}\left[\mathbf{y}_s\right]}{\sum_{s=1}^{n}\sum_{i=1}^{n_s}\lambda_{s,i}} \tag{46}$$

The optimal solution for $\boldsymbol{\mu}$ can be computed as in (35), (36) and (37). For the parameters λ, ψ and η of the Gaussian noise, and the term k in (37), we resort to numerical optimization based on the L-BFGS algorithm [31]. The speaker verification LLR can be computed as in (38) and (39), with the posterior parameters computed according to (43).

4. Experimental results

To assess the effectiveness of our proposed approach we consider three different test sets: NIST SRE19[3], NIST SRE24[4] and CN-Celeb [32]. Since the duration-aware models are aimed at scenarios with utterances of short and variable duration, for SRE19 we consider both the original and a "short" version of the benchmark obtained by randomly cutting enrollment and test segments to a duration between 3 and 30 seconds. Since both SRE24 and CN-Celeb already contain shorter segments, this procedure was not applied to these two sets. We employ two off-the-shelf, publicly available embedding extractors: a SimAM-ResNet-100[5] [33], pre-trained and fine-tuned on VoxBlink2 [34] and VoxCeleb2 [35], respectively, and a ReDimNet-B6[6] [36], pre-trained on VoxCeleb2. Speech segments have been extracted by means of energy-based Voice Activity Detection. Back-end training employs NIST CTS Superset [37] data for the SRE19 and SRE24 tests and VoxCeleb2 data for the CN-Celeb test, with a mix of short and long segments, obtained concatenating training segments belonging to a single session. Results are summarized in Table 1 for SimAM-ResNet-100 and Table 2 for ReDimNet-B6 embeddings, respectively. For all models we report Equal Error Rate (EER), minimum primary cost (min C_{prim}) as defined by NIST for SRE19 and SRE24, and the minimum cost of Log-Likelihood Ratio (min C_{llr}) [38, 39, 40], evaluated through the PAV algorithm [41] on the evaluation set. In each table the first two rows report the performance of cosine and T-PSDA baselines, with the best-performing baseline in bold. The following four rows refer to ISO-PLDA models, with and without length normalization ("Norm" column), with and without duration modeling ("Dur" column), whereas the last four rows report the SG-TPSDA performance with a similar organization. Results that match or outperform the best baseline are in bold. For each test set / front-end combination the subspace dimension was selected by optimizing the baseline T-PSDA results.

We start observing that T-PSDA is more effective than cosine scoring in almost all scenarios, in particular for the SimAM-ResNet-100 embeddings, where the model can even halve the EER on SRE24. For ReDimNet-B6 T-PSDA is also more effective, although the relative gains are smaller, and cosine scoring is able to match T-PSDA on CN-Celeb data. Regarding our proposed models, we note that SG-TPSDA with length-normalized embeddings achieves very similar performance as T-PSDA, confirming that the Gaussian likelihood can effectively be employed in place of the VMF likelihood. ISO-PLDA is also competitive when embeddings are normalized, providing slightly worse performance than T-PSDA but significantly outperforming cosine for SimAM-ResNet-100, and matching or providing minor improvement over T-PSDA for ReDimNet-B6. For both ISO-PLDA and SG-TPSDA the introduction of our duration model leads to a significant improvement in terms of EER and minimum C_{llr}, while minimum C_{prim} results are not significantly affected, except for ISO-PLDA on CN-Celeb, where we observe also an improvement for this metric.

In contrast with T-PSDA, both ISO-PLDA and SG-TPSDA can process non-normalized embeddings. Previous works have

[3] https://www.nist.gov/itl/iad/mig/nist-2019-speaker-recognition-evaluation

[4] https://www.nist.gov/itl/iad/mig/nist-2024-speaker-recognition-evaluation-sre24

[5] https://github.com/wenet-e2e/wespeaker/blob/master/docs/pretrained.md

[6] https://github.com/IDRnD/redimnet/blob/master/EVALUATION.md

Table 1: *Percent EER, min C_{prim} and min C_{llr} of the proposed back-ends, with and without length normalization, with and without duration modeling, SimAM-ResNet-100. The first two rows correspond to our baselines, with the results of the best system in bold. The following rows present ISO-PLDA and SG-TPSDA results, with bold values for models that match or outperform the best baseline.*

Backend	Norm	Dur	SRE 19 (Cuts) EER %	SRE 19 (Cuts) min C_{prim}	SRE 19 (Cuts) min C_{llr}	SRE 19 (Original) EER %	SRE 19 (Original) min C_{prim}	SRE 19 (Original) min C_{llr}	SRE 24 EER %	SRE 24 min C_{prim}	SRE 24 min C_{llr}	CNCeleb EER %	CNCeleb min C_{prim}	CNCeleb min C_{llr}
Cosine	✓	✗	12.9	0.61	0.43	6.2	0.37	0.23	14.3	0.81	0.46	9.9	0.39	0.32
T-PSDA	✓	✗	**9.2**	**0.55**	**0.33**	**4.4**	**0.33**	**0.18**	**7.1**	**0.58**	**0.26**	**6.7**	**0.36**	**0.24**
ISO-PLDA	✓	✗	10.0	0.56	0.35	4.8	0.34	0.19	7.5	0.60	0.28	7.2	0.37	0.26
		✓	**8.9**	**0.54**	**0.32**	4.6	**0.33**	**0.18**	7.5	0.59	0.28	**6.6**	**0.34**	**0.24**
	✗	✗	10.2	0.65	0.35	4.7	0.38	**0.18**	8.8	0.70	0.31	7.6	0.47	0.26
		✓	**8.0**	0.59	**0.28**	**4.4**	0.37	**0.17**	8.3	0.67	0.30	**6.6**	0.42	**0.22**
SG-TSPDA	✓	✗	**9.2**	**0.55**	**0.33**	**4.4**	**0.33**	**0.18**	**7.1**	**0.58**	**0.26**	**6.6**	**0.36**	**0.24**
		✓	**8.3**	0.56	**0.31**	**4.1**	**0.33**	**0.17**	**6.8**	**0.57**	**0.25**	**6.3**	**0.36**	**0.22**
	✗	✗	**8.7**	**0.55**	**0.31**	**4.1**	**0.33**	**0.16**	**6.9**	0.61	**0.26**	**5.9**	**0.36**	**0.21**
		✓	**8.3**	0.57	**0.30**	**4.1**	0.34	**0.16**	**6.5**	**0.58**	**0.24**	**5.5**	0.37	**0.19**

Table 2: *Percent EER, min C_{prim} and min C_{llr} of the proposed back-ends, with and without length normalization, with and without duration modeling, ReDimNet-B6. The first two rows correspond to our baselines, with the results of the best system in bold. The following rows present ISO-PLDA and SG-TPSDA results, with bold values for models that match or outperform the best baseline.*

Backend	Norm	Dur	SRE 19 (Cuts) EER %	SRE 19 (Cuts) min C_{prim}	SRE 19 (Cuts) min C_{llr}	SRE 19 (Original) EER %	SRE 19 (Original) min C_{prim}	SRE 19 (Original) min C_{llr}	SRE 24 EER %	SRE 24 min C_{prim}	SRE 24 min C_{llr}	CNCeleb EER %	CNCeleb min C_{prim}	CNCeleb min C_{llr}
Cosine	✓	✗	13.4	0.74	0.46	8.1	0.54	0.30	13.0	0.84	0.45	**11.2**	**0.49**	**0.37**
T-PSDA	✓	✗	**12.7**	**0.72**	**0.43**	**7.6**	**0.52**	**0.28**	**11.1**	**0.77**	**0.39**	**11.2**	**0.49**	**0.37**
ISO-PLDA	✓	✗	**12.5**	**0.72**	**0.43**	**7.6**	**0.52**	**0.28**	**10.8**	**0.77**	**0.39**	**11.1**	**0.49**	**0.37**
		✓	**11.6**	**0.70**	**0.40**	**7.3**	**0.51**	**0.27**	11.4	**0.76**	0.40	**10.5**	**0.43**	**0.34**
	✗	✗	13.9	0.74	0.46	7.9	0.53	**0.28**	12.5	0.83	0.43	12.8	0.52	0.42
		✓	**11.6**	**0.71**	**0.40**	**7.5**	**0.52**	**0.27**	12.4	0.80	0.42	**9.9**	**0.43**	**0.32**
SG-TSPDA	✓	✗	**12.7**	**0.72**	**0.43**	**7.6**	**0.52**	**0.28**	**11.1**	**0.77**	**0.39**	**11.2**	**0.49**	**0.37**
		✓	**11.8**	0.73	**0.41**	**7.4**	0.53	**0.27**	**10.9**	0.78	**0.38**	**10.0**	**0.47**	**0.34**
	✗	✗	13.1	0.73	0.44	**7.6**	**0.52**	**0.28**	11.6	0.80	0.41	11.5	0.50	0.38
		✓	**11.9**	0.74	**0.41**	**7.4**	0.53	**0.27**	**11.0**	0.78	**0.39**	**9.7**	**0.47**	**0.33**

suggested that length normalization may be less relevant for robust back-ends, especially when duration is accounted for at back-end or calibration level [18]. We therefore analyzed also the performance of our approaches without length normalization. For ISO-PLDA we can observe a degradation of performance when duration is not accounted for in almost all test sets for both front-ends. SG-TPSDA also shows a small degradation of performance in three out of four test sets with ReDimNet-B6 embeddings. On the other hand, with SimAM-ResNet-100 embeddings we observe improvements in terms of EER and C_{llr} when length normalization is not applied. The introduction of duration models results in significant gains in terms of EER and C_{llr} for both approaches, except for the original SRE19 test, where, due to the longer duration of the segments, the improvement is less significant. We can observe that, once duration is accounted for, the gap between models that employ unit-norm embeddings and models that employ the original, non-normalized embeddings reduce, confirming that length-normalization may be indirectly achieving a duration normalization effect.

5. Conclusions

We have presented a novel generative back-end for embedding-based speaker verification. SG-TPSDA combines the structured subspace and VMF speaker prior of T-PSDA with the Gaussian likelihood of an isotropic covariance PLDA model, achieving similar performance as T-PSDA with unit-norm embeddings, while being able, in some scenarios, to outperform T-PSDA when length normalization is not applied. We have also introduced an extension of the Gaussian likelihood that is able to incorporate a duration model that does not rely on ad-hoc embedding extractors, allowing us to further improve the SG-TPSDA performance. The model similarities and the relative performance of our ISO-PLDA and SG-TPSDA suggest that the structure of the speaker factor plays an important role. Future works will be devoted to extending SG-TPSDA to include structured duration information, evaluate a duration-aware T-PSDA model that can be constructed from the SG-TPSDA duration-aware model, and to further investigation of the structure of the speaker factor and its prior, with the goal of unifying ISO-PLDA, T-PSDA and SG-TPSDA.

6. References

[1] Q. Wang, K. A. Lee, and T. Liu, "Scoring of Large-Margin Embeddings for Speaker Verification: Cosine or PLDA?" in *Interspeech 2022*, 2022, pp. 600–604.

[2] Niko Brümmer and Albert Swart and Ladislav Mosner and Anna Silnova and Oldřich Plchot and Themos Stafylakis and Lukáš Burget, "Probabilistic spherical discriminant analysis: An alternative to PLDA for length-normalized embeddings," in *Interspeech 2022*, 2022, pp. 1446–1450.

[3] S. Malykh, A. Anikin, N. Khmelev, A. Korenevskaya, A. Zorkina, S. Novoselov, V. Marchevskiy, V. Volokhov, A. Shulipa, A. Kozlov, A. Melnikov, V. Galyuk, and T. Pekhovskiy, "STCON NIST SRE24 System: Composite Speaker Recognition Solution for Challenging Scenarios," in *Interspeech 2025*, 2025, pp. 3983–3987.

[4] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proceedings of the 9th European Conference on Computer Vision*, ser. ECCV'06, vol. Part IV, 2006, pp. 531–542.

[5] S. J. D. Prince and J. H. Elder, "Probabilistic Linear Discriminant Analysis for inferences about identity," in *Proceedings of 11th International Conference on Computer Vision*, 2007, pp. 1–8.

[6] P. Kenny, "Bayesian speaker verification with Heavy-Tailed Priors," in *The Speaker and Language Recognition Workshop (Odyssey 2010)*, 2010.

[7] J. Villalba, B. J. Borgstrom, S. Kataria, M. Rybicka, C. D. Castillo, J. Cho, L. P. García-Perera, P. A. Torres-Carrasquillo, and N. Dehak, "Advances in Cross-Lingual and Cross-Source Audio-Visual Speaker Recognition: The JHU-MIT System for NIST SRE21," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 213–220.

[8] S. Cumani, A. Silnova, S. Barahona, L. Mošner, O. Plchot, and J. Rohdin, "Analysis of the ABC classification backends for NIST SRE24," in *Interspeech 2025*, 2025, pp. 3978–3982.

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

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

[11] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[12] N. Brümmer and E. de Villiers, "The speaker partitioning problem," in *The Speaker and Language Recognition Workshop (Odyssey 2010)*, 2010, pp. 194–201.

[13] A. Sholokhov, N. Kuzmin, K. A. Lee, and E. S. Chng, "Probabilistic back-ends for online speaker recognition and clustering," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[14] A. Silnova, N. Brümmer, A. Swart, and L. Burget, "Toroidal probabilistic spherical discriminant analysis," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[15] M. I. Mandasari, R. Saeidi, M. McLaren, and D. A. van Leeuwen, "Quality measure functions for calibration of speaker recognition systems in various duration conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2425–2438, 2013.

[16] M. I. Mandasari, R. Saeidi, and D. A. van Leeuwen, "Quality measures based calibration with duration and noise dependency for speaker recognition," *Speech Communication*, vol. 72, pp. 126–137, 2015.

[17] A. Nautsch, R. Saeidi, C. Rathgeb, and C. Busch, "Robustness of quality-based score calibration of speaker recognition systems with respect to low-snr and short-duration conditions," in *Proceedings of Odyssey 2016*, 2016.

[18] S. Cumani and S. Sarni, "The distributions of uncalibrated speaker verification scores: A generative model for domain mismatch and trial-dependent calibration," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2204–2219, 2023.

[19] S. Cumani, O. Plchot, and P. Laface, "Probabilistic Linear Discriminant Analysis of i–vector posterior distributions," in *Proceedings of ICASSP 2013*, 2013, pp. 7644–7648.

[20] P. Kenny, T. Stafylakis, P. Ouellet, J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Proceedings of ICASSP 2013*, 2013, pp. 7649–7653.

[21] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P. Dumouchel, "Text-dependent speaker recognition using PLDA with uncertainty propagation," in *Proceedings of INTERSPEECH 2013*, 2013, pp. 3684–3688.

[22] S. Cumani, O. Plchot, and P. Laface, "On the use of i-vector posterior distributions in Probabilistic Linear Discriminant Analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 846–857, 2014.

[23] S. Cumani, "Fast scoring of full posterior PLDA models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 2036–2045, 2015.

[24] K. A. Lee, Q. Wang, and T. Koshinaka, "Xi-vector embedding for speaker recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 1385–1389, 2021.

[25] Q. Wang, K. A. Lee, and T. Liu, "Incorporating uncertainty from speaker embedding estimation to speaker verification," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[26] Q. Wang and K. A. Lee, "Cosine scoring with uncertainty for neural speaker embedding," *IEEE Signal Processing Letters*, vol. 31, pp. 845–849, 2024.

[27] S. Cumani, O. Glembek, N. Brümmer, E. Villiers, and P. Laface, "Gender independent discriminative speaker recognition in i-vector space," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4361–4364.

[28] L. Ferrer, M. McLaren, and N. Brümmer, "A speaker verification backend with robust performance across conditions," *Computer Speech & Language*, vol. 71, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230821000656

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

[30] J. BESAG, "A candidate's formula: A curious result in bayesian prediction," *Biometrika*, vol. 76, no. 1, pp. 183–183, 03 1989.

[31] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Math. Program.*, vol. 45, no. 1–3, pp. 503–528, Aug. 1989.

[32] L. Li, X. Li, H. Jiang, C. Chen, R. Hou, and D. Wang, "Cn-celeb-av: A multi-genre audio-visual dataset for person recognition," in *Interspeech 2023*, 2023, pp. 2118–2122.

[33] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 11 863–11 874.

[34] Y. Lin, M. Cheng, F. Zhang, Y. Gao, S. Zhang, and M. Li, "VoxBlink2: A 100k+ speaker recognition corpus and the open-set speaker-identification benchmark," in *Interspeech 2024*, 2024, pp. 4263–4267.

[35] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Interspeech 2018*, 2018, pp. 1086–1090.

[36] I. Yakovlev, R. Makarov, A. Balykin, P. Malov, A. Okhotnikov, and N. Torgashov, "Reshape Dimensions Network for Speaker Recognition," in *Interspeech 2024*, 2024, pp. 3235–3239.

[37] O. Sadjadi, "NIST SRE CTS superset: A large-scale dataset for telephony speaker recognition," 2021. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=933116

[38] N. Brümmer and J. A. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

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

[40] D. Van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," *Lecture Notes in Computer Science*, vol. 4343, pp. 330–353, 01 2007.

[41] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 08 2002.