Interspeech 2024
1-5 September 2024, Kos, Greece

Revisiting and Improving Scoring Fusion for Spoofing-aware Speaker Verification Using Compositional Data Analysis

Xin Wang[1], *Tomi Kinnunen*[2], *Kong Aik Lee*[3], *Paul-Gauthier Noé*[4], *Junichi Yamagishi*[1]

[1]National Institute of Informatics, Japan, [2]School of Computing, University of Eastern Finland, Finland, [3]Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong, [4]Laboratoire Informatique d'Avignon, Avignon Université, France

wangxin@nii.ac.jp

Abstract

Fusing outputs from automatic speaker verification (ASV) and spoofing countermeasure (CM) is expected to make an integrated system robust to zero-effort imposters and synthesized spoofing attacks. Many score-level fusion methods have been proposed, but many remain heuristic. This paper revisits score-level fusion using tools from decision theory and presents three main findings. First, fusion by summing the ASV and CM scores can be interpreted on the basis of compositional data analysis, and score calibration before fusion is essential. Second, the interpretation leads to an improved fusion method that linearly combines the log-likelihood ratios of ASV and CM. However, as the third finding reveals, this linear combination is inferior to a non-linear one in making optimal decisions. The outcomes of these findings, namely, the score calibration before fusion, improved linear fusion, and better non-linear fusion, were found to be effective on the SASV challenge database.

Index Terms: speaker verification, anti-spoofing, fusion, log-likelihood ratio, ternary classification

1. Introduction

Automatic speaker verification (ASV) systems are vulnerable to spoofed synthetic speech data that clone the target speakers' voices [1]. By combining ASV and spoofing countermeasure (CM), which ideally should reject any spoofed data, we expect to be able to construct a spoofing-aware ASV (SASV) system that only accepts speech data from real (bona fide) humans who match the target speaker's identity.

An SASV system can be a cascade of ASV and CM sub-systems [2]. The ASV and CM make separate binary decisions, and the input is accepted if it is accepted by both sub-systems. Another design used by many SASV systems is to produce one score and make one binary decision. Let the input be $\boldsymbol{x} = (\boldsymbol{x}^{(p)}, \boldsymbol{x}^{(r)}) \in \mathcal{X}$, where $\boldsymbol{x}^{(p)}$ and $\boldsymbol{x}^{(r)}$ are the probe (test) and reference (enrollment) samples, respectively, and can be waveforms, features, etc. An SASV system defines a scoring function $g : \mathcal{X} \rightarrow \mathbb{R}$ that maps $\boldsymbol{x}$ into a score s_{sasv}. If s_{sasv} surpasses a threshold, the system takes the action of `accept`, claiming that $\boldsymbol{x}_p$ is bona fide *and* matches the speaker identity in $\boldsymbol{x}_r$. Otherwise, it takes the action of `reject`.

This paper focuses on the SASV systems that produce s_{sasv} using score-level fusion. The idea is to use ASV and CM sub-systems to obtain an ASV score s_{asv} and CM score s_{cm} then use a score-level fusion function $h : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ to merge the scores into s_{sasv}. Note that there are feature-level fusion and end-to-end methods [6, 7, 3], which do not need to produce $\{s_{\text{asv}}, s_{\text{cm}}\}$. We argue that score-level fusion is useful because $\{s_{\text{asv}}, s_{\text{cm}}\}$ provides invaluable extra explanations on the model decision. Some evaluation metrics, such as tandem equal error rate (t-EER) [8], also require separated s_{asv} and s_{cm}.

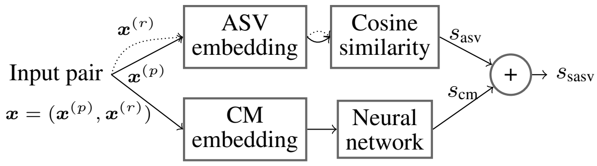


Figure 1: *Example SASV with score-level fusion (B1 in [3])*

An example SASV system using score-level fusion (baseline B1 of the SASV challenge [3]) is illustrated in Fig. 1. Its ASV sub-system extracts embeddings from $\boldsymbol{x}^{(p)}$ and $\boldsymbol{x}^{(r)}$ and computes their cosine similarity as s_{asv}, which indicates the degree that $\boldsymbol{x}^{(p)}$ and $\boldsymbol{x}^{(r)}$ match in terms of speaker identity. The CM sub-system extracts an embedding from $\boldsymbol{x}^{(p)}$ and produces a score $s_{\text{cm}} \in \mathbb{R}$ using a neural network. A higher s_{cm} indicates that $\boldsymbol{x}^{(p)}$ is more likely to be bona fide. Hence, it is intuitively reasonable to fuse the scores by $s_{\text{sasv}} = s_{\text{asv}} + s_{\text{cm}}$.

More advanced fusion functions have been proposed in the SASV and other biometrics fields [4, 2]. Although most are intuitively reasonable, for example, normalizing the numeric range of s_{asv} and s_{cm} before summation [3], the link between practice and theory is not always clear. Therefore, the implicit assumptions, link to optimal decisions, and room for improvement remain elusive.

Starting from the summation of the ASV and CM scores, we scrutinize the score fusion using tools from decision theory. After reformulating SASV as a classification task that involves three data classes but requires a binary decision, we interpret SASV scoring from the perspective of compositional data analysis and show that the summation of s_{asv} and s_{cm} is a reasonable choice if the ASV and CM scores are proper log-likelihood ratios (LLRs). Hence, the interpretation indicates that score calibration before fusion is beneficial. The interpretation also leads to an improved fusion method that uses probabilistic models to compute LLRs for summation. Finally, the summation-based fusion is analyzed from the perspective of optimal decision policy and found to be theoretically inferior to a non-linear fusion method.

The above interpretations of SASV have not been presented as far as the authors are aware. They differ from one existing study that treats $\{s_{\text{asv}}, s_{\text{cm}}\}$ as posterior probabilities [5]. Experiments were conducted on the SASV challenge database [3]. The calibration before fusion was found to be effective, especially for linear fusion by summation. The linear and non-linear fusion of LLRs outperformed baselines, and the latter outperformed the systems fusing posterior probabilities [5]. The results are reproducible. [1]

[1]Code site: https://github.com/nii-yamagishilab/SpeechSPC-mini.

10.21437/Interspeech.2024-422

2. Interpreting and improving SASV score fusion

2.1. SASV is not fusion of congruent binary classifiers

By definition, an SASV system should only `accept` if the input is uttered by the bona fide target speaker; otherwise, `reject`. Fusing ASV and CM scores for a binary decision in the SASV system seems to be similar to the score fusion in multi-modal biometric verification [9, 10].

Let $\mathbf{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_K\}$ be the inputs to a biometric verification system with K modalities (e.g., speech and face). The system determines whether inputs match the claimed identity (H_{tar}) or not (H_{non}). Assuming that $\boldsymbol{x}_k$ are statistically independent, we can use the Bayes' formula [11] to obtain

$$\log \frac{P(H_{\mathsf{tar}}|\mathbf{X})}{1 - P(H_{\mathsf{tar}}|\mathbf{X})} = \log \frac{\pi_{\mathsf{tar}}}{1 - \pi_{\mathsf{tar}}} + \sum_{k=1}^{K} \mathsf{llr}_{\mathsf{non}}^{\mathsf{tar}}(\boldsymbol{x}_k), \quad (1)$$

where $\mathsf{llr}_{\mathsf{non}}^{\mathsf{tar}}(\boldsymbol{x}_k) \triangleq \log \frac{p(\boldsymbol{x}_k|H_{\mathsf{tar}})}{p(\boldsymbol{x}_k|H_{\mathsf{non}})}$ is an LLR produced by the k-th sub-system, and π_{tar} is the prior probability of class tar. Assuming that the LLRs and priors are equal to the ground truth, our decision strategy for minimizing the decision error rate is to accept tar if and only if $\sum_{k=1}^{K} \mathsf{llr}_{\mathsf{non}}^{\mathsf{tar}}(\boldsymbol{x}_k) + \log \frac{\pi_{\mathsf{tar}}}{1-\pi_{\mathsf{tar}}} > 0$ [11]. The sum of LLRs is the weight of evidence to favor tar, which is known as *independent additivity* of LLRs [12].

We may hastily treat SASV as biometric verification with two modalities, namely, ASV and CM. Thus, fusion by $s_{\text{cm}} + s_{\text{asv}}$ can be interpreted as the sum of LLRs, given that the scores approximate the LLRs. However, unlike Eq. (1) where all sub-systems deal with $\{H_{\mathsf{tar}}, H_{\mathsf{non}}\}$, ASV and CM have different hypotheses: for CM, the input is either bona fide (H_{bona}) or spoofed (H_{spoof}). With four hypotheses involved, independent additivity of LLRs and Eq. (1) no longer hold [12, sec.4.3].

2.2. Interpreting SASV using compositional data analysis

A better starting point is to have three exhaustive and mutually exclusive hypotheses [2, 13]: the input probe is bona fide and matches the reference ($H_{\mathsf{tar.bf}}$), bona fide and unmatched ($H_{\mathsf{non.bf}}$), or spoofed (H_{spf}). [2] To interpret this classification task with three classes but two actions $\{\texttt{accept}, \texttt{reject}\}$, we use *compositional data analysis* [14, 15].

Let $\boldsymbol{P} = [P(H_{\mathsf{spf}}|\boldsymbol{x}), P(H_{\mathsf{non.bf}}|\boldsymbol{x}), P(H_{\mathsf{tar.bf}}|\boldsymbol{x})]^\top$ and $\boldsymbol{\pi} = [\pi_{\mathsf{spf}}, \pi_{\mathsf{non.bf}}, \pi_{\mathsf{tar.bf}}]^\top$ be the vectors of posterior and prior probabilities, respectively. Let the vector of likelihoods be $\boldsymbol{w} = [p(\boldsymbol{x}|H_{\mathsf{spf}}), p(\boldsymbol{x}|H_{\mathsf{non.bf}}), p(\boldsymbol{x}|H_{\mathsf{tar.bf}})]^\top$. Because $\sum_i \pi_i = \sum_i P_i = 1$, where i is the index of the vector's element, $\boldsymbol{\pi}$ and $\boldsymbol{P}$ are compositional vectors on the probability simplex $\mathbb{S}^3$. Hence, $\boldsymbol{P}$ (and $\boldsymbol{\pi}$) can be equivalently represented by a vector in 2D Euclidean space $\tilde{\boldsymbol{P}} = [\tilde{P}_1, \tilde{P}_2]^\top \in \mathbb{R}^2$.

Among many possible choices for $\mathbb{S}^3 \to \mathbb{R}^2$, an isometric-log-ratio (ILR) transformation [16] $\tilde{\boldsymbol{P}} = \text{ilr}(\boldsymbol{P})$ defines

$$\tilde{P}_1 = \frac{1}{\sqrt{2}} \log \frac{P_2}{P_1}, \quad \tilde{P}_2 = \frac{1}{\sqrt{6}} \log \frac{P_3 P_3}{P_1 P_2}, \quad (2)$$

where the scalars $\frac{1}{\sqrt{2}}$ and $\frac{1}{\sqrt{6}}$ are due to the Aitchison orthonormal basis [15, Sec.10.2][16]. The same transformation $\text{ilr}(\cdot)$ can be applied to $\boldsymbol{\pi}$ and $\boldsymbol{w}$. Particularly, for $\tilde{\boldsymbol{w}}$, we have

$$\tilde{w}_1 = \frac{1}{\sqrt{2}} \log \frac{w_3 w_2}{w_1 w_3} = \frac{1}{\sqrt{2}} \Big(\mathsf{llr}_{\mathsf{spf}}^{\mathsf{tar.bf}}(\boldsymbol{x}) - \mathsf{llr}_{\mathsf{non.bf}}^{\mathsf{tar.bf}}(\boldsymbol{x}) \Big), \quad (3)$$

[2] It is unnecessary to separate H_{spf} into matched and non-matched cases since both should be rejected [8].

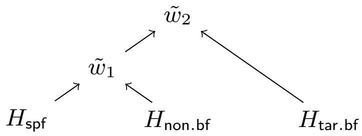


Figure 2: *Bifurcating tree for ternary hypothesis test based on compositional data analysis [15]*

$$\tilde{w}_2 = \frac{1}{\sqrt{6}} \log \frac{w_3 w_3}{w_1 w_2} = \frac{1}{\sqrt{6}} \Big(\mathsf{llr}_{\mathsf{non.bf}}^{\mathsf{tar.bf}}(\boldsymbol{x}) + \mathsf{llr}_{\mathsf{spf}}^{\mathsf{tar.bf}}(\boldsymbol{x}) \Big). \quad (4)$$

With the $\tilde{\boldsymbol{P}}, \tilde{\boldsymbol{\pi}}$, and $\tilde{\boldsymbol{w}}$, it is known that $\tilde{\boldsymbol{P}} = \tilde{\boldsymbol{\pi}} + \tilde{\boldsymbol{w}}$ [15, Sec.3.3.3]. This equation has a similar form of "posterior log-odds = prior log-odds + LLR" to Eq. (1). Similar to the sum of LLRs in Eq. (1), $\tilde{\boldsymbol{w}} = [\tilde{w}_1, \tilde{w}_2]^\top$ can be interpreted as weights of evidence, albeit in a hierarchical manner [15] as illustrated in Fig. 2. A larger $\tilde{w}_2$ favors $H_{\mathsf{tar.bf}}$ over the other two, and it is the $\tilde{w}_1$ that discriminates H_{spf} from $H_{\mathsf{non.bf}}$.

The above description indicates that $\tilde{w}_2$ is what an SASV system needs to produce. This observation paves the way for interpreting and improving the SASV score fusion methods.

2.3. Fusion method 1: summation of ASV and CM scores

Note that ASV is designed to discriminate $\mathsf{tar.bf}$ and $\mathsf{non.bf}$, while CM differentiates $\mathsf{tar.bf}$ from spf. **If** their scores approximate the LLRs, i.e., $s_{\text{asv}} \approx \mathsf{llr}_{\mathsf{non.bf}}^{\mathsf{tar.bf}}(\boldsymbol{x})$ and $s_{\text{cm}} \approx \mathsf{llr}_{\mathsf{spf}}^{\mathsf{tar.bf}}(\boldsymbol{x})$, [3] we can follow Eq. (4) and compute s_{sasv} as

$$\tilde{w}_2 \approx s_{\text{sasv}} = \frac{1}{\sqrt{6}} (s_{\text{asv}} + s_{\text{cm}}). \quad (5)$$

Hence, the summation-based score fusion is interpretable from the perspective of compositional data analysis. Note that the factor $\frac{1}{\sqrt{6}}$ does not affect the discrimination power of s_{sasv}.

The interpretation may be disappointing since Eq. (5) is the same as what we have done, but it suggests a practice that is either ignored or conducted arbitrarily: *the ASV and CM scores need to well approximate the LLRs before summation.* One way to do so is score calibration [10]. For example, let $\{f_{\text{asv}}, f_{\text{cm}}\}$ be calibration functions, both of which have a form of affine transformation $f(x) = ax + b$. We can learn parameters of $\{f_{\text{asv}}, f_{\text{cm}}\}$ on a development set through logistic regression [18], after which we compute $s_{\text{sasv}} = \frac{1}{\sqrt{6}} \big[f_{\text{asv}}(s_{\text{asv}}) + f_{\text{cm}}(s_{\text{cm}}) \big]$.

2.4. Fusion method 2: summation of LLRs

Rather than calibrating the ASV and CM scores to produce the LLRs (i.e., discriminative calibration [19]), we can treat the scores as features and carry out generative calibration [19]. Let us create a feature vector $\boldsymbol{s} = [s_{\text{asv}}, s_{\text{cm}}]^\top$ and introduce likelihood functions $p(\boldsymbol{s}; \boldsymbol{\theta}_{\mathsf{tar.bf}})$, $p(\boldsymbol{s}; \boldsymbol{\theta}_{\mathsf{spf}})$, and $p(\boldsymbol{s}; \boldsymbol{\theta}_{\mathsf{non.bf}})$ for the three classes. Assuming $\boldsymbol{s}$ encodes all the information of $\boldsymbol{x}$ and $\frac{p(\boldsymbol{x}|H_1)}{p(\boldsymbol{x}|H_2)} = \frac{p(\boldsymbol{s}|\boldsymbol{\theta}_1)}{p(\boldsymbol{s}|\boldsymbol{\theta}_2)}$, based on Eq. (4), we can compute

$$s_{\text{sasv}} = \frac{1}{\sqrt{6}} \Big(\mathsf{llr}_{\mathsf{non.bf}}^{\mathsf{tar.bf}}(\boldsymbol{s}) + \mathsf{llr}_{\mathsf{spf}}^{\mathsf{tar.bf}}(\boldsymbol{s}) \Big), \quad (6)$$

where $\mathsf{llr}_{\mathsf{non.bf}}^{\mathsf{tar.bf}}(\boldsymbol{s}) \triangleq \log \frac{p(\boldsymbol{s}; \boldsymbol{\theta}_{\mathsf{tar.bf}})}{p(\boldsymbol{s}; \boldsymbol{\theta}_{\mathsf{non.bf}})}$ and $\mathsf{llr}_{\mathsf{spf}}^{\mathsf{tar.bf}}(\boldsymbol{s}) \triangleq \log \frac{p(\boldsymbol{s}; \boldsymbol{\theta}_{\mathsf{tar.bf}})}{p(\boldsymbol{s}; \boldsymbol{\theta}_{\mathsf{spf}})}$. Each $p(\boldsymbol{s}; \boldsymbol{\theta})$ can be a Gaussian or another parametric distribution, and its parameter set $\boldsymbol{\theta}$ can be estimated by maximizing the likelihood on development data.

[3] For a speaker-independent CM, we can assume $p(\boldsymbol{x}|\mathsf{tar.bf}) \approx p(\boldsymbol{x}|\mathsf{non.bf}) \approx p(\boldsymbol{x}|\mathsf{bona})$ [17] and $\mathsf{llr}_{\mathsf{spf}}^{\mathsf{tar.bf}}(\boldsymbol{x}) \approx \mathsf{llr}_{\mathsf{spf}}^{\mathsf{bona}}(\boldsymbol{x})$.

During testing, we compute $p(\boldsymbol{s}; \boldsymbol{\theta}_{\text{tar.bf}})$, $p(\boldsymbol{s}; \boldsymbol{\theta}_{\text{spf}})$, and $p(\boldsymbol{s}; \boldsymbol{\theta}_{\text{non.bf}})$ of a test sample and plug them into Eq. (6). Since the selected parametric distributions may be different from the actual distribution of $\boldsymbol{s}$, the estimated LLRs may be inaccurate. Hence, we can optionally apply discriminative calibration on the LLRs before fusion.

2.5. Fusion method 3: nonlinear fusion of LLRs

The interpretation so far leads to the simple form of score fusion in Eqs. (5) and (6). Importantly, however, it also reveals a major limitation for making decisions. Recall from Sec. 2.1 that, for a single ASV system ($K = 1$), we accept tar if and only if $\text{llr}^{\text{tar}}_{\text{non}}(\boldsymbol{s}) + \log \frac{\pi_{\text{tar}}}{1-\pi_{\text{tar}}} > 0$. For SASV, a similar decision policy is to accept tar.bf if and only if $\tilde{P}_2 = \tilde{\pi}_2 + \tilde{w}_2 > 0$. Given $\tilde{P}_2$ defined in Eq. (2), this condition can be written as

$$P(H_{\text{tar.bf}}|\boldsymbol{x})^2 > P(H_{\text{non.bf}}|\boldsymbol{x})P(H_{\text{spf}}|\boldsymbol{x}), \quad (7)$$

or equivalently,

$$\text{llr}^{\text{tar.bf}}_{\text{non.bf}}(\boldsymbol{s}) + \text{llr}^{\text{tar.bf}}_{\text{spf}}(\boldsymbol{s}) > \log \frac{\pi_{\text{spf}}\pi_{\text{non.bf}}}{\pi^2_{\text{tar.bf}}}. \quad (8)$$

However, this decision policy cannot minimize the decision cost even if when the true LLRs and priors are given.

It can be shown that, for SASV with equal decision costs on false rejection of tar.bf and false acceptance of non.bf or spf, the optimal decision policy is to accept tar.bf when [4]

$$P(H_{\text{tar.bf}}|\boldsymbol{x}) > P(H_{\text{non.bf}}|\boldsymbol{x}) + P(H_{\text{spf}}|\boldsymbol{x}), \quad (9)$$

or with Bayes' formula and logarithm applied,

$$-\log[(1-\rho)e^{-\text{llr}^{\text{tar.bf}}_{\text{non.bf}}(\boldsymbol{s})} + \rho e^{-\text{llr}^{\text{tar.bf}}_{\text{spf}}(\boldsymbol{s})}] > -\log\beta, \quad (10)$$

where $\beta \triangleq \frac{\pi_{\text{tar.bf}}}{\pi_{\text{non.bf}}+\pi_{\text{spf}}}$ and $\rho \triangleq \frac{\pi_{\text{spf}}}{\pi_{\text{non.bf}}+\pi_{\text{spf}}}$. [5]

Inequality. (7) is a necessary but not sufficient condition to achieve Ineq. (9). For example, a sample with $[P(H_{\text{spf}}|\boldsymbol{x}), P(H_{\text{non.bf}}|\boldsymbol{x}), P(H_{\text{tar.bf}}|\boldsymbol{x})] = [0.05, 0.65, 0.3]$ will be rejected by Ineq. (9) but falsely accepted by (7). The overall decision cost cannot be lower than that based on the optimal decision policy in Ineq. (9) (or (10)), given that we know the true LLRs and priors.

To support the above argument, we generated ASV and CM scores using Gaussian distributions and a uniform prior, which enables us to compute the true LLRs and compared the linear and non-linear decision boundaries. As Fig. 3 indicates, the non-linear one based on Ineq. (10) leads to a lower decision cost, particularly, a lower false acceptance rate. Note that, Ineqs. (8) and (10) are agnostic to the data distributions, and we use Gaussians in simulation only for convenience.

On the basis of Ineq. (10), we define a new fusion method

$$s_{\text{sasv}} = -\log\left[(1-\rho)e^{-\text{llr}^{\text{tar.bf}}_{\text{non.bf}}(\boldsymbol{s})} + \rho e^{-\text{llr}^{\text{tar.bf}}_{\text{spf}}(\boldsymbol{s})}\right]. \quad (11)$$

The usage is similar to that in Sec. 2.4, but we plug the LLRs into Eq. (11). Unfortunately, the prior-dependent ρ cannot be decoupled from the LLRs, and it needs to be decided, for example, by grid search on a development set. The green solid line in Fig. 3 shows the decision boundary when ρ is computed from different priors. It performs reasonably well.

[4]It is different from the condition $P(H_{\text{tar.bf}}|\boldsymbol{x}) > P(H_{\text{non.bf}}|\boldsymbol{x})$ AND $P(H_{\text{tar.bf}}|\boldsymbol{x}) > P(H_{\text{spf}}|\boldsymbol{x})$ [15, sec.10.3], which maximizes the decision utility [20] with a scalar diagonal utility matrix.

[5]ρ is the same as the spoof prevalence prior in t-EER [8].

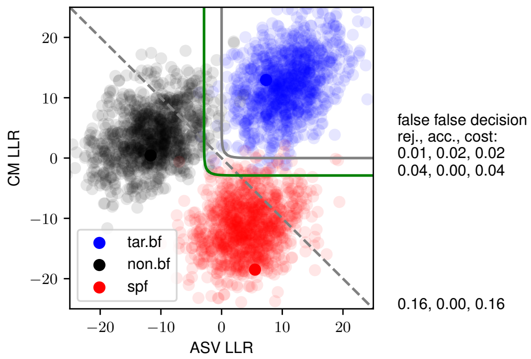


Figure 3: *Scatter plot of ASV* $\text{llr}^{\text{tar.bf}}_{\text{non.bf}}(\boldsymbol{s})$ *and CM* $\text{llr}^{\text{tar.bf}}_{\text{spf}}(\boldsymbol{s})$ *from simulated data. Dashed and solid lines are decision boundaries based on Ineqs. (8) and (10), respectively, given true flat prior. Green solid line is based on Ineq. (10) but mismatched priors* $\pi_{\text{spf}} = 0.05, \pi_{\text{non.bf}} = 0.05, \pi_{\text{tar.bf}} = 0.9$. *Note that LLRs rather than raw scores are plotted.*

As a special case, if each $p(\boldsymbol{s}; \boldsymbol{\theta})$ is implemented as a Gaussian distribution, Eq. (11) turns out to be equivalent to the so-called Gaussian back-end fusion [21, Eq.(1)]. The link between Gaussian back-end fusion and the optimal decision has not been shown before as far as the authors are aware. Detailed proof is available on the code site (see link on page 1).

3. Experiments

We verified the effectiveness of the three fusion methods on the SASV challenge dataset [3] and the official training, development, and evaluation splits.

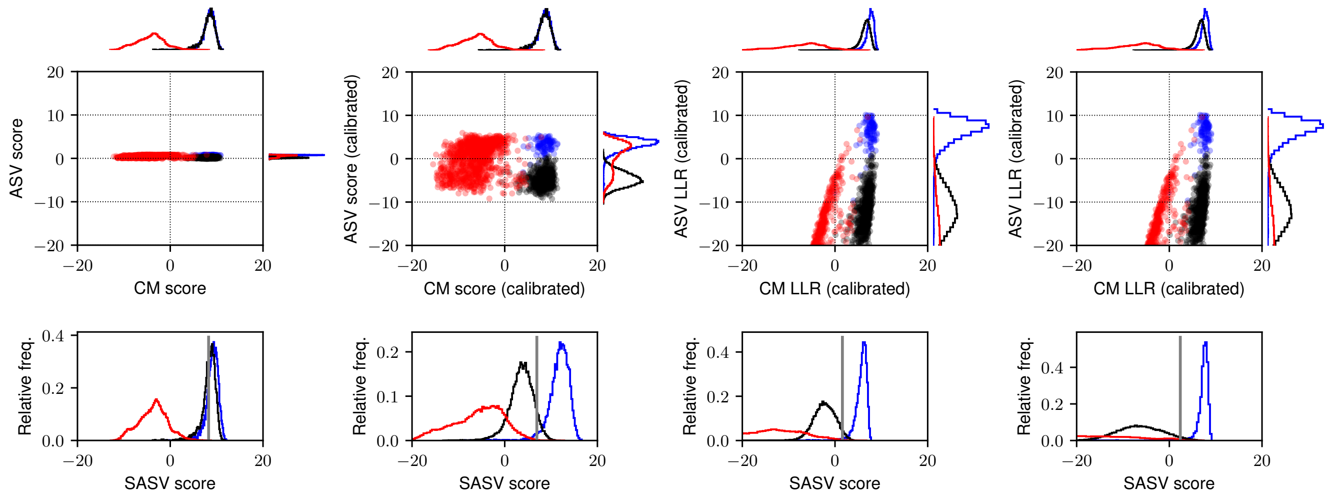
3.1. Model recipe and evaluation metrics

All eight experimental systems were based on the open-sourced SASV challenge baseline B1 (Fig. 1) and used the pre-trained ECAPA-TDNN-based ASV [22] and AASIST-based CM [23] embedding extractors, which were provided by the challenge organizers. The differences in the systems are

- `B1` is a replicate of SASV B1. The ASV branch requires no training, and the neural network in the CM branch is trained using binary cross-entropy given s_{cm} and the labels for CM. The SASV score is computed by $s_{\text{sasv}} = s_{\text{cm}} + s_{\text{asv}}$.
- `L2` is similar to `B1` but uses Eq. (6) for score fusion. Each $p(\boldsymbol{s}; \boldsymbol{\theta})$ is a Gaussian with a full covariance matrix.
- `L3` follows `L2` but uses Eq. (11) for score fusion.
- `B1c`, `L2c`, and `L3c` are variants of `B1`, `L2`, and `L3`, respectively, but calibrated scores or LLRs before fusion. The calibration function is $f(x) = ax + b$.
- `B1v2` (B2-v2 in [3]) and `Post` (PR-S-F in [5]) were included for reference. They interpret scores as posterior probabilities and fuse by $s_{\text{sasv}} = \sigma(s_{\text{cm}}) + s_{\text{asv}}$ and $s_{\text{sasv}} = \sigma(s_{\text{cm}}) \times \sigma(s_{\text{asv}})$, respectively [24]. The $\sigma(\cdot)$ is a sigmoid function.

The parameters $\boldsymbol{\theta}$, ρ, and $\{a, b\}$ for calibration were tuned on the development set.

All systems have about 82.5 k trainable parameters in the the neural network for CM, which has three hidden layers with 258, 128, and 64 neurons, respectively. All layers used a leaky ReLU activation with a negative slope value of 0.3. All systems were trained on a Tesla V100 card using the Adam optimizer with a learning rate of 5×10^{-5}, batch size of 24, and



(a) B1 (b) B1c (c) L2c (d) L3c

Figure 4: *Distributions of CM, ASV, and fused SASV scores. Bona fide data of target speakers (*`tar.bf`*), those of non-target speakers (*`non.bf`*), and spoofed data (*`spf`*) are in different colors. Each vertical line in bottom plane marks SASV-EER threshold.*

Table 1: *Results on SASV evaluation set. Each number is averaged over six training and evaluation rounds. Calibration (discriminative) is not on s_{sasv} but on scores or LLRs to be fused. Darker cell color indicates worse result in each row.*

ID	B1	B1c	L2	L2c	L3	L3c	B1v2	Post
Fusion	Eq. (5)		Eq. (6)		Eq. (11)		[3]	[5]
Calibration	×	✓	×	✓	×	✓	×	×
SASV-EER (%)	20.46	2.73	3.31	1.56	1.44	1.43	1.60	1.55
conf. ($\alpha = 5\%$)	±0.40	±0.27	±0.31	±0.23	±0.23	±0.23	±0.22	±0.24
Cllr	2.17	1.09	1.04	0.14	0.18	0.16	0.96	0.84
$\mathrm{Cllr}_{\mathrm{min}}$	0.52	0.11	0.13	0.07	0.06	0.07	0.08	0.07
$\mathrm{Cllr}_{\mathrm{calib}}$	1.64	0.98	0.91	0.07	0.11	0.10	0.88	0.78
t-EER (%)	2.10	2.10	1.68	1.68	1.68	1.68	2.19	2.21

weight decay with a coefficient of 1×10^{-3}. The maximum number of training epochs was 10, and the checkpoint with the best CM EER on the development set was used for evaluation. The implementation was done using SpeechBrain [25], and the training took less than one day per system.

The SASV-EER [3] was used to gauge the discrimination power of s_{sasv}. The log-likelihood-ratio cost Cllr [26] was also computed and decomposed into $\mathrm{Cllr}_{\mathrm{min}}$ and $\mathrm{Cllr}_{\mathrm{calib}}$ to measure the discrimination and calibration performance of s_{sasv}. The t-EER [8] was computed given $\{s_{\mathrm{asv}}, s_{\mathrm{cm}}\}$ or the LLRs. Each system was trained and evaluated for six rounds, where each round used a different seed to randomly initialize the neural network in the CM sub-system. The averaged results are listed in Table 1. The confidence interval of SASV-EER was using the Interspeech official toolkit with the default setting.

3.2. Results

Calibration before fusion was useful: As reported in the original paper [3], `B1` achieved an SASV-EER of around 20%. As Fig. 4(a) shows, due to the large difference in the dynamic ranges of ASV and CM scores, the latter dominated the fused SASV score. The values of Cllr and $\mathrm{Cllr}_{\mathrm{calib}}$ were higher than 1 bit, indicating that the SASV scores were not useful for making decisions. By calibrating s_{asv} and s_{cm} before fusion, `B1c` reduced the SASV EER to 2.73%, and Cllr-based metrics also improved. From Fig. 4(b), we can observe that the distribution of s_{asv} was 'stretched out', and the fused s_{sasv} better separated the three classes of data.

When fusing LLRs, the system pairs {`L2`, `L2c`} and {`L3`, `L3c`} showed that the gain of the calibration was diminishing. This is understandable because the dynamic range of ASV and CM scores are handled by the covariance matrix of $p(\boldsymbol{s}; \boldsymbol{\theta})$. However, calibrating before fusion was still helpful, especially in the case of linear fusion in Eq. (6). Compared with `L2`, `L2c`'s EER was reduced to 1.56%, and its Cllr was only 0.14 bits.

Non-linear fusion of LLRs was better: Rather than fusing the ASV and CM scores, `L2`, `L3`, and their variants fused the estimated LLRs. The nonlinear fusion of LLRs in `L3c` led to the best SASV-EER in our experiments. The difference between `L3c` and `L2c` was small, possibly because the test data set of the SASV challenge can be well separated using a linear decision boundary in the space of LLRs, as Fig. 4(c) and (d) show.

Compared with `B1v2` and `Post`, which treat the ASV and CM scores as posterior probabilities, `L2c`, `L3`, and `L3c` performed similarly or better in terms of SASV-EER, and their $\mathrm{Cllr}_{\mathrm{calib}}$s were much lower. The results suggest that the fusion methods based on compositional data analysis are also practically useful.

Finally, the t-EER measures the discrimination power of $\{s_{\mathrm{asv}}, s_{\mathrm{cm}}\}$ or LLRs and is agnostic to their calibration and fusion. Hence, the t-EERs of `B1` and `B1c` are the same. Similarly, the four systems fusing LLRs have the same t-EER. It was observed that using the LLRs led to a lower t-EER than raw or sigmoid-transformed s_{asv} and s_{cm}. This is evidence to encourage the sub-systems to produce LLRs.

4. Conclusions

This paper presented an alternative interpretation of SASV score fusion based on compositional data analysis. We suggested practices and improved fusion methods that may have been ignored or used ad hoc. First, the calibration of scores or LLRs before fusion was found to be helpful, especially when linearly fusing raw scores or LLRs. Second, rather than fusing raw scores, estimating and fusing LLRs turned out to be better in experiments. Third, the linear fusion of LLRs was found to be inferior to the non-linear one in both analysis and experiments. What is presented in this paper is by no means the only way to interpret SASV fusion, thus is expected to foster other studies.

5. Acknowledgements

This work was supported by JST, PRESTO Grant Number JPMJPR23P9, Japan. It was supported in part by the Academy of Finland under Grant 349605, project "SPEECHFAKES".

6. References

[1] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, Feb. 2015.

[2] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, "Biometrics systems under spoofing attack: An evaluation methodology and lessons learned," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 20–30, 2015.

[3] J.-w. Jung, H. Tak, H.-j. Shim, H.-S. Heo, B.-J. Lee, S.-W. Chung, H.-J. Yu, N. Evans, and T. Kinnunen, "SASV 2022: The first spoofing-aware speaker verification challenge," in *Proc. Interspeech*, 2022, pp. 2893–2897.

[4] I. Chingovska, A. Anjos, and S. Marcel, "Anti-spoofing in action: Joint operation with a verification system," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 98–104.

[5] Y. Zhang, G. Zhu, and Z. Duan, "A probabilistic fusion framework for spoofing aware speaker verification," in *Proc. Odyssey*, Jun. 2022, pp. 77–84.

[6] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, "Joint speaker verification and antispoofing in the i-vector space," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 821–832, 2015.

[7] A. Gomez-Alanis, J. A. Gonzalez-Lopez, S. P. Dubagunta, A. M. Peinado, and M. Magimai.-Doss, "On joint optimization of automatic speaker verification and anti-spoofing in the embedding space," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1579–1593, 2021.

[8] T. H. Kinnunen, K. A. Lee, H. Tak, N. Evans, and A. Nautsch, "T-eer: Parameter-free tandem evaluation of countermeasures and biometric comparators," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–16, 2023.

[9] K. Nandakumar, Y. Chen, S. C. Dass, and A. Jain, "Likelihood ratio-based biometric score fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 342–347, 2007.

[10] G. S. Morrison, "Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio," *Australian Journal of Forensic Sciences*, vol. 45, no. 2, pp. 173–197, 2013.

[11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[12] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge university press, 2003.

[13] I. Chingovska, A. R. Dos Anjos, and S. Marcel, "Biometrics evaluation under spoofing attacks," *IEEE transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2264–2276, 2014.

[14] J. Aitchison, "The statistical analysis of compositional data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 44, no. 2, pp. 139–160, 1982.

[15] P.-G. Noé, "Representing evidence for attribute privacy : Bayesian updating, compositional evidence and calibration," Theses, Université d'Avignon, Apr. 2023.

[16] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelo-Vidal, "Isometric logratio transformations for compositional data analysis," *Mathematical geology*, vol. 35, no. 3, pp. 279–300, 2003.

[17] T. Kinnunen, H. Delgado, N. Evans, K. A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2195–2210, 2020.

[18] N. Brümmer and G. R. Doddington, "Likelihood-ratio calibration using prior-weighted proper scoring rules," in *Proc. Interspeech*, 2013, pp. 1976–1980.

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

[20] X. He, C. E. Metz, B. M. Tsui, J. M. Links, and E. C. Frey, "Three-class ROC analysis-A decision theoretic approach under the ideal observer framework," *IEEE Transactions on Medical Imaging*, vol. 25, no. 5, pp. 571–581, 2006.

[21] M. Todisco, H. Delgado, K. A. Lee, M. Sahidullah, N. Evans, T. Kinnunen, and J. Yamagishi, "Integrated presentation attack detection and automatic speaker verification: Common features and Gaussian back-end fusion," in *Proc. Interspeech*, 2018, pp. 77–81.

[22] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[23] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. ICASSP*, 2022, pp. 6367–6371.

[24] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[25] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021.

[26] N. Brummer, "Measuring, refining and calibrating speaker and language information extracted from speech," Ph.D. dissertation, Stellenbosch: University of Stellenbosch, 2010.

[27] A. Nautsch, "Speaker recognition in unconstrained environments." Ph.D. dissertation, Darmstadt University of Technology, Germany, 2019.

[28] D. R. Castro, "Forensic evaluation of the evidence using automatic speaker recognition systems," Ph.D. dissertation, Universidad autónoma de Madrid, 2007.

[29] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Introducing the VoicePrivacy Initiative," in *Proc. Interspeech*, ISCA, Oct. 2020, pp. 1693–1697.