Certification of Speaker Recognition Models to Additive Perturbations

Dmitrii Korzh[1,2,*], Elvir Karimov[1,2], Mikhail Pautov[1,4], Oleg Y. Rogov[1,2,3], Ivan Oseledets[1,2]
[1]AIRI; [2]Skoltech; [3]MTUCI; [4]ISP RAS; Moscow, Russia. *korzh@airi.net

TLDR

- **Problem:** one can maliciously fail voice biometrics with additive adversarial perturbations. We aim to defend voice biometrics against such attacks **provably.**
  Our solution: novel analytical randomized smoothing-based approach with a special mapping and a practical statistical numerical scheme.
  Our contributions: first voice biometrics certification, SOTA few-shot certification results.

Intro

- Deep Learning models are vulnerable to **adversarial attacks**, special perturbations that might be insignificant to the human's perception but can drastically affect the model's performance. For example, the Fast Gradient Sign Attack $x' = x - \epsilon\,\text{sign}(\nabla_x L(\theta, x, y))$ is quite effective.
- Many adversarial attacks (white-box, black-box, targeted, untargeted) and empirical defences (e.g., adversarial training) exist, resembling **cat-and-mouse games** as **empirical defences** are likely to fail under the new attack.
- Thus, it might be valuable to develop **certification** methods that provide provable guarantees on the stability of the model against some constrained set of input perturbations such as ℓ-norm bounded.
- **Certification methods have not been proposed for the speaker recognition task before.**
- Consider $f: \mathbb{R}^n \to \mathbb{R}^d$, $\|f(\cdot)\|_2 = 1$, a **speaker recognition (embedder)** model. It is trained in a way to map the speaker into a vector-embedding; for the different audio of the same speaker, corresponding embeddings should be close (and vice-versa).
- For the **speaker identification (ASI)**, the database of speakers' enrolment vectors (centroids) $S^c = \{c_j\}_{j=1}^{j=K}$, $c_k = \frac{1}{M}\sum_{x \in S_k^e} f(x)$, $\|c_k\|_2 = 1$, is given, where S^e is a set of audios used for construction of centroids.
- During **inference**, a new sample $x \in S^i$ is classified by assigning it to the speaker whose enrolment vector is the closest $i_1 = \arg\min_{k \in [1,\ldots,K]} \rho(f(x), c_k)$.
- However, **small-norm additive adversarial perturbation** ε can lead to incorrect or even malicious authentication: $\arg\min_{k\in[1,\ldots,K]} \rho(f(x), c_k) \neq \arg\min_{k\in[1,\ldots,K]} \rho(f(x+\varepsilon), c_k)$

Method

- The described model f is said to be **certified** at x against additive perturbations of bounded magnitude, if for all δ: $\|\delta\|_2 \le R$ the condition $\arg\min_{k\in[1,\ldots K]} \rho(f(x), c_k) = \arg\min_{k\in[1,\ldots K]} \rho(f(x+\delta), c_k)$ is satisfied.
- Unfortunately, this cannot be achieved directly for the f, but f can be substituted with a *smoothed model* g. This technique is called a *randomized smoothing (RS)*, and it was initially proposed for the certification of classification models and later extended for the vector functions: $g(x) = \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,\sigma^2 I)} f(x+\varepsilon)$.
- Note, that $g(x)$ is **not normalized** in contrast to $f(x)$ and centroids.
- Smoothed model have an important property of **Lipschitz continuity**: outputs' perturbation can be limited for the fixed input's perturbation level, making the model more **robust.**
- Unfortunately, $g(x)$ **cannot be evaluated analytically.** Thus, for practical application, **Monte-Carlo sampling is required,** and robustness guarantee is **probabilistic** (nonetheless, it has a high confidence).

Code and Details

Theoretical Result

- **Theorem.** Suppose that input audio x is correctly assigned to the class i_1 represented by a centroid c_{i_1}. Assume that c_{i_2} is the second closest to $g(x)$ centroid. If we introduce scalar mapping $\phi: \mathbb{R}^d \to [0,1]$ in the form $\phi = \phi(g(x), c_{i_1}, c_{i_2}) = \frac{\langle g(x), c_{i_1} - c_{i_2}\rangle}{2\|c_{i_1} - c_{i_2}\|_2} + \frac{1}{2}$. Then for all additive perturbations $\delta: \|\delta\|_2 \le R(\phi,\sigma) = \sigma\Phi^{-1}(\phi)$ **the following is satisfied**: $\arg\min_{k\in[1,\ldots K]} \|g(x) - c_k\|_2 = \arg\min_{k\in[1,\ldots K]} \|g(x+\delta) - c_k\|_2$, where $R(\phi,\sigma)$ is called **certified radius** of g at x.

Experimental Results

- We considered **ECAPA-TDNN, pyannote** voice-biometrics models and **VoxCeleb1,2** datasets for our experiments. The primary evaluation metric is **certified accuracy (CA)** — a fraction of correctly identified speakers having a certified robust radius above the current threshold: $CA(S^c, S^i, \varepsilon) = \frac{|(x,y) \in S^i : R(x) > \varepsilon \ \wedge\ i_1(x) = y|}{|S^i|}$.

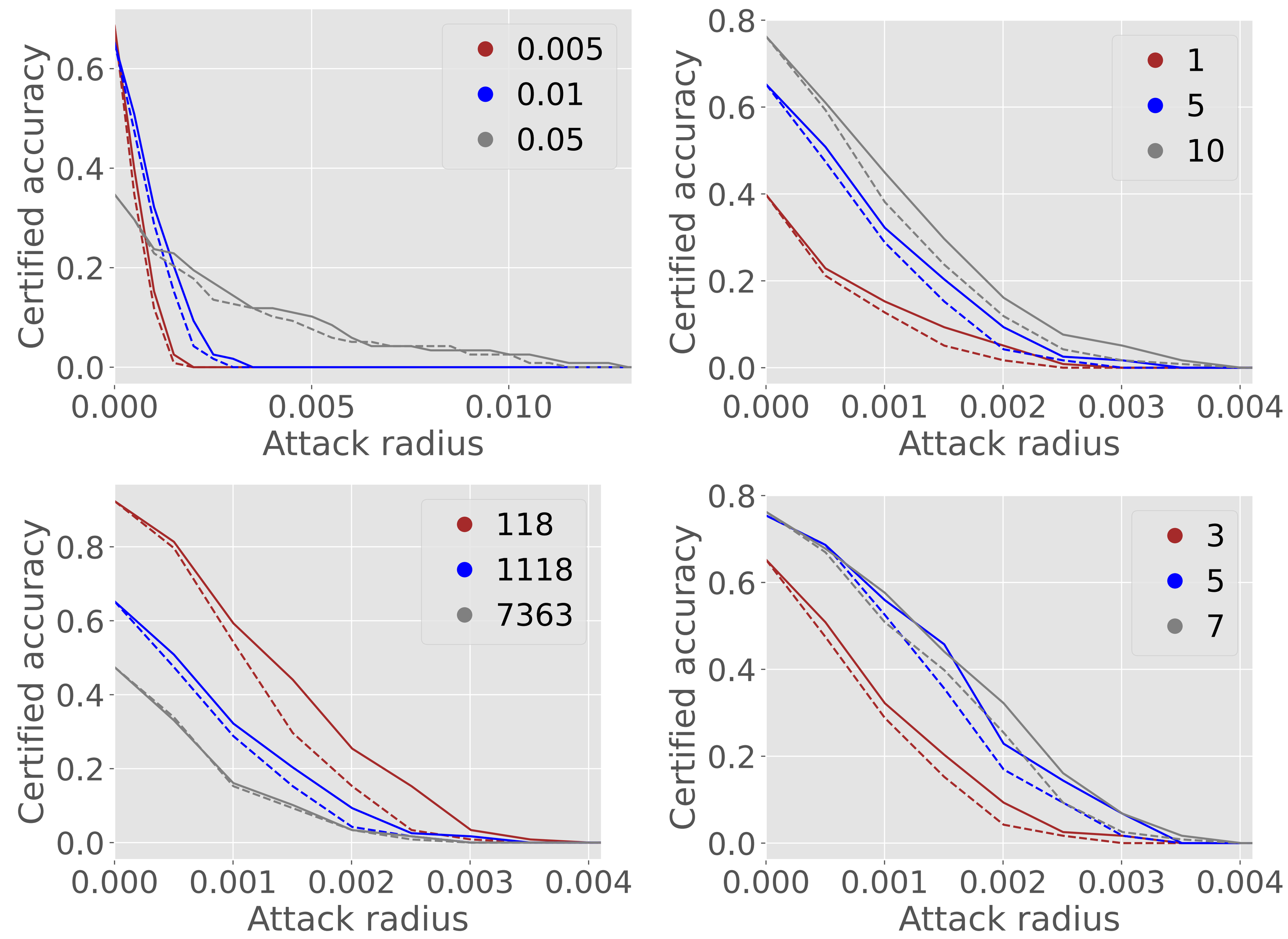

Fig 1. Pyannte model. Dependency of certified accuracy on the: **(top-left)** smoothing standard deviation σ; **(top-right)** number of enrollment audios per speaker M; **(bottom-left)** number of total enrolled speakers; **(bottom-right)** audio length in seconds in comparison with the competitors (SE). The dashed lines represent results for SE, while the solid lines correspond to our method.

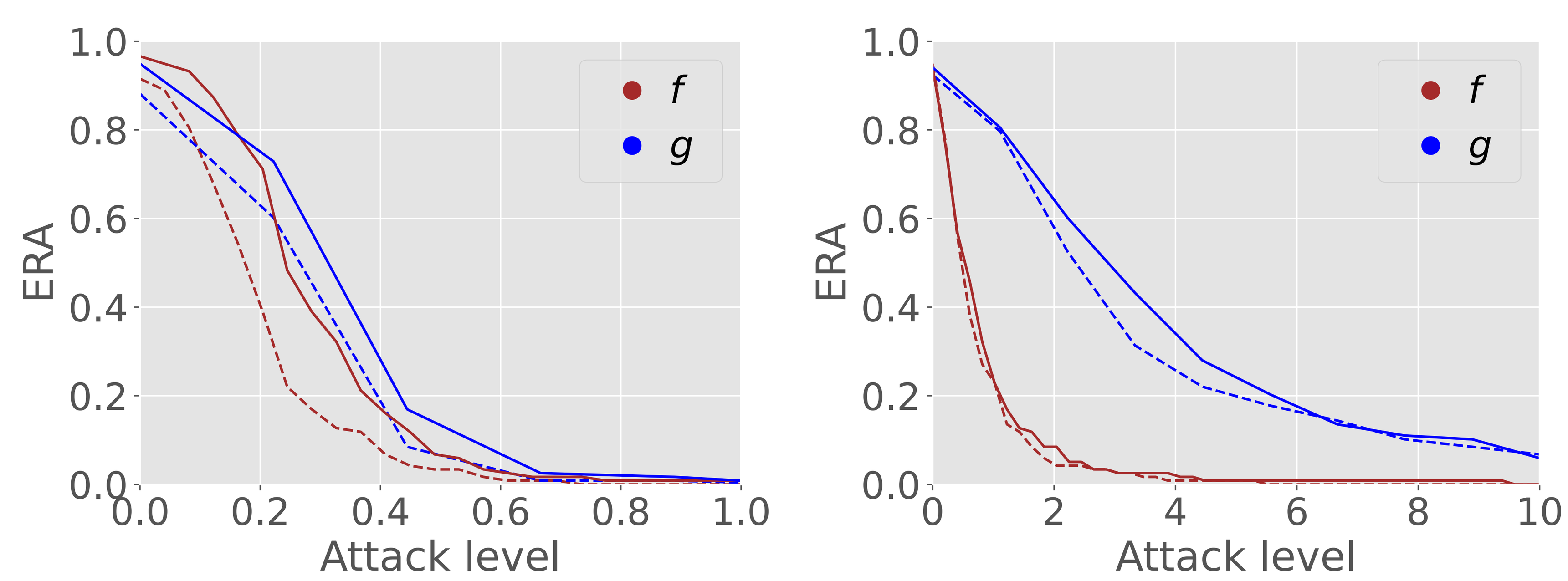

Fig 2. Pyannote model. Empirically Robust Accuracy (ERA) of base f and smoothed g models in the presence of additive perturbations: **(left)** PGD adversarial attack; **(right)** Universal Adversarial Patch. Dashed lines: audio length is 3s, solid lines - 5s.

Discussion

- Our method outperforms the competitor's method SE.
- Our method can be applied to other neural embedding tasks such as few-shot classification, requiring only the two closest centroids for certification. However, consequently, it cannot be applied directly to the speaker verification task.
- Results depend on audio length, as the longer the audio, the lower the signal-to-noise ratio of the noise of a fixed ℓ_2 norm.