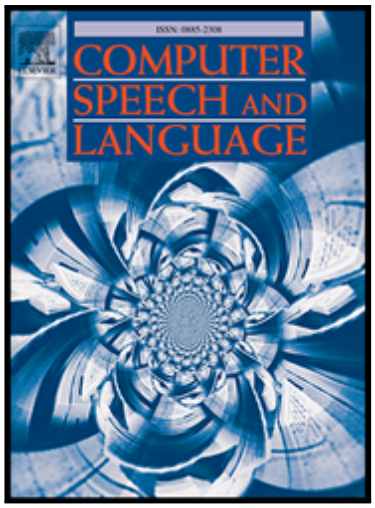
Journal Pre-proof



Voice Biometrics Security: Extrapolating False Alarm Rate via Hierarchical Bayesian Modeling of Speaker Verification Scores

Alexey Sholokhov, Tomi Kinnunen, Ville Vestman, Kong Aik Lee

PII: S0885-2308(19)30268-2
DOI: https://doi.org/10.1016/j.csl.2019.101024
Reference: YCSLA 101024

To appear in: *Computer Speech & Language*

Received date: 6 May 2019
Revised date: 9 July 2019
Accepted date: 25 September 2019

Please cite this article as: Alexey Sholokhov, Tomi Kinnunen, Ville Vestman, Kong Aik Lee, Voice Biometrics Security: Extrapolating False Alarm Rate via Hierarchical Bayesian Modeling of Speaker Verification Scores, *Computer Speech & Language* (2019), doi: https://doi.org/10.1016/j.csl.2019.101024

Highlights

- Security evaluation of automatic speaker verification for arbitrarily large population
- Bayesian generative model of non-target scores
- From random-impostor evaluation to worst-case impostor false alarm rate
- Analysis of i-vector and x-vector systems up to 100000 simulated impostors

Voice Biometrics Security: Extrapolating False Alarm Rate via Hierarchical Bayesian Modeling of Speaker Verification Scores

Alexey Sholokhov[a,b,c], Tomi Kinnunen[c,*], Ville Vestman[c,**], Kong Aik Lee[d]

[a] *St. Petersburg Electrotechnical University, ul. P. Popova 5, St. Petersburg, 197376, Russia*
[b] *ITMO University, Kronverkskiy pr. 49, St. Petersburg, 197101, Russia*
[c] *School of Computing, University of Eastern Finland, FI-80101, Joensuu, Finland*
[d] *Biometrics Research Laboratories, NEC Corp., Tokyo, Japan*

Abstract

How secure automatic speaker verification (ASV) technology is? More concretely, given a specific target speaker, how likely is it to find another person who gets falsely accepted as that target? This question may be addressed empirically by studying naturally confusable pairs of speakers within a large enough corpus. To this end, one might expect to find at least some speaker pairs that are indistinguishable from each other in terms of ASV. To a certain extent, such aim is mirrored in the standardized ASV evaluation benchmarks, for instance, the series of speaker recognition evaluation (SRE) organized by the National Institute of Standards and Technology (NIST). Nonetheless, arguably the number of speakers in such evaluation benchmarks represents only a small fraction of all possible human voices, making it challenging to extrapolate performance beyond a given corpus. Furthermore, the impostors used in performance evaluation are usually selected *randomly*. A potentially more meaningful definition of an impostor — at least in the context of security-driven ASV applications — would be *closest* (most confusable) other speaker to a given target.

We put forward a novel performance assessment framework to address both the inadequacy of the random-impostor evaluation model and the size limitation of evaluation corpora by addressing ASV security against closest impostors on arbitrarily large datasets. The framework allows one to make a prediction of the safety of given ASV technology, in its current state, for arbitrarily large speaker database size consisting of virtual (sampled) speakers. As a proof-of-concept, we analyze the performance of two state-of-the-art ASV systems, based on i-vector and x-vector speaker embeddings (as implemented in the

popular Kaldi toolkit), on the recent VoxCeleb 1 & 2 corpora, containing a total of 7,365 speakers. We fix the number of target speakers to 1000, and generate up to $N = 100,000$ virtual impostors sampled from the generative model. The model-based false alarm rates are in a reasonable agreement with empirical false alarm rates and, as predicted, increase substantially (values up to 98%) with $N = 100,000$ impostors. Neither the i-vector or x-vector system is immune to increased false alarm rate at increased impostor database size, as predicted by the model.

Keywords: Speaker verification, population size, security, false alarm rate, random impostor, closest impostor, Bayesian score modeling, VoxCeleb

1. Introduction

Some have predicted that voice-operated user interfaces will be the next paradigm of human-machine interaction. Given that the consumer market already provides various *virtual assistants* — Google Home, Apple Siri, and Amazon Alexa to name a few — it might be a reasonable prediction. Such services are intended to provide human-to-human like user experience leveraging from speech and speaker recognition technology, dialogue modeling and speech synthesis. An increasing number of smart services also enable users to log-in or authenticate payments using voice (or other biometric traits), for both increased security and user convenience — there is no need to consult, *e.g.*, printed key-lists (or other stealable or copiable accessories). The co-evolution of smart device technology and machine learning [1, 2] has substantially broadened the landscape of *automatic speaker verification* (ASV) [3] use cases from its traditional, highly specialized applications — forensics and survaillance — to our living rooms and everyday mobile environments. For instance, nowadays, smart phones, virtual assistants and other devices with powerful processors and wireless connectivity enable efficient on-device or cloud-based voice data processing, including ASV-based user authentication with algorithms that would have been difficult to execute on portable devices of the past decades. Early

*Corresponding Author
**A part of the work of the third author was carried out during an intership at NEC.
Email addresses: sholokhovalexey@gmail.com (Alexey Sholokhov), tkinnu@cs.joensuu.fi (Tomi Kinnunen), vvestman@cs.uef.fi (Ville Vestman), k-lee@ax.jp.nec.com (Kong Aik Lee)

ASV technology, such as [3], was developed with the aid of far less powerful computers and smaller datasets. The increase in dataset sizes and computing power has not only enabled the research community to address increasingly more challenging ASV tasks, but enabled running more powerful models in portable devices. Much of the progress in the underlying core ASV technology has been facilitated by coordinated technology benchmarks, pioneered by *National Institute of Standards and Technology* (NIST) in their evaluation campaigns [4, 5, 6].

Increased awareness of the possibilities of voice-based interaction also raises concern about the security of the technology. The possibility to invoke malicious voice commands from a distance in another user's phone [7] (potentially even using *inaudible* sounds [8]), and the potential to masquerade oneself as another targeted speaker through various *spoofing attacks* [9] is widely acknowledged. The latter includes *replay*, *text-to-speech*, and *voice conversion* attacks. Many of these *technology*-aided attacks can be combated through various *countermeasures* ranging from knowledge-based approaches to classification approaches, known within biometric technology standardization bodies as *presentation attack detection* (PAD) [10] methods. For instance, specialized binary detector could be used to verify liveness of a voice sample before being passed to a speaker verification system. Detection of attacks is possible since replayed speech, introduced through loudspeakers, has different frequency characteristics than live human; and since synthetic and converted voices contain processing artifacts due to training data limitations and modeling imperfections. More details of different attacks, their effectiveness, detection, and evaluation metrics are discussed elsewhere [11] in more detail. In this study, we focus on core ASV technology.

While recent efforts have capitalized the importance of preparing ASV systems against spoofing attacks, another, more fundamental question remains: how *unique* the human voice is? Note that even the performance of an ASV system equipped with *perfect* PAD subsystem will be upper bounded by the performance of the underlying core technology [12]. This raises fundamental, yet thus far conclusively unanswered questions such as,

- Given a large-enough population of speakers (such as 7.6 billion), how likely is it to find two speakers that are confusable with each other? In other words, *how many unique voices there are?*

- Conversely, assuming that we wish to maintain a certain minimum level of non-confusability between speakers, is there some maximum population (speaker database) size for which it can be guaranteed?

Answers would enable both technology vendors and users of ASV technology to have increased confidence to the expected reliability of such systems. By drawing analogy from the security of passwords, some studies [13] based on biometric information measures [14] have assessed the strength of speech representations in terms of their speaker information, though the viewpoint is rarely neither on the population size nor attacks.

Before proceeding further, it is necessary to constrain the scope. First, the question of voice uniqueness is, clearly, ill-posed. *In theory*, the number of different human voices is, if not infinite, some very large number: both the organic (physiological) and learnt traits vary greatly across individuals thanks to differences in the anatomy and kinematics of our articulatory systems — it would be extremely unlikely to find another *voice clone* with perfectly-matched voice production systems and learned traits. *In practice*, when working with real-world acoustic speech waveforms, we are bounded both by *extrinsic* and *intrinsic* signal variations. Extrinsic variation refers to the inability to accurately measure 'pure' speaker characteristics from imperfect acoustic observations (for instance, due to imperfect transducer, lossy communication channel, background noise, or reverberant environment). Intrinsic variation, in turn, refers to linguistic and non-linguistic variation induced by the speaker him/herself, some of which can be substantial [15, 16]. The main focus of the ASV research community for the past several decades [3, 17] has been on improving ASV technology to handle extrinsic variations of increased complexity, though specific intrinsic factors, such as *vocal effort*, have also been addressed in the context of NIST SREs [18].

Neither the extrinsic nor intrinsic variations are deterministic, fixed operations. Therefore, there are practical limits as to how accurately one can discriminate two voices from each other. As these limits are clearly a function of the specific types of variations and distortions (as the ASV community is well aware of), it would be meaningless to attempt to answer the unconditional question of voice uniquess. The answer depends on both data conditions and the employed hypothesis tester (*e.g.*, a specific human listener or a specific ASV system). We might even say that uniqueness of voices is a *subjective* matter;

a pair of speakers that is confusable for one hypothesis tester A (for instance, a human) may not be so for another hypothesis tester B (for instance, a machine).

We, therefore, constrain the focus on *statistical methods* to address questions such as the above *empirically* for given data. In particular, we are interested in the relation of corpus size (number of speakers) and the probability of a false alarm (P_{FA}) for a given ASV system, under a specific model detailed in Section 2. The input data to our proposed model consists of detection scores (log-likelihood ratios or uncalibrated raw scores) of *any* ASV system on a specific corpus. This makes the method widely applicable for the analysis of any ASV system, treated as a black-box.

The reader familiar with performance assessment of ASV systems may wonder if there is anything new to say about detection scores of a given system on a given corpus. Indeed, measuring detection errors (including P_{FA}) and calibrating speaker recognition systems is a fairly standardized activity [4, 19]. So, what is new here? The answer, in brief, is that in the NIST-style ASV evaluations, the *non-target* speaker trials (pairwise comparisons of test utterances against a hypothesized speaker model with disjoint speaker identities) are, essentially, random pairs of speakers. We use more effort to model situation of more confusable (closest) pairs of speakers; one could argue a recognizer that handles the 'worst cases' (closest competing) speakers well may exhibit improved generalization.

In our model, 'closest' speakers are in fact *none* of the non-target speakers in the training set, but *virtual speakers* sampled from the distribution that models random sampling of speakers. Specifically, speakers are represented implicitly by distributions of scores corresponding to pairs of speakers. This allows us to *extrapolate* beyond the given evaluation corpus to arbitrarily large virtual speaker populations. Assuming that the observed speaker pairs are sampled from a same underlying generative process, we can get an idea of how the ASV system scales up with corpus size, *without collecting new speech data*.

While the technical voice conversion spoofing attacks have received a lot of attention in the recent years, it might be appropriate time to re-address worst-case impostors in the context of regular ASV as well. The initial spark for this work stems from our recent work [20] (inspired by [21]) where we addressed a specific research hypothesis relating to potentially emerging, yet cursorily addressed vulnerability of ASV technology *against*

itself. The idea was that an attacker could use (public-domain) ASV system as a voice search engine to identify suitable target speaker (specifically, the closest one), such as a celebrity or any person who uploads a lot of his/her voice or video samples to the Internet. After identifying a suitable target, the attacker would attempt to attack another ASV system (*e.g.*, at bank) using natural (possibly mimicked[1]) voice. Despite the relatively large VoxCeleb corpus with more than 7000 target speakers, none of our attackers were successful in getting falsely accepted[2]. While good news concerning security of ASV, the finding was on specific ASV systems, attackers and target corpus. One reason why the finding in [20] might have been negative is that the attacker's ASV (designed to be purposefully different from the attacked one) was not powerful enough. Nonetheless, we saw transferability across our two ASV systems in terms of relative target speaker rankings, suggesting that the attacks might be successful with a scaled-up database. We argue that there *must be* a speaker database size (possibly very large) where one is likely to locate closely-matched non-target voices — effect which we were unable to *observe* under the specific experimental conditions. For these reasons, we wanted to re-address the problem by using a more principled and re-usable setup that requires neither two ASV systems (attacker's ASV and targeted ASV) nor fresh recordings. To be precise, the framework proposed in this study addresses a *worst-case* attack scenario with the following two assumptions:

1. **Assumption 1: known ASV system.** The adversary's ASV system (used for identifying closest targets to attack) is the same as the attacked ASV system.
2. **Assumption 2: access to target's enrollment data.** The adversary has access to the target speaker's enrollment data (alternatively, no domain mismatch exist between target's public-domain and enrollment recordings).

[1]Mimicry is a special skill, based on the idea of a listener trying to match his or her acoustic profile with that of another person. As the acoustic correlates of speaker identity, as learned by current ASV systems, remain largely unknown, human mimicry is generally an inconsistent strategy to spoof ASV systems. This is why [20] included ASV system to first identify targets that are similar to attacker's voice.

[2]To be more precise, in [20], we did not consider hard binary decisions but analyzed changes in the log-likelihood ratio (LLR) scores of the ASV systems. The nontarget LLRs, whether or not originating from zero-effort or mimicry trials were far below the range of target LLRs.

The generative model presented in this work enables us to increase the corpus size indefinitely to establish empirical performance bounds on the false alarm rate, under these two assumptions. As search queries to the attacked system can be limited and the enrollment utterances can be protected by template protection techniques, neither assumption is necessarily realistic from the perspective of the adversary. An evaluation corpus designer, technology vendor, or a bank, however, may still want to assess worst-case performance. Importantly, the above assumptions greatly simplify the set-up over the scenarios addressed in [20]. The methods developed in this study can be seen as an extension of the arsenal of statistical performance evaluation tools. We address each of the two assumptions in the empirical part.

We summarize our two main contributions as follows. First, we propose a general-purpose performance metric, *worst-case false alarm rate with N impostors* (P_{FA}^N). It is the probability of accepting the closest impostor among N available candidate impostors selected randomly for each enrolled speaker. As will be discussed below, the proposed metric reduces to the 'conventional' probability of a false alarm (P_{FA}) if $N = 1$. Second, we devise a hierarchical Bayesian generative model of non-target score distribution to enable prediction of P_{FA}^N for arbitrarily large values of N that can exceed the number of non-target speakers in a given corpus. The proposed model allows one to make a prediction of the safety of given ASV technology, in its current state, for arbitrarily large speaker database size consisting of virtual (sampled) speakers. Importantly, as the training data consists of detection scores only, the framework is widely applicable for the analysis of arbitrary ASV system (or even other biometric systems). Further, all the model parameters are automatically inferred from data, leaving no manually-tunable control parameters to be set. As a representative snapshot of the current ASV technology and evaluation databases, our proof-of-concept experiments include two widely-used ASV methods based on *i-vector* [22] and *x-vector* [23] embeddings, evaluated on the combined VoxCeleb1 [24] and VoxCeleb2 [25] corpora.

2. Measuring and Extrapolating False Alarm Rates

An automatic speaker verification (ASV) system is a hypothesis testing machine that takes a pair of speech utterances $\mathcal{X} = (\mathcal{X}_{\text{e}}, \mathcal{X}_{\text{t}})$ — one for enrollment, one for test — and

produces a numerical detection score $s \in \mathbb{R}$, with the convention that higher values (in relative terms) indicate stronger support for the *same speaker* (null) hypothesis and low scores for the *different speaker* (alternative) hypothesis. Speech utterances are typically represented as fixed-sized *speaker embeddings* such as *i-vectors* [22] or *x-vectors* [23] and the detection score is a *logarithmic likelihood ratio* (LLR) produced by a statistical back-end model, such as the *probabilistic linear discriminant analysis* (PLDA) [26, 17].

2.1. False alarm rate

The detection score s can be interpreted as a realization of a continuous random variable that admits an underlying probability density $p(s)$, with $p(s) \geq 0$ and $\int_{s=-\infty}^{\infty} p(s)\,\mathrm{d}s = 1$. In the conventional ASV set-up (as in NIST SREs [4, 5]), the performance of an ASV system is assessed using two types of users, *targets* and *nontargets*. The former means that speaker identities of $\mathcal{X}_\mathrm{e}$ and $\mathcal{X}_\mathrm{t}$ match, while the latter means that they differ. We denote the class-conditional score densities of targets and nontargets by $p(s|\mathrm{tar})$ and $p(s|\mathrm{non})$, respectively.

Our focus is on ASV security against impostors, characterized by the nontarget score distribution. In specific, an ASV system is characterized by the probability of accepting a random impostor (sometimes known as *zero-effort* impostor), known as *false alarm rate* (or *false acceptance rate*). It is defined as the following non-increasing function of detection threshold $\tau \in \mathbb{R}$,

$$\mathrm{P}_\mathrm{FA}(\tau) = \int_{\tau}^{\infty} p(s|\mathrm{non})\,\mathrm{d}s, \tag{1}$$

where τ is fixed in advance to set $P_\mathrm{FA}(\tau)$ to a desirable level (increasing τ reduces false alarm rate but increases target rejection rate, also known as *miss rate*).

As we do not have access to $p(s|\mathrm{non})$, in practice $\mathrm{P}_\mathrm{FA}(\tau)$ is usually approximated using *Monte-Carlo* (MC) methods [27]. Monte-Carlo integration is a class of numerical methods that can be used to evaluate expected values of complicated functions. It replaces integrals in expectations by finite sums with the help of independent samples drawn from the underlying probability distribution. By using $\mathbb{I}\{\cdot\}$ to denote an indicator function that equals 1 for a true proposition and 0 otherwise, we write the MC-approximated false

alarm rate as,

$$\begin{aligned} \mathrm{P}_{\mathrm{FA}}(\tau) &= \int_{-\infty}^{\infty} p(s|\mathrm{non})\mathbb{I}\{s > \tau\}\,\mathrm{d}s \\ &= \mathbb{E}_{s\sim p(s|\mathrm{non})}[\mathbb{I}\{s > \tau\}] \approx \frac{1}{R}\sum_{r=1}^{R}\mathbb{I}\{s_r > \tau\}, \quad s_r \sim p(s|\mathrm{non}), \end{aligned} \tag{2}$$

by assuming one is able to obtain R independent samples s_r from the non-target score distribution. Here, $\mathbb{E}_{s\sim p(s)}[g(s)]$ denotes expected (average) value of function $g(s)$ w.r.t. the distribution $p(s)$. Usually we have just a finite collection of detection scores $\{s_r\}_{r=1}^{R}$ with no further knowledge of $p(s|\mathrm{non})$.

2.2. Reinterpreting False Alarm Rate as Averaged Speaker-Pair Conditioned False Alarm Rate

In the following, we provide an alternative view of the false alarm rate as an average of speaker-pair specific false alarm rates, useful in paving way towards a new performance metric and a generative model designed to extrapolate its values beyond available datasets. To that end, note first that the detection scores $\{s_r\}$ are obtained through an ASV system that processes some pre-defined *trial list* formed from a finite set of pairwise speaker comparisons. Thus, the terms in (2) can be divided into groups corresponding to unique pairs of speakers. In the special case when these groups are of equal size, we can rewrite the sum in (2) as

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{I}\{s_r > \tau\} = \frac{1}{T}\sum_{i=1}^{T}\underbrace{\frac{1}{L_i}\sum_{l=1}^{L_i}\mathbb{I}\{s_{i,l} > \tau\}}_{\substack{\text{speaker-pair specific} \\ \text{probability of a false alarm}}}, \tag{3}$$

where $s_{i,l}$ denotes the l^{th} trial score from speaker pair i, L_i is the total number of scores for the i^{th} speaker pair, and T is the total number of speaker pairs, such that $L_1 = L_2 = ... = L_T = L$ and $R = T \cdot L$. Here, the inner sum can be interpreted as the probability of a trial from a given pair of speakers being incorrectly accepted, with the outer sum forming average of the speaker-pair specific false alarm probabilities.

The above simple reformulation provides a bridge towards our proposed framework detailed below. As our approach enables extrapolation of $P_{\mathrm{FA}}(\tau)$ estimates beyond a given speech corpus, it is necessary to proceed from the empirical averaged false alarm

rate (3) towards a continuous-space formulation. In specific, as illustrated in Fig. 1, we require a model that enables sampling both speakers and speaker-pair specific scores from continuous distributions. Note, first, that the distribution of non-target scores $p(s|\text{non})$ can be seen as a continuous mixture of score distributions between all possible pairs of speakers,

$$p(s|\text{non}) = \iint p(s|\boldsymbol{y}_\text{e}, \boldsymbol{y}_\text{t}) p(\boldsymbol{y}_\text{e}) p(\boldsymbol{y}_\text{t}) \, \mathrm{d}\boldsymbol{y}_\text{e} \, \mathrm{d}\boldsymbol{y}_\text{t}, \tag{4}$$

where we have introduced two new vector-valued variables $\boldsymbol{y}_\text{e}$ and $\boldsymbol{y}_\text{t}$, viewed as so-called *latent identity variables* [26, 17]. Let $\boldsymbol{y} \in \mathcal{Y}$ be an element of some space $\mathcal{Y}$. The latent identity variable framework [26] assumes that $\boldsymbol{y}$ is a pure representation of a person's identity and that there is a distribution on $\mathcal{Y}$ with known probability density function $p(\boldsymbol{y})$. Given a likelihood function for the latent identity variable (*e.g.*, *meta-embedding* [28]), one can make inferences about speaker identities within a set of speech utterances. Examples of such tasks include speaker verification, identification and clustering [29]. For instance, speaker verification involves testing whether two sets of utterances belong to the same or to different speakers. In this setup the unit of observations, a speech utterance, corresponds to a single speaker identity.

The same framework can also be used in the score domain where observations correspond to pairs of identities. Given a pair of (unknown) identity variables $\boldsymbol{y}_\text{e}, \boldsymbol{y}_\text{t} \in \mathcal{Y}$, one can describe the distribution of similarity scores between the corresponding speakers by the density function $p(s|\boldsymbol{y}_\text{e}, \boldsymbol{y}_\text{t})$. This allows to conduct a test of alternative hypotheses such as: (i) two sets of scores belong to different pairs of speakers, (ii) two sets of scores share one common speaker, (iii) two sets of scores belong to the same pair of speakers.

The representation (4) allows us to rewrite the false alarm probability $\mathrm{P}_\text{FA}(\tau)$ akin to (3), namely,

$$\begin{aligned}
\mathrm{P}_\text{FA}(\tau) &= \int_\tau^\infty p(s|\text{non}) \, \mathrm{d}s = \int_\tau^\infty \left(\iint p(s|\boldsymbol{y}_\text{e}, \boldsymbol{y}_\text{t}) p(\boldsymbol{y}_\text{e}) p(\boldsymbol{y}_\text{t}) \, \mathrm{d}\boldsymbol{y}_\text{e} \, \mathrm{d}\boldsymbol{y}_\text{t} \right) \mathrm{d}s \\
&= \iint \underbrace{\left(\int_\tau^\infty p(s|\boldsymbol{y}_\text{e}, \boldsymbol{y}_\text{t}) \, \mathrm{d}s \right)}_{\substack{\text{speaker-pair specific} \\ \text{probability of a false alarm}}} p(\boldsymbol{y}_\text{e}) p(\boldsymbol{y}_\text{t}) \, \mathrm{d}\boldsymbol{y}_\text{e} \, \mathrm{d}\boldsymbol{y}_\text{t},
\end{aligned} \tag{5}$$

where the inner integral is the speaker-pair specific probability of a false alarm and the outer two integrals correspond to summing over all possible speaker pairs.

Given a trial list with speaker IDs, one can obtain the estimate of $\mathrm{P}_{\mathrm{FA}}(\tau)$ using so-called *nested* Monte-Carlo [30]. It uses MC estimate of the inner integral in (5) to compute MC estimate of the outer integral. The corresponding nested sampling scheme consists of sampling a pair of speakers, followed by sampling a set of scores from the speaker-pair specific score distribution. In practice, any trial list consisting of T unique speaker pairs and the corresponding scores can be thought as being generated according to this scheme. For instance (see Figure 1), the following generative process produces the scores suitable for computing the nested MC estimate of $\mathrm{P}_{\mathrm{FA}}(\tau)$:

1. sample an enrolled speaker $\boldsymbol{y}_{\mathrm{e}}^{(i)} \sim p(\boldsymbol{y})$
2. sample a test speaker $\boldsymbol{y}_{\mathrm{t}}^{(i)} \sim p(\boldsymbol{y})$
3. sample n_{e} utterances of the enrolled speaker $\boldsymbol{x}_{e,j} \sim p(\boldsymbol{x}|\boldsymbol{y}_{\mathrm{e}}^{(i)}), j = 1, 2, \ldots, n_{\mathrm{e}}$
4. sample n_{t} utterances of the test speaker $\boldsymbol{x}_{\mathrm{t},k} \sim p(\boldsymbol{x}|\boldsymbol{y}_{\mathrm{t}}^{(i)}), k = 1, 2, \ldots, n_{\mathrm{t}}$
5. compute $L_i = n_{\mathrm{e}} \cdot n_{\mathrm{t}}$ pairwise scores $s_{j,k} = \mathrm{score}(\boldsymbol{x}_{e,j}, \boldsymbol{x}_{\mathrm{t},k})$ using an ASV system.

Here, the index i runs over all speaker pairs and $p(\boldsymbol{x}|\boldsymbol{y})$ denotes the conditional distribution of speech utterances $\boldsymbol{x}$ belonging to speaker $\boldsymbol{y}$. Here, the last step can be equivalently re-formulated as sampling from the distribution of scores conditioned on a pair of speakers:

1. sample an enrolled speaker $\boldsymbol{y}_{\mathrm{e}}^{(i)} \sim p(\boldsymbol{y})$
2. sample a test speaker $\boldsymbol{y}_{\mathrm{t}}^{(i)} \sim p(\boldsymbol{y})$
3. sample L_i scores $s_l \sim p(s|\boldsymbol{y}_{\mathrm{e}}^{(i)}, \boldsymbol{y}_{\mathrm{t}}^{(i)}), l = 1, 2, \ldots, L_i$.

Now, the nested MC estimate of (5) can be found as

$$\mathrm{P}_{\mathrm{FA}}(\tau) \approx \frac{1}{T} \sum_{i=1}^{T} \mathrm{P}_{\mathrm{FA}}^{(i)}(\tau), \tag{6}$$

where

$$\begin{aligned} \mathrm{P}_{\mathrm{FA}}^{(i)}(\tau) &= \int_{\tau}^{\infty} p(s|\boldsymbol{y}_{\mathrm{e}}^{(i)}, \boldsymbol{y}_{\mathrm{t}}^{(i)})\, \mathrm{d}s \\ &\approx \frac{1}{L_i} \sum_{l=1}^{L_i} \mathbb{I}\{s_l > \tau\}, \quad s_l \sim p(s|\boldsymbol{y}_{\mathrm{e}}^{(i)}, \boldsymbol{y}_{\mathrm{t}}^{(i)}), \quad \boldsymbol{y}_{\mathrm{e}}^{(i)}, \boldsymbol{y}_{\mathrm{t}}^{(i)} \sim p(\boldsymbol{y}). \end{aligned} \tag{7}$$

We refer to $\mathrm{P}_{\mathrm{FA}}^{(i)}(\tau)$ as *speaker-pair conditioned* false alarm rate. It is the fraction of similarity scores between these speakers being above the decision threshold τ.

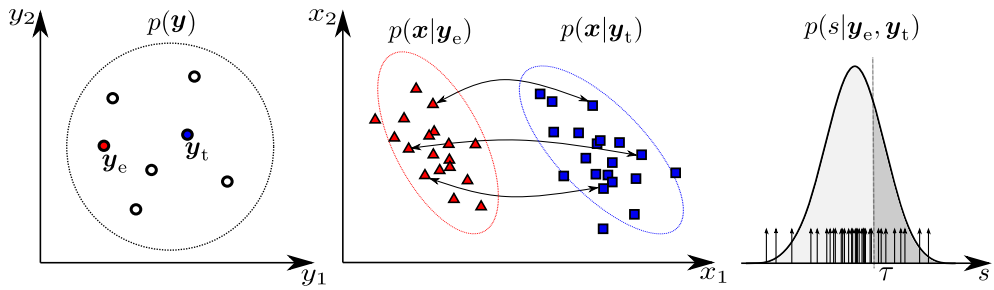


Figure 1: Illustration of the steps to obtain the speaker-pair conditioned score distribution. (Left) Latent speaker identity space. Each element of this space corresponds to unique identity. Small circles represent a dataset consisting of 7 speakers. (Middle) Observation space. Here, $p(\boldsymbol{x}|\boldsymbol{y})$ is the distribution of utterances $\boldsymbol{x}$ of the speaker $\boldsymbol{y}$. (Right) Score space. Here, $p(s|\boldsymbol{y}_\mathrm{e},\boldsymbol{y}_\mathrm{t})$ is the distribution of similarity scores between utterances of the pair of speakers. Samples from this distribution are shown as vertical arrows. Shaded area corresponds to the speaker-pair conditioned false alarm probability which depends on the decision threshold τ.

The $\mathrm{P}_\mathrm{FA}(\tau)$ can be estimated based on either a model or available empirical data. In the former case one needs a probabilistic model of between-speaker similarity scores and an algorithm to generate samples from this model. In specific, one must be able to obtain samples from the distribution of speaker identities $p(\boldsymbol{y})$ and from the distribution of similarity scores $p(s|\boldsymbol{y}_\mathrm{e},\boldsymbol{y}_\mathrm{t})$ given an arbitrary speaker pair $(\boldsymbol{y}_\mathrm{e},\boldsymbol{y}_\mathrm{t})$. An example of such a model will be described in Section 2.4. In the latter case, the distribution $p(\boldsymbol{y})$ is a uniform distribution over speakers' IDs and the observed between-speaker scores can be viewed as being samples drawn from an unknown distribution $p(s|\boldsymbol{y}_\mathrm{e},\boldsymbol{y}_\mathrm{t})$. That is, the $\mathrm{P}_\mathrm{FA}(\tau)$ can be estimated by repeated selection of random pairs of speakers from a dataset and computing similarity scores between random subsets of their sessions. Algorithm 1 summarizes a procedure to estimate the probability of accepting a zero-effort impostor, $\mathrm{P}_\mathrm{FA}(\tau)$, given a set of utterances with speaker labels.

One should note that in general case, *i.e.*, when speaker-pair specific subsets have different number of scores, L_i, the estimators defined by (2) and (6) produce different results. The former estimator relies on the unrealistic i.i.d. assumption and does not take into account data dependencies resulting from multiple appearances of the same speaker in a given trial list. In practice, however, limited resources usually do not allow to collect sufficiently many unique pairs of speakers to satisfy this assumption. As a result, the

Algorithm 1

Input: Dataset with speaker labels

Result: $\mathrm{P}_{\mathrm{FA}}(\tau)$

for $i = 1...T$ **do**

Select random enrolled (target) speaker, $\boldsymbol{y}_{\mathrm{e}}^{(i)}$

Select random test speaker, $\boldsymbol{y}_{\mathrm{t}}^{(i)}$

Compute L_i scores $\{s_l\}$ between $\boldsymbol{y}_{\mathrm{e}}^{(i)}$ and $\boldsymbol{y}_{\mathrm{t}}^{(i)}$

Compute the MC estimate of the speaker-pair conditioned false alarm probability:

$$\mathrm{P}_{\mathrm{FA}}^{(i)}(\tau) \approx \frac{1}{L_i} \sum_{l=1}^{L_i} \mathbb{I}\{s_l > \tau\},$$

end

Compute the MC estimate of $\mathrm{P}_{\mathrm{FA}}(\tau)$:

$$\mathrm{P}_{\mathrm{FA}}(\tau) \approx \frac{1}{T} \sum_{i=1}^{T} \mathrm{P}_{\mathrm{FA}}^{(i)}(\tau)$$

estimate may be biased if some speaker pairs have disproportionately large number of trials compared to the rest. The estimator in (6) compensates this bias by assigning weights to the terms in the sum which are inversely proportional to the number of trials. A more in-depth discussion of data dependence in speaker recognition evaluation can be found in [6].

2.3. Worst-Case False Alarm Rate With N Impostors

As (6) suggests, the probability of accepting an impostor speaker can be estimated by averaging the speaker-pair conditioned false alarm probabilities. In particular, Algorithm 1 repeats simulation of the zero-effort attack scenario where an impostor speaker is selected at random from the general population.

We propose a new characteristic of ASV systems which generalizes $\mathrm{P}_{\mathrm{FA}}(\tau)$ to attack scenarios where an impostor speaker is selected among N speakers with the intention to fool an ASV system. We call it the *worst-case false alarm rate with N impostors*, denoted by $\mathrm{P}_{\mathrm{FA}}^{N}(\tau)$. Algorithm 2 outlines the steps to estimate $\mathrm{P}_{\mathrm{FA}}^{N}(\tau)$. Here, similarity$(\cdot,\cdot)$ is an arbitrary similarity measure between speakers. The similarity function could be defined, for instance, in a speaker embedding space. In this work, all our models are defined in

the score domain. One possible strategy to select the closest speaker is to sample N sets of scores from $p(s|\boldsymbol{y}_{\mathrm{e}}^{(i)}, \boldsymbol{y}_{\mathrm{t},j}^{(i)})$ for $j = 1, \ldots, N$ and select the set with the highest mean value. We adopt this strategy. Figure 2 illustrates progression of Algorithm 2.

Algorithm 2

Input: Dataset with speaker labels

Result: $\mathrm{P}_{\mathrm{FA}}^{N}(\tau)$

for $i = 1...T$ **do**

Select random enrolled (target) speaker, $\boldsymbol{y}_{\mathrm{e}}^{(i)}$

Select N random test speakers, $\boldsymbol{y}_{\mathrm{t},1}^{(i)}, \boldsymbol{y}_{\mathrm{t},2}^{(i)}, ..., \boldsymbol{y}_{\mathrm{t},N}^{(i)}$

Find the closest speaker $\boldsymbol{y}_{\mathrm{t},k}^{(i)}$, where

$$k = \arg\max_{j} \text{similarity}(\boldsymbol{y}_{e}^{(i)}, \boldsymbol{y}_{\mathrm{t},j}^{(i)})$$

Compute L_i scores $\{s_l\}$ between $\boldsymbol{y}_{\mathrm{e}}^{(i)}$ and $\boldsymbol{y}_{\mathrm{t},k}^{(i)}$

Compute the MC estimate of the speaker-pair conditioned false alarm probability:

$$\mathrm{P}_{\mathrm{FA}}^{(i)}(\tau) \approx \frac{1}{L_i} \sum_{l=1}^{L_i} \mathbb{I}\{s_l > \tau\},$$

end

Compute the MC estimate of $\mathrm{P}_{\mathrm{FA}}^{N}(\tau)$:

$$\mathrm{P}_{\mathrm{FA}}^{N}(\tau) \approx \frac{1}{T} \sum_{i=1}^{T} \mathrm{P}_{\mathrm{FA}}^{(i)}(\tau)$$

Algorithm 2 reduces to the zero-effort imposture case if $N = 1$, or if one selects a *random* (among N available) test speaker, rather than the closest one to the enrolled speaker. Figure 3 demonstrates differences between these cases.

2.4. Performance Extrapolation Through Generative Model of Scores

Note that in the above strategy, the value of N is limited by the number of speakers in the dataset. Here, we describe an approach to extrapolate $\mathrm{P}_{\mathrm{FA}}^{N}(\tau)$ for values of N greater than the number of speakers in a dataset. Our main assumption is to approximate the speaker-pair conditioned score distribution $p(s|\boldsymbol{y}_{\mathrm{e}}, \boldsymbol{y}_{\mathrm{t}})$ as a (univariate) Gaussian. It should be noted that this assumption, by itself, does not put too many constraints on the shape of the distribution $p(s|\mathrm{non})$ which can be asymmetric and/or heavy-tailed.

In the sequel we describe a probabilistic model of between-speaker scores which follows

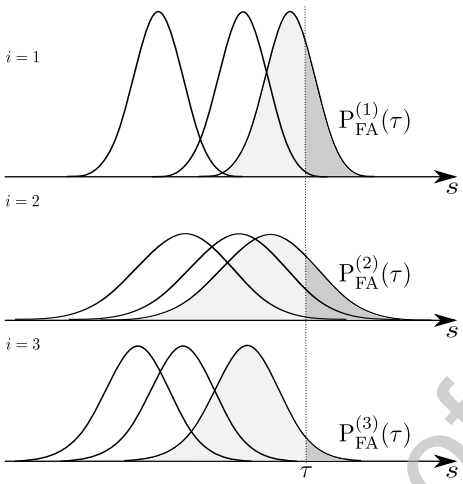


Figure 2: Illustration of a few iterations of the Algorithm 2 in the case of $N = 3$. At each iteration the algorithm samples N non-target speakers. The corresponding speaker-pair conditioned score distributions are depicted as Gaussians. The algorithm selects a distribution with the largest mean value. This distribution is shown as the one with shaded area under the curve. Finally, the algorithm computes the probability of the score being above the decision threshold τ for the selected distribution. This probability, denoted as $\mathrm{P}_{\mathrm{FA}}^{(i)}$, equals the area under the curve to the right of τ. Since the score distributions are not available in practice, one can only compute the empirical estimates of $\mathrm{P}_{\mathrm{FA}}^{(i)}$.

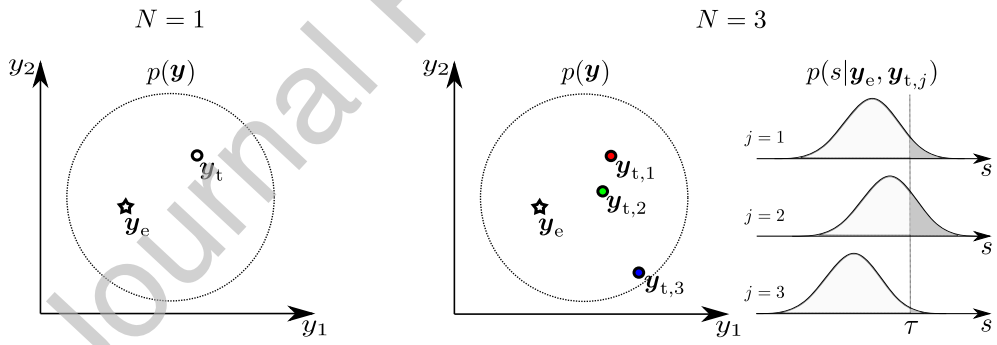


Figure 3: Illustration of two evaluation scenarios where an impostor speaker is selected among $N = 1$ (zero-effort attack) and $N = 3$ impostor speakers. (Left and Middle) Latent speaker identity space. Star represents an enrolled speaker and circles correspond to the impostor speakers. (Right) Distributions of scores between the enrolled speaker $\boldsymbol{y}_{\mathrm{e}}$ and each of the N impostor speakers $\boldsymbol{y}_{\mathrm{t},j}$ for $j = 1, \ldots, N$.

the generative process in Algorithm 2. It will allow to obtain estimates of $\mathrm{P}_{\mathrm{FA}}^{N}(\tau)$ for arbitrary values of N. We introduce two sets of latent variables: $\boldsymbol{\eta}_{\mathrm{e}}$ and $\boldsymbol{\eta}_{\mathrm{t}}$. The variables $\boldsymbol{\eta}_{\mathrm{e}}$ are *shared* among N speaker pairs and represent individual characteristics

of the enrolled speaker $\boldsymbol{y}_{\mathrm{e}}$. The variables $\boldsymbol{\eta}_{\mathrm{t},j}$, in turn, are responsible for differences between score distributions within a set of test speakers $\boldsymbol{y}_{\mathrm{t},j}$.

The proposed probabilistic model consists of the distribution of observations $p(s|\boldsymbol{\eta}_{\mathrm{e}}, \boldsymbol{\eta}_{\mathrm{t}})$, which is assumed to be Gaussian, and the prior distribution of latent variables $p(\boldsymbol{\eta}_{\mathrm{e}}, \boldsymbol{\eta}_{\mathrm{t}}) = p(\boldsymbol{\eta}_{\mathrm{t}}|\boldsymbol{\eta}_{\mathrm{e}})p(\boldsymbol{\eta}_{\mathrm{e}})$. Assuming that one can generate random samples of these variables, sampling scores from the model can be done according to the following steps (index i is omitted for clarity):

1. sample $\boldsymbol{\eta}_{\mathrm{e}} \sim p(\boldsymbol{\eta}_{\mathrm{e}})$
2. sample $\boldsymbol{\eta}_{\mathrm{t},j} \sim p(\boldsymbol{\eta}_{\mathrm{t}}|\boldsymbol{\eta}_{\mathrm{e}})$ for $j = 1...N$
3. sample N sets of scores $\mathcal{S}_j = \{s_{j,l}\}$, where $s_{j,l} \sim p(s|\boldsymbol{\eta}_{\mathrm{e}}, \boldsymbol{\eta}_{\mathrm{t},j})$

We consider a particular instance of such model where $\boldsymbol{\eta}_{\mathrm{t}} = \{\mu\}$, $\boldsymbol{\eta}_{\mathrm{e}} = \{m, \lambda, \sigma^2\}$ and the joint probability density function of the observed score and latent variables is factorized as follows

$$p(s, \boldsymbol{\eta}_{\mathrm{e}}, \boldsymbol{\eta}_{\mathrm{t}}) = p(s|\mu, \sigma^2)p(\mu|m, \lambda, \sigma^2)p(m)p(\lambda)p(\sigma^2).$$

The individual factors are outlined below:

$$p(s|\mu, \sigma^2) = \mathcal{N}(s|\mu, \sigma^2)$$
$$p(\mu|m, \lambda, \sigma^2) = \mathcal{N}(\mu|m, \sigma^2/\lambda)$$
$$p(m) = \mathcal{N}(m|\mu_0, \sigma_0^2)$$
$$p(\lambda) = \mathrm{Gam}(\lambda|\alpha_\lambda, \beta_\lambda)$$
$$p(\sigma^2) = \mathrm{InvGam}(\sigma^2|a_\sigma, b_\sigma).$$

Here, $\boldsymbol{\theta} = \{\mu_0, \sigma_0^2, a_\sigma, b_\sigma\, \alpha_\lambda, \beta_\lambda\}$ are hyper-parameters which can be estimated on the training set of scores formed according to Algorithm 2. Given hyper-parameters, the model can be used to predict $\mathrm{P}_{\mathrm{FA}}^N(\tau)$ for arbitrary values of N using Algorithm 3. It differs from Algorithm 2 in a way that the observed scores are replaced by samples from a generative model meant to approximate the unknown distribution of scores. In the special case of the proposed model the $\mathrm{P}_{\mathrm{FA}}^N(\tau)$ can be estimated without explicit sampling of scores. The assumption of the speaker-pair conditioned score distribution

being Gaussian allows to compute the estimate as

$$\mathrm{P}_{\mathrm{FA}}^{N}(\tau) \approx \frac{1}{T}\sum_{i=1}^{T} 1 - \Phi(\tau|\max_{j=1\ldots N}(\{\mu_{i,j}\}), \sigma_i^2),$$

where m_i, λ_i, σ_i^2 and $\mu_{i,j}$ are sampled from the corresponding distributions. Here, $\Phi(\cdot)$ denotes cumulative distribution function of the Gaussian distribution.

Algorithm 3

Input: Generative model of scores

Result: $\mathrm{P}_{\mathrm{FA}}^{N}(\tau)$

for $i = 1\ldots T$ **do**

Sample N sets of scores from the model: $\mathcal{S}_j$ for $j = 1\ldots N$

Find the set with the highest mean score

$$k = \arg\max_j \mathrm{mean}(\mathcal{S}_j)$$

Compute the MC estimate of the speaker-pair conditioned false alarm probability:

$$\mathrm{P}_{\mathrm{FA}}^{(i)}(\tau) \approx \frac{1}{|\mathcal{S}_k|}\sum_{l=1}^{|\mathcal{S}_k|}\mathbb{I}\{s_l > \tau\}, s_l \in \mathcal{S}_k$$

end

Compute the MC estimate of $\mathrm{P}_{\mathrm{FA}}^{N}(\tau)$:

$$\mathrm{P}_{\mathrm{FA}}^{N}(\tau) \approx \frac{1}{T}\sum_{i=1}^{T}\mathrm{P}_{\mathrm{FA}}^{(i)}(\tau)$$

This model assumes shared variance among score distributions $p(s|\boldsymbol{y}_{\mathrm{e}}, \boldsymbol{y}_{\mathrm{t},j})$ for $j = 1, \ldots, N$ given a target speaker $\boldsymbol{y}_{\mathrm{e}}$. This assumption as well as the choice of specific distributions are primarily motivated by the convenience of computing the posterior distribution of the latent variables. In particular, using *conjugate pairs* of distributions [31] as building blocks in the model allows to devise efficient algorithms to obtain approximate posterior distribution. This leads to closed-form updates in the *expectation-maximization* (EM) algorithm [32] used to estimate the model hyper-parameters, with the details provided in Appendix I. Further insight to the form of the score distributions implied by our model (including its limitations) is provided in Appendix II. In Section 5 we provide discussion of the adequacy of the model assumptions and potential alternatives.

Figure 4 depicts the Bayesian network of the proposed model. A Bayesian network is a directed graphical model [1] that represents a set of random variables and their condi-

tional dependencies via a directed acyclic graph. Empty circles denote latent variables, shaded circles denote observed variables and nodes without circles denote deterministic parameters. A group of nodes surrounded by a box, called a plate, labeled with T indicates that the subgraph inside a plate is duplicated T times [33]. The arrows between the nodes point from the parent variables to their children variables and represent the conditional dependencies between these variables.

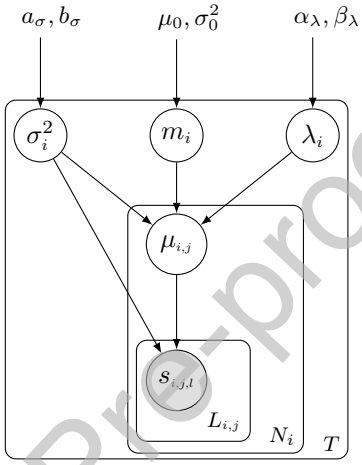


Figure 4: A graphical representation of the generative model. Here, T is the number of target speakers, N_i is the number of non-target speakers for the i^{th} target speaker, and $L_{i,j}$ is the number of similarity scores between the i^{th} target speaker and the j^{th} non-target speaker.

3. Experimental Setup

This section describes the ASV systems, protocols, and the dataset we use for the experiments with the proposed worst-case false alarm rate with N impostors (P_{FA}^N) metric.

3.1. Dataset

A suitable dataset for our experiments has to fulfill two requirements. First, it must have a large number of speakers to not only train well-performing ASV systems, but to have enough speakers in the evaluation side to produce good P_{FA}^N estimates from the ASV scores. Second, each speaker in the evaluation side should have enough utterances to produce a sufficiently large number of scores between each pair of speakers, required for

reliable P_{FA}^{N} estimation. For these reasons, we chose the VoxCeleb datasets (VoxCeleb1 [24] & VoxCeleb2 [25]). When combined, the datasets contain 7365 speakers and, on average, each speaker has well over 100 utterances, which typically originate from about 20 sessions.

We divided the available speakers into three disjoint sets containing 5345, 40, and 2000 speakers. The first set of 5345 speakers is used to train the ASV systems. The second set of 40 speakers consists of the test speakers in the standard VoxCeleb1 ASV evaluation protocol, which is used for evaluating performance of our ASV systems. The third, gender-balanced set contains 1000 male and 1000 female speakers and is used for the experiments with P_{FA}^{N} estimation. The speakers in this last set were chosen so that each had utterances from at least 18 different sessions; otherwise the split between the first and the last set was random.

3.2. Automatic Speaker Verification Systems

We provide experimental results for two different ASV systems, based on the two most commonly used speaker embeddings, i-vectors [22] and x-vectors [23]. We trained both systems using Kaldi [34] recipes for VoxCeleb using our custom train-test data division. Both systems use mel-frequency cepstral coefficients (MFCCs) as acoustic features and a combination of *linear discriminant analysis* (LDA) and *probabilistic* LDA (PLDA) in the scoring backend. The fundamental difference between the i-vector and x-vector systems is that the former is based on Gaussian generative model, while the latter is trained discriminatively and utilizes longer time context via time-delay neural network. Another major difference is that the x-vector system is trained with a larger training set leveraging from data augmentation. For further details of the systems, refer to Table 1.

3.3. Evaluation Protocols

We used two ASV protocols to serve two different purposes. First, we adopted the standard VoxCeleb1 ASV protocol to assess the performance of our ASV systems. This protocol contains 40 speakers, and 37720 evaluation trials with a balanced number of target (same speaker) and non-target (different speaker) trials. The second protocol is used to obtain a large number of non-target scores for a large number of speaker pairs to estimate P_{FA}^{N}. For each of the 2000 speakers in the testing set, we randomly chose 18

Table 1: Details of the ASV systems used in this study.

	i-vector system	x-vector system
Acoustic features	24-dimensional MFCCs + delta + double-delta coefficients; energy based speech activity detection	30-dimensional MFCCs; energy based speech activity detection
Background model	Gaussian mixture model of 2048 components with full covariance matrices; trained using the whole training data	—
Embedding extractor	Trained with 100 000 longest utterances in the training set	Trained using the whole training data plus 1 000 000 utterances obtained by data augmentation (reverb, noise, babble, music)
Embeddings	400-dimensional i-vectors	512-dimensional x-vectors
LDA and PLDA	Both trained using the whole training data; dimensionality reduction to 200-D with LDA	Both trained using the whole training data; dimensionality reduction to 200-D with LDA

utterances so that all the utterances were from different sessions. Then, for each pair of speakers, we obtained $18^2 = 324$ trials by forming all the utterance pairs between the two speakers. In total, we had $1\,999\,000 \cdot 324 = 647\,676\,000$ trials, where $1\,999\,000$[3] is the total number of unique speaker pairs. The above number includes cross-gender trials. Including only speaker pairs within one gender, we have 161 838 000 trials for both males and females.

[3] $2000!/(2!(2000-2)!) = 1\,999\,000$ (number of 2-combinations in a set of 2000 speakers)

4. Results

4.1. Performance of Speaker Verification Systems

Table 2: Parameters of three different detection cost functions (DCF) used in this study. The system thresholds (τ) that mimimize these DCFs are used to estimate false alarm rates in the following experiments.

	P_{target}	C_{miss}	C_{fa}
minDCF_1	0.5	10	1
minDCF_2	0.5	1	1
minDCF_3	0.5	1	10

Before proceeding to our proposed generative approach, we validate correctness of the ASV implementations through standard performance metrics. To this end, we report *equal error rate* (EER) and *minimum normalized detection cost function* (minDCF) [35]. EER is obtained by setting the system threshold τ so that false alarm and miss rates equal each other. The threshold selection for minDCF, in turn, is governed by the parameters P_{target} (prior probability of target speaker), C_{miss} (cost of missing the target speaker), and C_{fa} (cost of falsely accepting a non-target speaker). For this study, we adopt three different sets of parameters (Table 2): the first set has high cost for misses, the second set has equal costs for misses and false alarms, and the last one penalizes false alarms more. From the security perspective, DCF_3 is the most relevant, whereas the other two DCFs can be utilized in applications where high security is not required.

Table 3 shows the EERs and minDCFs for i-vector and x-vector systems using VoxCeleb test protocol (category 'all'). In addition, we split the protocol based on genders and also report a result for the 'pooled' category, which does not contain inter-gender trials making the original test protocol more difficult. Our results are in line with the results reported in the original Kaldi recipes. As we used about 2000 speakers less for system training, our EER for x-vector system is about 0.5% (absolute) higher than what is reported in the original recipe.

In addition to the overall performance difference between i-vector and x-vector systems, these systems differ in their ability to recognize speakers from different genders.

For the i-vector system, the performance for males is considerably better, whereas for the x-vector system the difference between the genders is smaller.

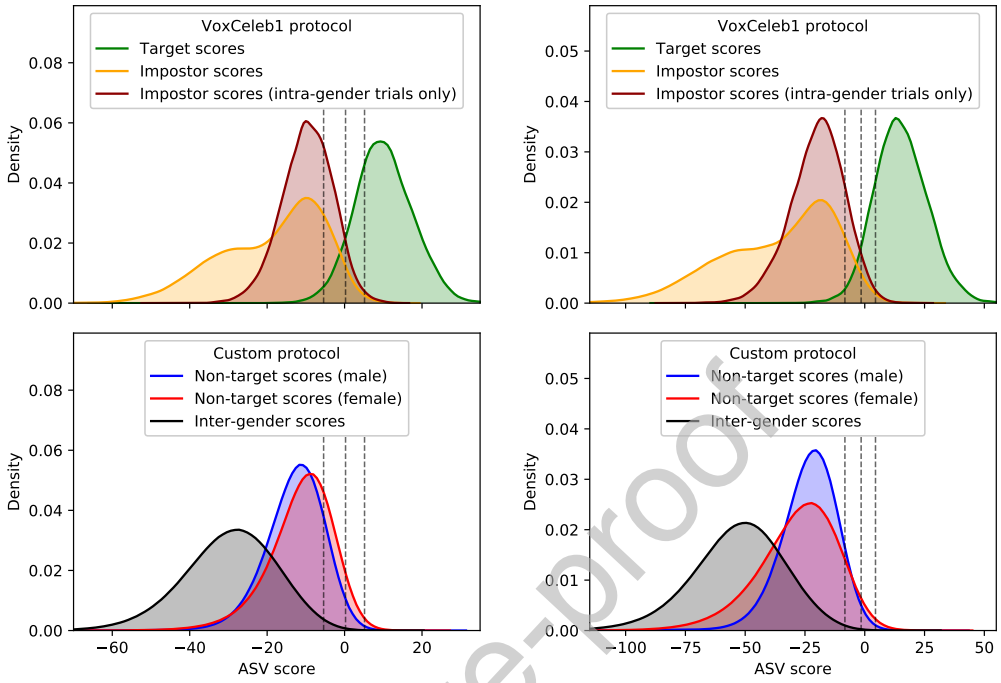
In Figure 5, we display score distributions for the VoxCeleb1 test protocol and for our custom protocol containing non-target scores only. The VoxCeleb1 protocol is used to set the system thresholds τ for the $\mathrm{P}_{\mathrm{FA}}^{N}$ estimation experiments presented in the next section. In these experiments, we use gender-specific thresholds obtained via minimizing DCF separately on male and female trials. Note that for clarity, Figure 5 does not show gender-specific thresholds, but instead it shows the thresholds for 'pooled' category.

4.2. Estimation of Worst-Case False Alarm Rates

We estimated worst-case false alarm rates empirically using Algorithm 2 by randomly selecting enrolled speaker $T = 1000$ times. Similarly, we use $T = 1000$ in Algorithm 3 to obtain model-based estimates. The estimates are shown in Figure 6 for both ASV systems using three different thresholds obtained using the DCF parameter sets in Table 2. We find that model-based approaches give good estimates when the threshold is low

Table 3: Performance of i-vector and x-vector systems on VoxCeleb1 test protocol. The original protocol ('all') contains both intra- and inter-gender trials. The numbers under the category 'pooled' are computed using only intra-gender trials from both genders.

	minDCF_1	minDCF_2	minDCF_3	EER (%)
i-vector				
male	0.43	0.14	0.31	6.97
female	0.53	0.17	0.37	8.80
pooled	0.44	0.14	0.34	7.18
all	0.30	0.11	0.27	5.62
x-vector				
male	0.27	0.09	0.24	4.71
female	0.30	0.10	0.24	5.19
pooled	0.28	0.09	0.23	4.75
all	0.21	0.07	0.19	3.61

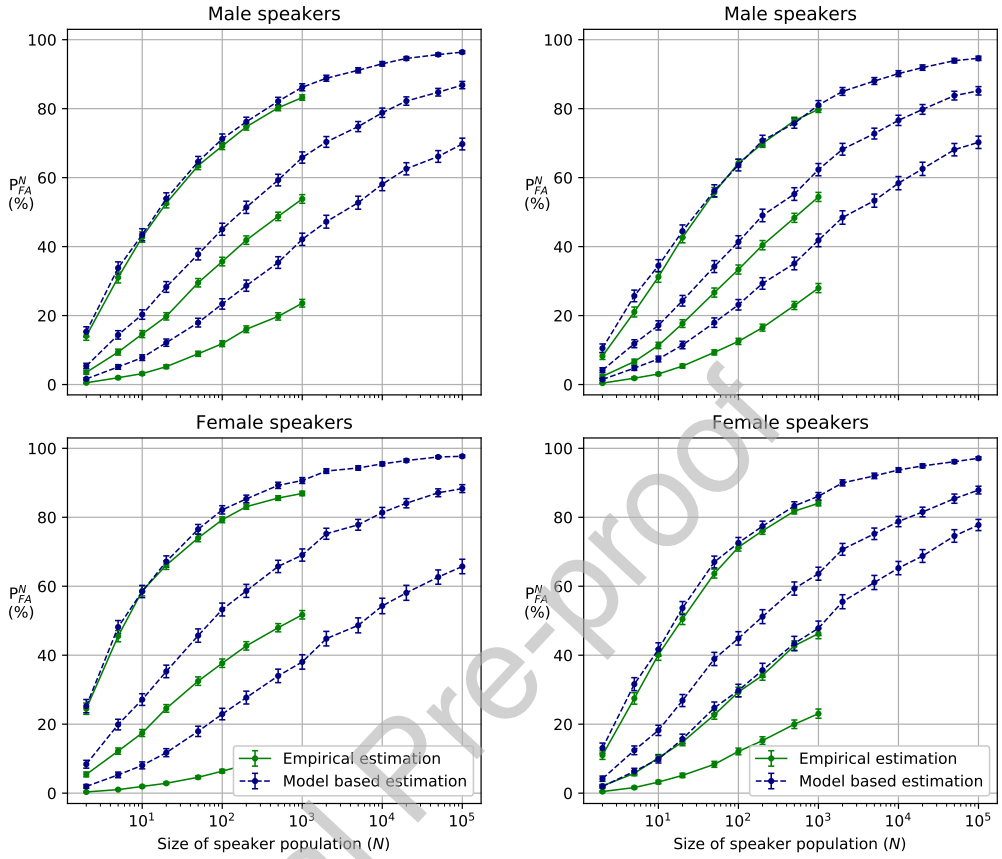


(a) i-vector. (b) x-vector.

Figure 5: Score distributions for the standard VoxCeleb protocol and the custom protocol obtained using i-vector and x-vector systems. Dashed lines represent minDCF thresholds for 'pooled' scores (see Table 3) using different sets of cost parameters presented in Table 3.

(higher cost for misses). When the threshold is higher than in the minDCF_1 case, the model-based estimates can be seen as a conservative upper bounds for the P_{FA}^N rates. We also find that the differences between the empirical and the model-based estimates are greater for females than for the males. To obtain further insight, we depict the score distributions of the closest impostors for population size of $N = 1000$ in Figure 7. The figure indicates that especially for females, the model-based score distributions tend to be too wide and slightly shifted to the right, which causes higher false acceptance rates when a high threshold value is used.

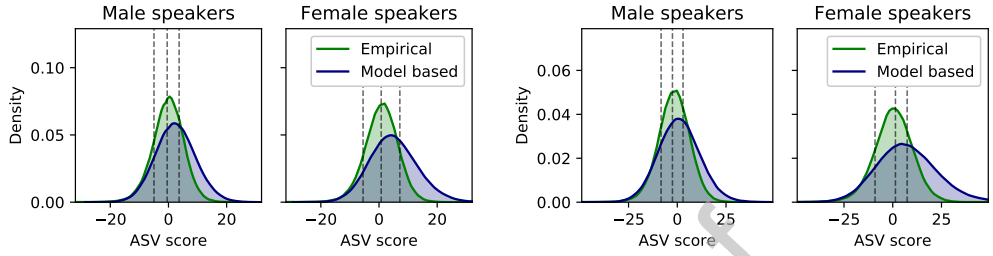
Using the estimates, we can predict that for the minDCF_1 threshold, P_{FA}^N is $95 - 98\%$ for an impostor population of size 100 000. For minDCF_3 threshold, we can rely only on the empirical estimates, which tell us that, depending on the system and gender, P_{FA}^N



(a) i-vector. (b) x-vector.

Figure 6: Empirical and model-based estimates of worst-case false alarm rates with N impostors for various sizes of speaker populations. In each plot, three empirical and model-based estimates are shown for three different thresholds. These thresholds are obtained separately for each plot using cost parameters defined in Table 2. The curves from top to bottom correspond to the thresholds of minDCF_1, minDCF_2, and minDCF_3, respectively. The model-based estimates follow closely emprical estimates for low threshold values, but as the threshold gets stricter, the difference between the empirical and model-based estimates grows. The curves are obtained using $T = 1000$ in Algorithms 2 and 3. The mean values obtained using these algorithms are shown together with their 99% confidence intervals.

rate of 12 – 28% is obtained for a population of size 1000. Note that our populations contain only speakers from one gender. If the population would contain speakers from both genders, the false alarm rates would be lower.



(a) i-vector. (b) x-vector.

Figure 7: Score distributions of the closest impostors ($N = 1000$) pooled together from $T = 1000$ samplings/simulations for the empirical and model-based approaches. Dashed lines represent minDCF thresholds obtained using different sets of cost parameters, which are presented in Table 3. For the strictest threshold, the area under the density curves on the right side of the threshold is larger for the model-based estimation, which explains why the model-based estimation lead to higher false alarm rates as shown in Figure 6.

5. Discussion

Before concluding, the authors would like to address two relevant concerns, the over-estimated false alarm rates at high threshold, and the worst-case attack assumption.

5.1. Analysis of the model-based worst-case false alarm estimation

From Figure 7, we observe two apparent problems in the score distributions given by the model for the closest impostors: a) they are shifted to the right and b) they have too large variances. As a result, some of the generated scores of the closest impostors are too high, which results in over-estimated false alarm rates given by the model. We have identified three causes for the problems.

First, we found that the empirical distribution of μ, which is the distribution of score means of speaker pairs, is skewed to the left (negative skewness, see Table 4). Consequently, the fitted normal distribution (assumed in the model) has longer tail on

Table 4: Sources of mismatch between the observed and generated scores. The generative score model assumes that pairwise non-target scores and means of pair-wise scores (μ) are normally distributed, while the analysis shows that they are skewed to the left. Additionally, scores with the closest impostors tend to have smaller variances than scores with random impostors, which is not factored into the model.

	i-vector system		x-vector system	
	males	females	males	females
Avg. skewness of pairwise scores	-0.20	-0.29	-0.20	-0.27
Skewness of μ	-0.86	-0.99	-0.99	-0.62
Avg. STDEV* of scores with the closest impostors	5.1	5.5	7.5	9.4
Avg. STDEV* of scores with random impostors	6.2	7.1	9.6	13.7

* Computed as a square root of an average of variances.

the right than what the original score data had. The right tail of the distribution of μ is where we will find the closest impostors in Algorithm 3. As a result, the scores of the closest impostors are shifted to the right.

Additionally, we observed that the variation in target-vs-impostor scores is smaller for speaker pairs with the closest impostors than for random impostors. In other words, the closer the impostor's voice is to the target speaker's voice, the smaller is the variance in scores between the two speakers. As our model does not take this into account, the speaker pairs with closest impostors tend to have too large score variances.

Finally, the scores between the speaker pairs are also skewed to the left, which is another source of mismatch between empirical scores and scores generated by the model.

These observations open two potential directions towards increasing the prediction accuracy. The first direction is to revise the proposed generative model to take into account the skew of score distributions, as well as by relaxing the assumption of a shared variance. This can be done at the cost of losing conjugacy between distributions in the model, leading to increased computational complexity of hyper-parameter estimation. An alternative, second direction would be supervised fine-tuning to optimize some loss function between the empirical and the model-based estimates. As our model is parameterized only by six numbers, we believe that a good hyper-parameter configuration can be

found in reasonable time using one of the *derivative-free* optimization methods [36]. To give some empirical evidence for this claim, Fig. 8 displays an example where we tuned our model parameters *manually*. As seen, the model itself is actually flexible enough to fit the empirical false alarm rates accurately — but the purely generative training criterion does not find the parameter values that achieve this. With the manually corrected model, we obtained 54% worst-case false acceptance rate estimate for population size of 100.000 for the strictest minDCF_3 threshold, whereas the original model clearly over-estimated this by giving FA rate of 70% as shown in the top-right panel of Figure 6.

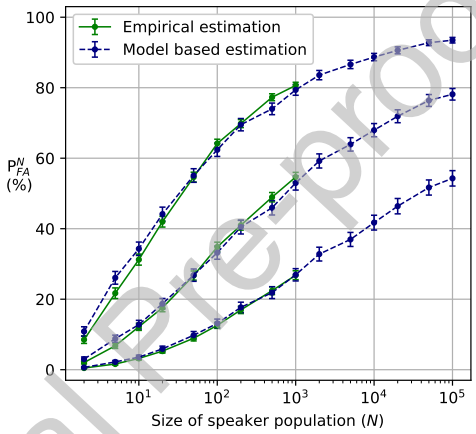
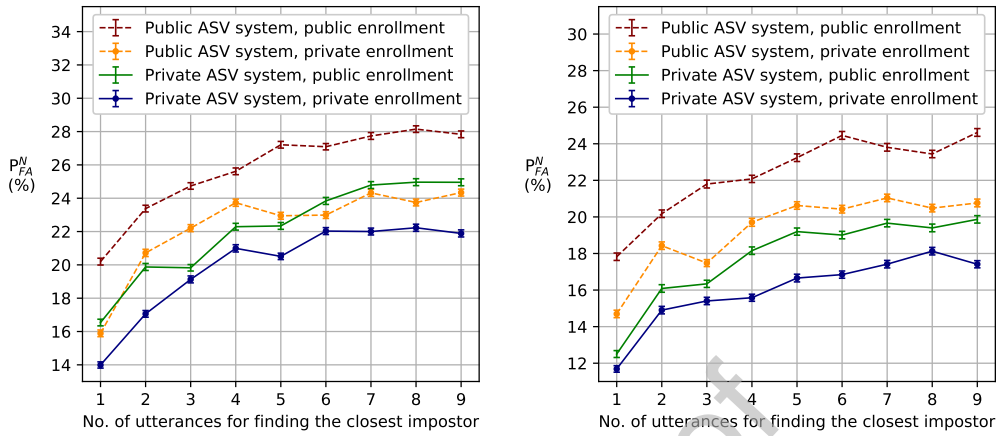


Figure 8: Worst-case false alarm estimates for male scores given by x-vector system after tweaking the model hyper-parameters manually. First, the parameter α_λ was adjusted until the variances of distributions in Figure 7 matched and then μ_0 was adjusted to fix the shifting misalignment. As a result, the model-based estimates follow closely the empirical estimates unlike in Figure 6.

5.2. False alarm estimation in simulated attack scenarios

So far we have considered worst-case false alarm estimation from the system deployer's perspective. From the presented results, we can gain understanding on how many enrolled speakers systems with specific thresholds can handle without starting to confuse speakers to each other too much.

Next, let us consider a scenario, in which a malicious attacker is utilizing ASV technology to find similar sounding speakers to the enrolled target speaker's voice to break



(a) Male speakers.

(b) Female speakers.

Figure 9: Empirical estimation of P^N_{FA} ($N = 1000$) in various scenarios for x-vector system with minDCF_3 threshold (see Table 2). 'Public ASV system' refers to a case where the closest impostors in Algorithm 2 are selected using the same x-vector system. To simulate a scenario, where attacker uses another ASV system for impostor selection due to not knowing the details of the deployed system ('private ASV system'), an i-vector system is used to select the closest impostors. Further, 'public enrollment' and 'private enrollment' refer to such cases, where the attacker has access (public) or does not have access (private) to the enrolled target speaker's enrollment data. If the enrollment data and the ASV system are public, the selection of the closest impostor is easier, which results in higher false acceptance rates.

the ASV system. As discussed in Section 1, the previously presented results can be considered as the worst-case situation, where the attacker has access to both the deployed ASV system as well as to the target speaker's enrollment data. In reality, the attacker would be unlikely to have access to either of them. Instead, the attacker would first have to set up another ASV system and then collect some speech data from the target speaker to perform the speaker search. These steps will make the attack more difficult as the closest impostor obtained using attacker's system and data might not be the same as what would be the closest impostor when using the attacked system and the real enrollment data.

To study the effect of system/data mismatch in the impostor selection, we set up a following experiment. First, we divided the available 18 utterances for each speaker into two disjoint sets of nine utterances. The first one was used for impostor selection and

the second for speaker enrollment. We compared this setup to a case, where the same set of nine utterances was used both for impostor selection and enrollment to address the effect of data mismatch. Further, we also varied the number of utterances used for impostor selection from one to nine to see the effect of the amount of data used for impostor search. We simulated the ASV system mismatch by using i-vector system to select closest impostors, while the x-vector system was considered to be the attacked system. This was compared to the case where the impostor search was done using the same attacked x-vector system.

The results are shown in Figure 9, which reveals the expected patterns: when there is no data mismatch or ASV system mismatch, false acceptance rates are highest, which means that the attacks are most successful. If there is either data mismatch or system mismatch, the false acceptance rates drop. The lowest false acceptance rates are obtained, when both types of mismatches are present and when the number of utterances available for impostor search is low.

As another future direction, we consider designing a model for joint modeling of scores from two different ASV systems suitable for more realistic scenario where the attacker does *not* have access to the target speaker's enrollment data and the deployed ASV system.

6. Conclusions

Seamless integration of artificial intelligence to our daily lives, including speech technology products, raises growing concern of their trustworthiness and safety. Our study resides in the landscape of automatic speaker verification (ASV), or voice biometrics security. One unique feature of voice (and face) biometrics is that, unlike traditional physical biometrics — fingerprints, iris, retina, DNA to name a few — is that much of the biometric data is *publicly available in the Internet* through social media, news, interviews, lectures, and workplace websites to name a few. An important concern is the relation of false alarm (false acceptance) and database size: regardless of the selected ASV technology, given a large enough database, one will eventually have speaker collisions. A technology-aware attacker may increase the likelihood of such collisions through the use of public-domain ASV system to identify target speakers from a public database

[20]. Even without dedicated attacks, however, the number of different voices (subject to extrinsic and intrinsic speech variations) is not infinite. Therefore, eventually, ASV performance will be capped at some database size.

The methodology concept put forward in this study gives us novel tool to address the dependency of false alarm rate and database size beyond the size of a given evaluation corpus. The proposed model produced reasonable match with empirical scores and displayed the expected trend of increasing false alarm rate as a function of database size. Our model is general and can be applied to analyze (and optimize) any black-box ASV system to produce graphs similar to those in Fig. 6, based on detection scores only. As such, these graphs are *predictions* by the model — how the ASV system will behave if one were able to collect more speakers assuming the speaker sampling process remains the same. Even if it is not easy to experimentally validate the model beyond a given training corpus size, the general trends what we saw in our pilot experiments with VoxCeleb corpus are deemed as expected: false alarm rates increase as a function of database size and will eventually saturate.

Our work has a number of limitations as well. First, as the results indicate, the generative model overestimated the false alarm rates, especially for the high-security operating region (high threshold). In real-world deployment of ASV, the detection threshold τ needs to be optimized to achieve a desirable security–convenience trade-off based on some development data and application (DCF setting). If the false alarm rate is overestimated, the threshold τ would have to be increased (relative to the value it would have been set with precise knowledge of the false alarm rate), leading to decreased user convenience due to increased miss rate. Nonetheless, as Section 5 indicates, our proposed generative model is flexible enough to be adjusted so that the empirical and predicted false alarm rates will match closely. Our model design philosophy has been simplicity: all the parameters are automatically learned from data and we leverage from conjugate families of distributions to enable efficient inference. The suggested future improvements include revising our distributional assumptions, and combining generative modeling with discriminative fine-tuning.

Second, our model assumes a worst-case scenario where the attacker has access to the target speaker's enrollment data, as well as the attacked ASV system. This is no

different from standard NIST SRE style evaluations where an evaluator reports standard evaluation metrics (such as EER, minDCF, or P_{FA}) on a given, fixed evaluation corpus with known trial key. In future, we are interested in extending our generative model to model interaction between two different ASV systems and across different data domains.

Furthermore, it would be interesting to compare different ASV systems and database qualities. VoxCeleb data was selected for the experiments primarily due to the large number of speakers and the amount of intra-speaker scores. Nonetheless, being representative of *found Internet data*, VoxCeleb contains many style, channel and environment variations. At least for academic curiosity, it would be interesting to repeat our simulations on more controlled database for reference purposes. Further, it will be interesting to analyze the impacts of i-vector and x-vector dimensionality, dimensionality reduction of these embeddings, and speaker subspace size in PLDA. Finally, it would be interesting to apply the proposed methods to speaker diarization or other use cases within ASV, such as score normalization with increased speaker cohort size.

Acknowledgement

The work has been supported by Academy of Finland (proj. no. 309629 entitled "NOTCH: NOn-cooperaTive speaker CHaracterization") and by the Doctoral Programme in Science, Technology and Computing (SCITECO) of the UEF. The work of the first author was partially financially supported by the Government of the Russian Federation (Grant 08-08). A part of the work of the third author was supported by NEC internship program. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

Appendix I: parameter inference

In the following we describe the algorithm used to estimate hyper-parameters of the proposed generative model. We find the values of hyper-parameters $\boldsymbol{\theta}$ that maximize the likelihood function – the joint probability density of the observed data viewed as a

function of $\boldsymbol{\theta} = \{\mu_0, \sigma_0^2, a_\sigma, b_\sigma, \alpha_\lambda, \beta_\lambda\}$:

$$\mathscr{L}(\boldsymbol{\theta}) = \int \prod_{i=1}^{T} \mathcal{N}(m_i|\mu_0, \sigma_0^2)\mathrm{Gam}(\lambda_i|\alpha_\lambda, \beta_\lambda)\mathrm{InvGam}(\sigma_i^2|a_\sigma, b_\sigma)$$
$$\prod_{j=1}^{N_i} \mathcal{N}(\mu_{i,j}|m_j, \lambda_i, \sigma_i^2) \prod_{l=1}^{L_{i,j}} \mathcal{N}(s_{i,j,l}|\mu_{i,j}, \sigma_i^2)\, \mathrm{d}m_i\, \mathrm{d}\lambda_i\, \mathrm{d}\sigma_i^2\, \mathrm{d}\mu_{i,j}$$

For many probabilistic models with latent variables, including the proposed one, this objective function is intractable (cannot be evaluated). A commonly adopted strategy to avoid this obstacle is to use the *expectation-maximization* (EM) algorithm [32], an iterative optimization method to find the local extrema of the likelihood function. The EM algorithm alternates between two steps: **expectation** step (E-step) and **maximization** step (M-step). On the E-step it computes (approximate) posterior distribution of the latent variables and on the M-step it updates all the hyper-parameters of the model.

Inference for latent variable models can be conducted through *variational Bayes* [1, Chapter 10] or *Monte-Carlo* techniques [1, Chapter 11]. We choose the former approach due to its better scalability in terms of computational costs. The variational Bayesian inference approximates the exact posterior distribution $p(\{m_i\}, \{\lambda_i\}, \{\sigma_i^2\}, \{\mu_{i,j}\}|\{s_{i,j,l}\})$ by a *variational* distribution q from a restricted family of distributions. The variational distribution is found by minimizing the Kullback-Leibler divergence from the posterior distribution.

One of the commonly adopted strategies, known as *black-box variational inference* (BBVI) [37], is to explicitly define the family of variational distributions and use stochastic optimization [38] to minimize the objective. Another strategy to define q is the *mean-field* approximation [1, 39] which assumes the variational distribution to be fully factorized:

$$q(\{m_i\}, \{\lambda_i\}, \{\sigma_i^2\}, \{\mu_{i,j}\}) = \prod_{i=1}^{T} q(m_i)q(\lambda_i)q(\sigma_i^2) \prod_{j=1}^{N_i} q(\mu_{i,j}) \tag{8}$$

but with no further assumptions imposed on the functional forms of the factors. For conditionally conjugate models [40] this approach leads to closed form solutions in a coordinate descent optimization algorithm which iteratively updates the parameters of one factor while holding the others fixed. Since the proposed model is conditionally conjugate we choose to use the mean-field approach due to its lower computational complexity. The

inference algorithm performs the following updates performed until convergence.

Expectation step (E-step):

- Updating $q(m_i)$:

$$q(m_i) = \mathcal{N}\left(m_i|\hat{m}_i, s_i^2\right)$$
$$s_i^2 = \left(N_i\mathbb{E}\left[\lambda_i\right]\mathbb{E}\left[\frac{1}{\sigma_i^2}\right] + \frac{1}{\sigma_0^2}\right)^{-1}$$
$$\hat{m}_i = \left(N_i\mathbb{E}[\lambda_i]\mathbb{E}\left[\frac{1}{\sigma_i^2}\right] + \frac{1}{\sigma_0^2}\right)^{-1}\left(\mathbb{E}[\lambda_i]\mathbb{E}\left[\frac{1}{\sigma_i^2}\right]\left(\sum_{j=1}^{N_i}\mathbb{E}\left[\mu_{i,j}\right]\right) + \frac{\mu_0}{\sigma_0^2}\right)$$
$$\mathbb{E}[m_i] = \hat{m}_i,\ \mathbb{E}[m_i^2] = \hat{m}_i^2 + s_i^2$$

- Updating $q(\sigma_i^2)$:

$$q(\sigma_i^2) = \text{InvGam}\left(\sigma_i^2|\hat{a}_i, \hat{b}_i\right)$$
$$\hat{a}_i = a_\sigma + \frac{N_i}{2} + \sum_{j=1}^{N_i} L_{i,j}$$
$$\hat{b}_i = b_\sigma + \frac{1}{2}\mathbb{E}\left[\sum_{j=1}^{N_i}\sum_{l=1}^{L_{i,j}}(s_{i,j,l} - \mu_{i,j})^2\right] + \frac{1}{2}\mathbb{E}[\lambda_i]\mathbb{E}\left[\sum_{j=1}^{N_i}(\mu_{i,j} - m_i)^2\right]$$
$$\mathbb{E}\left[\frac{1}{\sigma_i^2}\right] = \frac{\hat{a}_i}{\hat{b}_i},\ \mathbb{E}[\log\sigma_i^2] = \log\hat{b}_i - \psi(\hat{a}_i)$$

- Updating $q(\lambda_i)$:

$$q(\lambda_i) = \text{Gam}\left(\lambda_i|\hat{\alpha}_i, \hat{\beta}_i\right)$$
$$\hat{\alpha}_i = \alpha_\lambda + \frac{N_i}{2}$$
$$\hat{\beta}_i = \beta_\lambda + \frac{1}{2}\mathbb{E}\left[\frac{1}{\sigma_i^2}\right]\mathbb{E}\left[\sum_{j=1}^{N_i}(\mu_{i,j} - m_i)^2\right]$$
$$\mathbb{E}[\lambda_i] = \frac{\hat{\alpha}_i}{\hat{\beta}_i},\ \mathbb{E}[\log\lambda_i] = \psi(\hat{\alpha}_i) - \log\hat{\beta}_i.$$

- Updating $q(\mu_{i,j})$:

$$q(\mu_{i,j}) = \mathcal{N}\left(\mu_{i,j}|\hat{\mu}_{i,j}, \hat{s}^2_{i,j}\right)$$
$$\hat{\mu}_{i,j} = \frac{\sum_{l=1}^{L_{i,j}} + \mathbb{E}[\lambda_i]}{L_{i,j} + \mathbb{E}[\lambda_i]}$$
$$\hat{s}^2_{i,j} = \left(\mathbb{E}\left[\frac{1}{\sigma_i^2}\right](L_{i,j} + \mathbb{E}[\lambda_i])\right)^{-1}$$
$$\mathbb{E}[\mu_{i,j}] = \hat{\mu}_{i,j},\ \mathbb{E}[\mu^2_{i,j}] = \hat{\mu}_{i,j} + \hat{s}^2_{i,j}$$

Here, $\mathbb{E}[\cdot]$ denotes the expected value of a random variable.

Maximization step (M-step):

Given an approximate posterior distribution found on the E-step, the M-step proceeds by updating the hyper-parameters $\boldsymbol{\theta}$ as follows:

- Updating μ_0:

$$\mu_0 = \frac{1}{T}\sum_{i=1}^{T}\mathbb{E}[m_i]$$

- Updating σ_0^2:

$$\sigma_0^2 = \frac{1}{T}\sum_{i=1}^{T}(\mathbb{E}[m_i] - \mu_0)^2$$

- Updating α_λ and β_λ:

$$\alpha_\lambda, \beta_\lambda = \arg\max_{\alpha,\beta} T(\alpha\log\beta - \log\Gamma(\alpha)) + (\alpha-1)\sum_{i=1}^{T}\mathbb{E}[\log\lambda_i] - \beta\sum_{i=1}^{T}\mathbb{E}[\lambda_i]$$

- Updating a_σ and b_σ:

$$a_\sigma, b_\sigma = \arg\max_{a,b} T(a\log b - \log\Gamma(a)) + (a-1)\sum_{i=1}^{T}\mathbb{E}[\log\sigma_i^2] - b\sum_{i=1}^{T}\mathbb{E}\left[\frac{1}{\sigma_i^2}\right]$$

The last two updates are two-dimensional convex optimization problems. Their solutions can be obtained using numerical optimization algorithms specialized to these tasks [41, 42]. Our approach is a less elaborate version of [41] where we use general-purpose root-finding algorithms which can be found in any commonly-adopted mathematical library.

The EM algorithm repeats the E- and M-steps outlined above until the convergence. In our experiments we found that a few iterations are sufficient to reach a point where any further iterations do not substantially change the values of hyper-parameters.

Appendix II: Score Distribution of the Model is Approximately Gaussian

In the sequel we show how to obtain the marginal distribution of the observations

$$p(s) = \int p(s|\mu, \sigma^2)p(\mu|m, \lambda, \sigma^2)p(m)p(\lambda)p(\sigma^2)\,\mathrm{d}\mu\,\mathrm{d}m\,\mathrm{d}\lambda\,\mathrm{d}\sigma^2$$

by integrating out all the latent variables in the model one-by-one. We begin by noting that convolution of two Gaussians is another Gaussian with summed variances, to integrate out μ:

$$p(s) = \int \mathcal{N}(s|m, \sigma^2 + \sigma^2/\lambda)\mathcal{N}(m|\mu_0, \sigma_0^2)\mathrm{Gam}(\lambda|\alpha_\lambda, \beta_\lambda)\mathrm{InvGam}(\sigma^2|a_\sigma, b_\sigma)\,\mathrm{d}m\,\mathrm{d}\lambda\,\mathrm{d}\sigma^2$$

Further, since the inverse gamma distribution is a conjugate prior for Gaussian distribution with fixed mean, we arrive at the following:

$$p(s) = \int t_{2a_\sigma}(s|m, b_\sigma/a_\sigma(1 + 1/\lambda))\mathcal{N}(m|\mu_0, \sigma_0^2)\mathrm{Gam}(\lambda|\alpha_\lambda, \beta_\lambda)\,\mathrm{d}m\,\mathrm{d}\lambda$$

where $t_\nu(s|\eta, \varsigma^2)$ denotes the non-standardized t-distribution with ν degrees of freedom, mean η and variance ς^2. Since the t-distribution can be closely approximated by a Gaussian distribution, which is its limiting case when $\nu \to \infty$, even for moderate values of ν, we can approximate the score distribution by a continuous mixture of Gaussians with gamma as the mixing distribution:

$$p(s) \approx \int \mathcal{N}(s|\mu_0, \sigma_0^2 + b_\sigma/a_\sigma(1 + 1/\lambda))\mathrm{Gam}(\lambda|\alpha_\lambda, \beta_\lambda)\,\mathrm{d}\lambda$$

Note that the distributions inside the integral resemble a conjugate pair, which would lead to $p(s)$ being the t-distribution. Therefore, we speculate that the distribution $p(s)$ can be roughly approximated by a Gaussian. In fact, our simulations indicate that sampling scores from the model results in bell curve shaped histograms. The analysis above reveals a potential limitation of the proposed model – the assumption that the distribution is symmetric around the mean.

References

References

[1] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag, Berlin, Heidelberg, 2006.

[2] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, The MIT Press, 2016.

[3] D. A. Reynolds, Speaker identification and verification using Gaussian mixture speaker models, Speech Communication 17 (1-2) (1995) 91–108.

[4] G. R. Doddington, M. A. Przybocki, A. F. Martin, D. A. Reynolds, The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective, Speech Communication 31 (2-3) (2000) 225–254.

[5] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, J. Hernandez-Cordero, The 2016 NIST speaker recognition evaluation, in: Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017, 2017, pp. 1353–1357.

[6] J. C. Wu, A. F. Martin, C. S. Greenberg, R. Kacker, The impact of data dependence on speaker recognition evaluation, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (2017) 5–18.

[7] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, W. Zhou, Hidden voice commands, in: 25th USENIX Security Symposium (USENIX Security 16), Austin, TX, 2016.

[8] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, W. Xu, Dolphinattack: Inaudible voice commands, in: ACM Conference on Computer and Communications Security, ACM, 2017, pp. 103–117.

[9] N. K. Ratha, J. Connell, R. M. Bolle, Enhancing security and privacy in biometrics-based authentication systems, IBM Systems Journal 40 (3) (2001) 614–634.

[10] ISO/IEC 30107-1:2016, Information technology – Biometric presentation attack detection – Part 1: Framework, https://www.iso.org/obp/ui/#iso:std:iso-iec:30107:-1:ed-1:v1:en, [Online; accessed 06-May-2019] (2016).

[11] M. Sahidullah, H. Delgado, M. Todisco, T. Kinnunen, N. W. D. Evans, J. Yamagishi, K. A. Lee, Introduction to voice presentation attack detection and recent advances, in: S. Marcel, M. Nixon, J. Fierrez, N. Evans (Eds.), Handbook of Biometric Anti-Spoofing: Presentation Attack Detection, Springer, 2018, https://arxiv.org/abs/1901.01085.

[12] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, D. A. Reynolds, t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification, in: Proc. Odyssey 2018: The Speaker and Language Recognition Workshop, 2018, pp. 312–319.

[13] A. Nautsch, C. Rathgeb, R. Saeidi, C. Busch, Entropy analysis of i-vector feature spaces in duration-sensitive speaker recognition, in: Proc. IEEE ICASSP, Brisbane, Queensland, Australia, 2015, pp. 4674–4678.

[14] M. Lim, P. C. Yuen, Entropy measurement for biometric verification systems, IEEE Trans. Cybernetics 46 (5) (2016) 1065–1077.

[15] J. H. L. Hansen, M. K. Nandwana, N. Shokouhi, Analysis of human scream and its impact on text-independent speaker verification, The Journal of the Acoustical Society of America 141 (4) (2017) 2957–2967.

[16] R. González Hautamäki, M. Sahidullah, V. Hautamäki, T. Kinnunen, Acoustical and perceptual

study of voice disguise by age modification in speaker verification, Speech Communication 95 (2017) 1–15.

[17] P. Kenny, Bayesian speaker verification with heavy-tailed priors.

[18] C. S. Greenberg, A. F. Martin, B. N. Barr, G. R. Doddington, Report on performance results in the NIST 2010 speaker recognition evaluation, in: Proc. Interspeech, Florence, Italy, 2011, pp. 261–264.

[19] N. Brümmer, J. du Preez, Application-independent evaluation of speaker detection, Computer Speech & Language 20 (2) (2006) 230 – 275.

[20] V. Vestman, T. Kinnunen, G. Hautamäki, M. Sahidullah, Voice mimicry attacks assisted by automatic speaker verification, Computer Speech & Language 59 (2020) 36–54.

[21] Y. Lau, M. Wagner, D. Tran, Vulnerability of speaker verification to voice mimicking, in: Proc. Int. Symp on Intelligent Multimedia, Video & Speech Processing (ISIMP'2004), Hong Kong, 2004, pp. 145–148.

[22] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Trans. Audio, Speech & Language Processing 19 (4) (2011) 788–798.

[23] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-vectors: Robust DNN embeddings for speaker recognition, in: Proc. IEEE ICASSP, Calcary, Canada, 2018, pp. 5329–5333.

[24] A. Nagrani, J. S. Chung, A. Zisserman, VoxCeleb: A large-scale speaker identification dataset, Proc. Interspeech 2017 (2017) 2616–2620.

[25] J. S. Chung, A. Nagrani, A. Zisserman, VoxCeleb2: Deep speaker recognition, in: INTERSPEECH, 2018.

[26] S. J. D. Prince, J. Aghajanian, U. Mohammed, M. Sahani, Latent identity variables: Biometric matching without explicit identity estimation, in: Advances in Biometrics, International Conference, ICB 2007, Seoul, Korea, August 27-29, 2007, Proceedings, 2007, pp. 424–434.

[27] C. P. Robert, G. Casella, Monte Carlo Statistical Methods (Springer Texts in Statistics), Springer-Verlag, Berlin, Heidelberg, 2005.

[28] N. Brümmer, A. Silnova, L. Burget, T. Stafylakis, Gaussian meta-embeddings for efficient scoring of a heavy-tailed PLDA model, in: Proceedings of Odyssey 2018, Vol. 2018, International Speech Communication Association, 2018, pp. 349–356.

[29] N. Brümmer, E. de Villiers, The speaker partitioning problem, in: Odyssey, ISCA, 2010.

[30] T. Rainforth, R. Cornish, H. Yang, A. Warrington, F. Wood, On nesting Monte Carlo estimators, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, Vol. 80 of Proceedings of Machine Learning Research, PMLR, Stockholmsmässan, Stockholm Sweden, 2018, pp. 4267–4276.

[31] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, D. Rubin, Bayesian Data Analysis, 3rd Edition, Chapman and Hall/CRC, London, 2013.

[32] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, Series B 39 (1) (1977) 1–38.

[33] W. L. Buntine, Operations for learning with graphical models, J. Artif. Int. Res. 2 (1) (1994) 159–225.

[34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., The Kaldi speech recognition toolkit, Tech. rep., IEEE Signal Processing Society (2011).

[35] M. P. Alvin, A. Martin, NIST speaker recognition evaluation chronicles, in: Proc. Odyssey 2004, The Speaker and Language Recognition Workshop, 2004.

[36] J. Larson, M. Menickelly, S. M. Wild, Derivative-free optimization methods, Acta Numerica 28 (2019) 287–404.

[37] R. Ranganath, S. Gerrish, D. M. Blei, Black box variational inference, in: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014, 2014, pp. 814–822.

[38] H. Robbins, S. Monro, A Stochastic Approximation Method, The Annals of Mathematical Statistics 22 (3) (1951) 400–407.

[39] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, L. K. Saul, An introduction to variational methods for graphical models, Machine Learning 37 (2) (1999) 183–233.

[40] C. Wang, D. M. Blei, Variational inference in nonconjugate models, J. Mach. Learn. Res. 14 (1) (2013) 1005–1031.

[41] T. Minka, Estimating a gamma distribution, https://tminka.github.io/papers/minka-gamma.pdf (2002).

[42] A. Llera, C. F. Beckmann, Estimating an inverse gamma distribution, https://arxiv.org/abs/1605.01019 (2016).