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, Moscow, Russia
[2]Skolkovo Institute of Science and Technology, Moscow, Russia
[3]Moscow Technical University of Communications and Informatics, Moscow, Russia
[4]ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russia
korzh@airi.net, karimov@airi.net, pautov@airi.net, rogov@airi.net, oseledets@airi.net

Abstract

Speaker recognition technology is applied to various tasks, from personal virtual assistants to secure access systems. However, the robustness of these systems against adversarial attacks, particularly to additive perturbations, remains a significant challenge. In this paper, we pioneer applying robustness certification techniques to speaker recognition, initially developed for the image domain. Our work covers this gap by transferring and improving randomized smoothing certification techniques against norm-bounded additive perturbations for classification and few-shot learning tasks to speaker recognition. We demonstrate the effectiveness of these methods on VoxCeleb 1 and 2 datasets for several models. We expect this work to improve the robustness of voice biometrics and accelerate the research of certification methods in the audio domain.

Code — https://github.com/AIRI-Institute/asi-certification
Extended version — https://arxiv.org/abs/2404.18791

Introduction

This work addresses the issues of robustness and privacy in deep learning voice biometrics models (Snyder et al. 2018; Wan et al. 2018). Although deep learning models excel in various applications, they are unreliable and susceptible to specific perturbations. These perturbations may be imperceptible to humans but can dramatically affect the model's performance (Szegedy et al. 2014; Kaviani, Han, and Sohn 2022). Researchers have developed various methods to compute adversarial perturbations and defenses against them, recently becoming necessary to provide provable guarantees on model behavior under constrained perturbations (Li, Xie, and Li 2023; Cohen, Rosenfeld, and Kolter 2019). However, the audio domain has not received as much attention as the image domain. Given the escalating levels of speech fraud due to advancements in adversarial models and deepfake technologies (Qin et al. 2023), significant security risks could arise in biometric systems or even in creating personalized scams in social networks. Thus, this article focuses on the certification of automatic speaker recognition models. The certified speaker recognition model is the one in which prediction does not change under additive perturbations of the input audio.

Automatic speaker recognition models (Desplanques, Thienpondt, and Demuynck 2020; Bredin et al. 2020; Wang et al. 2023b) typically utilize spectrograms (such as Mel spectrograms) or raw-waveform frontends to address several vital tasks. The first task is automatic speaker identification (ASI), where the model determines the speaker's identity in an audio recording. The second task is automatic speaker verification (ASV), which involves verifying whether two audio samples are from the same speaker. The third task is speaker diarization, where the model segments audio into parts corresponding to different speakers.

Voice biometric models convert speech into vector representations, ensuring that utterances from the same speaker generate closely aligned vectors while those from different speakers are widely separated. These properties should hold even for speakers not encountered during training. Several training strategies exist for the encoder that maps audio x to these embeddings. One approach uses metric learning with triplet (Hermans, Beyer, and Leibe 2017) or contrastive loss (Wang and Liu 2021). Another strategy involves training an embedder combined with a classifier on a fixed set of speakers, with variations of cross-entropy loss that was initially developed for face biometrics (Meng et al. 2021) to enhance the expressiveness and separation of embeddings, even for unseen speakers. During inference, cosine similarity, cosine distance, or other distance metrics are used to match the embedding of the inference audio to the closest reference speaker's embedding (enrollment vector).

Our work explores the certified robustness of speaker recognition models against any additive perturbation constrained by the l_2 norm value. Such perturbations can be created via various adversarial attacks, whether targeted or untargeted, white-box or black-box scenarios, in which the attacker may know the model's architecture, parameters, and gradients or may only have input and output access.

Our contributions can be summarized as follows:

- We introduce a novel randomized smoothing-based approach to certify few-shot embedding models against additive, norm-bounded perturbations. Our approach provides state-of-the-art certification results in a few-shot setting.
- We derive robustness certificates and demonstrate their ad-

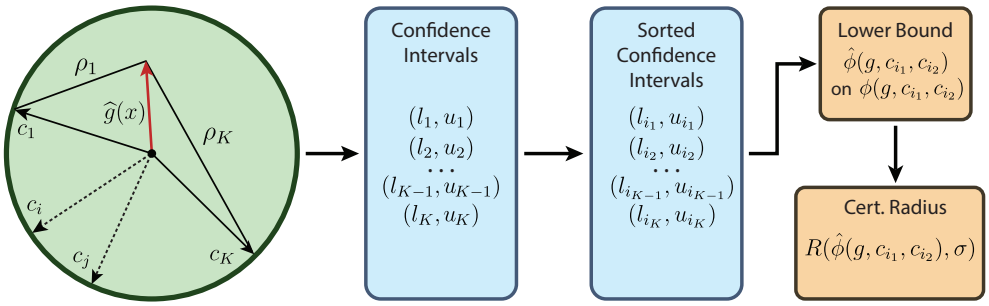


Figure 1: The scheme illustrating the proposed algorithm. The algorithm requires an audio sample x, base model f, and the set of centroids $S^c = \{c_1, \ldots, c_K\}$. In the Figure, $\hat{g}(x)$ corresponds to the estimation of the smoothed embedding $g(x)$ from Eq. (8) computed in the form from Eq. (11). When executed, Algorithm 1 computes the confidence interval (l_i, u_i) for the distance between $\hat{g}(x)$ and corresponding centroid c_i for all $i \in [1, \ldots, K]$. Then, given sorted confidence intervals $\{(l_{i_1}, u_{i_1}), \ldots, (l_{i_K}, u_{i_K})\}$, two closest centroids, c_{i_1} and c_{i_2}, are determined. The last step of the algorithm is the computation of the lower bound $R(\hat{\phi}(\hat{g}, c_{i_1}, c_{i_2}))$ on the certified radius $R(\phi(g, c_{i_1}, c_{i_2}))$ from the Theorem 1.

vantages over those obtained using existing competitors' methods. Our theoretical claims are supported by experimental results on the VoxCeleb datasets using several well-known speaker recognition models.

- To the best of our knowledge, there are no previous works that present the provable robustness of speaker recognition models. We highlight this issue and provide starting baselines that others can improve in future research.

Related Work

Speaker Recognition

Recently, speaker recognition (Snyder et al. 2018; Wan et al. 2018; Desplanques, Thienpondt, and Demuynck 2020; Wang et al. 2023a,b) has made significant progress. The x-vector system, based on Time Delay Neural Network (TDNN) technology, has been particularly influential and further developed in many other models. This system uses one-dimensional convolution to pick up important time-related features in speech. For example, ECAPA-TDNN (Desplanques, Thienpondt, and Demuynck 2020) uses techniques that allow the model to consider a wider range of time-related information, combining features recursively from several previous states for the next hidden state. Later, a densely connected TDNN (D-TDNN) (Yu and Li 2020) was presented, which reduced the number of parameters needed. Additionally, the Context-Aware Masking (CAM) module, a type of pooling, was combined with D-TDNN, and the model CAM++ improves the performance regarding verification metrics (such as Equal Error Rate and Detection Cost Function) and the inference time.

Adversarial Attacks

It has long been known (Szegedy et al. 2014; Goodfellow, Shlens, and Szegedy 2015) that deep learning models are vulnerable to small additive perturbations of input. In recent years, many approaches have been proposed to generate adversarial examples, for example, (Athalye, Carlini, and Wagner 2018; Khrulkov and Oseledets 2018; Su, Vargas, and Sakurai 2019; Yuan et al. 2021; Wang et al. 2023c). These methods expose different conceptual vulnerabilities of models: some generate attacks using information about the model's gradient, while others deploy separate networks to produce malicious input. Moreover, adversarial examples can be transferred across models (Inkawhich et al. 2019), which limits the application of neural networks in various practical scenarios. This vulnerability poses significant risks in contexts such as biomedical image segmentation (Apostolidis and Papakostas 2021), industrial face recognition (Komkov and Petiushko 2021) and detection (Kaziakhmedov et al. 2019), self-driving car systems (Deng et al. 2020), and speaker recognition systems (Zhang et al. 2023; Lan et al. 2022; Li et al. 2020). Additionally, speaker anonymization systems aim to conceal identity features while preserving other information (text, emotions) from the speech, and often based on the generation of additive perturbations (Deng et al. 2023; Liu et al. 2024).

Empirical and Certified Defenses

Numerous defensive approaches have recently been proposed (Li, Xie, and Li 2023; Fan et al. 2023) to mitigate the effects of the attacks. Among these, adversarial training (Goodfellow, Shlens, and Szegedy 2015; Andriushchenko and Flammarion 2020) is arguably the best technique to enhance the robustness of models in practice. The method is straightforward – during training, each batch of data is augmented with adversarial examples generated by a specific method. Consequently, the model becomes more resistant to the type of attack used during the training process. However, the model may easily become overfitted to the provided attacks and unable to be robust against new types of adversarial perturbations. Additionally, adversarial training is time-consuming and often leads to notable performance degradation. Despite this, several prominent fast adversarial training approaches exist. Data augmentation with ordinary transforms and noises (e.g., Gaussian) and regularization techniques (e.g., consistency loss (Jeong and Shin 2020)) are the most straightforward, cheapest, and prominent approaches to increase empirical ro-

bustness. Additionally, (Castan et al. 2017; Zhou et al. 2023; Wu et al. 2021) are improved empirical guarantees in speaker recognition using unlabeled data, adversarial training, and self-supervised methods.

Another research direction is the development of methods that provide provable certificates on the model's prediction under certain transformations. Mainly, approaches are based on Satisfiability Modulo Theory (Pulina and Tacchella 2010) and Mixed Integer Linear Programming (Cheng, Nührenberg, and Ruess 2017) solvers, on the interval (Gowal et al. 2019) and polyhedra (Lyu et al. 2020) relaxation, on analysis of Lipschitz continuity (Salman et al. 2019) and the curvature of the decision boundary of the network (Singla and Feizi 2020).

Nowadays, randomized smoothing (Cohen, Rosenfeld, and Kolter 2019; Salman et al. 2019) forms the basis for many certification approaches, offering defenses against both norm-bounded (Yang et al. 2020) and semantic perturbations (Li et al. 2021; Muravev and Petiushko 2022; Hao et al. 2022). This method is simple, effective, and scalable to large models and datasets. Notably, it can also be theoretically applied to certify automatic speech recognition systems (Olivier and Raj 2021).

Methodology

In this section, we define the problem statement, provide an overview of the techniques used, and describe the proposed method for certifying embedding models against norm-bounded additive perturbations.

Speaker Recognition as a Few-Shot Problem

Few-shot learning is a machine learning paradigm where models are trained to generalize effectively from only a few examples of each class, addressing the challenge of limited data availability (Koch et al. 2015; Snell, Swersky, and Zemel 2017), that is highly relevant to biometrics systems. Consider $f : \mathbb{R}^n \to \mathbb{R}^d$ as the base model that maps input audios to normalized embeddings, where $\|f(\cdot)\|_2 = 1$, n is an input dimension, d is an embedding dimension. After training the embedding model, we need to enroll new speakers we want to authorize later in our biometrics system.

For every enrolled speaker, the enrollment vector or centroid is established as the mean or weighted sum of embeddings derived from collected audio samples of the speaker. These centroids create the basis for calculating the similarity with the embeddings of new audio samples during inference authorization. The enrollment dataset, denoted as $S^e = \{(x_1, y_1), \ldots, (x_l, y_l)\}$, consists of audio samples $x_i \in \mathbb{R}^n$ assigned to corresponding speakers $y_i \in [1, \ldots, K]$. Depending on the application, this dataset may consist of speakers not encountered during training or a mix of seen and unseen speakers. For a given class k, the subset $S^e_k = \{(x_i, y_i) \in S^e : y_i = k\}$ comprises the audios belonging to the speaker k. Although in practice, the number of available audios $M(k)$ in every subset S^e_k can vary from speaker to speaker, in the few-shot setting, the number $M(k)$ is fixed to the pre-defined number M of audios used to construct the speaker's enrollment vector $\forall k \mapsto |S^e_k| = M$ for the fair comparison. The normalized speaker enrollment embedding (speaker centroid, prototype) is then can be formalized as follows:

$$c_k = \frac{1}{M} \sum_{x \in S^e_k} f(x), \;\; \|c_k\|_2 = 1, \tag{1}$$

and a database $S^c = \{c_j\}_{j=1}^{j=K}$ of centroid vectors is constructed. During inference, a new sample $x \in S^i$ is classified by assigning it to the speaker whose enrollment vector from S^c is the closest in terms of some distance function ρ:

$$i_1 = \underset{k \in [1,\ldots,K]}{\operatorname{argmin}} \; \rho(f(x), c_k). \tag{2}$$

Although few-shot usually implies $M \in [1, 2, 3]$ only, we consider $M \in [1, 5, 10]$ following common biometrics practice. We equate the speaker recognition (ASI) and few-shot models to emphasize that our method is also applicable to other few-shot scenarios.

Problem Statement and Certification for Vector Functions

Certification guarantees against additive perturbations of a bounded magnitude can be formulated as follows. Suppose that f is the base vector (embedding) model, c_k is defined as in Eq. (1), and $R > 0$ is the norm threshold. Then, the model f is said to be certified at x, if for all $\|\delta\|_2 \leq R$,

$$\underset{k \in [1,\ldots K]}{\operatorname{argmin}} \; \rho(f(x), c_k) = \underset{k \in [1,\ldots K]}{\operatorname{argmin}} \; \rho(f(x + \delta), c_k). \tag{3}$$

Unfortunately, this cannot be achieved directly for the f, but f can be substituted with *smoothed model* g. This technique is called a *randomized smoothing (RS)*, and it was initially proposed for the classification (Lecuyer et al. 2019; Cohen, Rosenfeld, and Kolter 2019) and g has an important property of Lipschitz continuity (Salman et al. 2019): outputs' perturbation can be limited for the fixed input's perturbation level. Given the classifier model $f_{\text{clf}} : \mathbb{R}^n \to [0, 1]^K$ and the smoothing distribution $\mathcal{N}(0, \sigma^2 I)$ the smoothed model takes the form

$$g_{\text{clf}}(x) = \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 I)} f_{\text{clf}}(x + \varepsilon), \tag{4}$$

here $g_{\text{clf}}(x)$ is the vector of class probabilities with K components. As it is shown in (Cohen, Rosenfeld, and Kolter 2019), when the model from Eq. (4) is confident in predicting the correct class i_1 for the input x,

$$g_{\text{clf}}(x)_{i_1} = p_{i_1} \geq p_{i_2} = \max_{i \neq i_1} g_{\text{clf}}(x)_i \tag{5}$$

then it is robust in l_2-ball around x of radius

$$R = \frac{\sigma}{2} \left(\Phi^{-1}(p_{i_1}) - \Phi^{-1}(p_{i_2}) \right), \tag{6}$$

$$\forall \delta : \|\delta\|_2 < R \mapsto \operatorname{argmax} g_{\text{clf}}(x) = \operatorname{argmax} g_{\text{clf}}(x + \delta), \tag{7}$$

where $\Phi^{-1}(\cdot)$ is the inverse of the standard Gaussian cumulative density function.

For the vector functions, let us consider the base model $f : \mathbb{R}^n \to \mathbb{R}^d$ that maps input to normalized embeddings, the associated smoothed model $g : \mathbb{R}^n \to \mathbb{R}^d$ is defined as

$$g(x) = \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 I)} f(x + \varepsilon). \tag{8}$$

Here $g(x)$ is $d-$dimensional smoothed embedding. Note that f and centroids c_k are normalized while g is not. Suppose that input audio x is correctly assigned to class i_1 represented by 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

$$\phi = \phi(g(x), c_{i_1}, c_{i_2}) = \frac{\langle g(x), c_{i_1} - c_{i_2} \rangle}{2\|c_{i_1} - c_{i_2}\|_2} + \frac{1}{2}, \quad (9)$$

then the following robustness guarantee holds:

Theorem 1 (Main result)**.** *For all additive perturbations* $\delta : \|\delta\|_2 \leq R(\phi, \sigma) = \sigma\Phi^{-1}(\phi)$

$$\underset{k\in[1,\dots K]}{\operatorname{argmin}} \|g(x) - c_k\|_2 = \underset{k\in[1,\dots K]}{\operatorname{argmin}} \|g(x + \delta) - c_k\|_2, \quad (10)$$

where $R(\phi, \sigma)$ *is called certified radius of* g *at* x*.*

Remark 1. The detailed proof is provided in the Appendix of the full manuscript version.

Remark 2. The method is generalizable to open setups and other neural embedding tasks, requiring only the two closest centroids for certification. Thus, it cannot be applied to ASV certification. Note that cosine distance is as suitable as l_2 norm.

Implementation Details

In this section, we describe the numerical implementation of the proposed method.

Sample Mean Instead of Expectation

Notably, the prediction of the smoothed model from Eq. (8) is an expected value of the random variable that is the function of the base classifier. Hence, it is impossible to evaluate it exactly in the case of nontrivial f. Consequently, evaluating the mapping ϕ from Eq. (9) is impossible. A conventional way to deal with this problem is to replace the smoothed model with its unbiased estimation – sample mean computed over N samples, namely

$$\hat{g}(x) = \frac{1}{N} \sum_{i=1}^{N} f(x + \varepsilon_i), \quad (11)$$

where ε_i are the independent identically distributed normal random variable. However, it is also impossible to exactly determine which two centroids c_{i_1} and c_{i_2} are the closest ones to the true value of the smoothed classifier from Eq. (8). We solve the issues mentioned above in the following manner:

1. Firstly, we compute interval estimations of the distances between $g(x)$ and all the centroids using Hoeffding inequality (Hoeffding 1994). It is done to determine the two closest centroids with sufficient confidence.
2. Secondly, given the two closest centroids, we compute the lower confidence bound $\hat{\phi}$ of ϕ from Eq. (9).
3. Finally, when $\hat{\phi}$ is computed, the value $R(\hat{\phi}, \sigma)$ from Theorem 1 is treated as the certified radius of g at x.

Algorithm 1: Computation of the certified radius.

Input: f, x
Parameter: $N, N_{\text{max}}, \sigma, \alpha$
Output: R

isFinished $\leftarrow$ False
$N_0 \leftarrow N$
while not isFinished or $N \leq N_{\text{max}}$ **do**
  $\varepsilon_1, \dots, \varepsilon_N \sim \mathcal{N}(0, \sigma^2 I)$
  $\varepsilon_{N+1}, \dots, \varepsilon_{2N} \sim \mathcal{N}(0, \sigma^2 I)$
  $\hat{g}_1(x) = \frac{1}{N} \sum_{j=1}^{N} f(x + \varepsilon_j)$
  $\hat{g}_2(x) = \frac{1}{N} \sum_{j=1}^{N} f(x + \varepsilon_{N+j})$
  for $i \in \{1, \dots, K\}$ **do**
    $v_i^1 = \hat{g}_1(x) - c_i$
    $v_i^2 = \hat{g}_2(x) - c_i$
    $(l_i, u_i) \leftarrow$ HOEFFDINGCI(v_i^1, v_i^2, α) ▷ Computation of two-sided CI using Hoeffding inequality, namely $(l_i, u_i) : \mathbb{P}\left(\|g(x) - c_i\|_2 \in (l_i, u_i)\right) \geq 1 - \alpha$
  $i_1 \leftarrow \text{argmin}\{l_1, \dots, l_K\}$
  $i_2 \leftarrow \text{argmin}\{l_1, \dots, l_K \setminus l_{i_1}\}$
  $i_q \leftarrow \text{argmin}\{l_1, \dots, l_K \setminus \{l_{i_1}, l_{i_2}\}\}$
  if $u_{i_1} < l_{i_2} \wedge u_{i_2} < l_{i_q}$ **then**
    isFinished $\leftarrow$ True
    $\hat{g}(x) = \frac{\hat{g}_1(x) + \hat{g}_2(x)}{2}$
    $\tilde{\phi} \leftarrow \frac{\langle \hat{g}(x), c_{i_1} - c_{i_2} \rangle}{2\|c_{i_1} - c_{i_2}\|_2} + \frac{1}{2}$
    $\hat{\phi} \leftarrow$ HOEFFDINGLOWERBOUND$(\tilde{\phi}, \alpha)$
    $R \leftarrow \sigma\Phi^{-1}(\hat{\phi})$
    return R
  else
    if $2N > N_{\text{max}}$ **then**
      return Abstain
    else
      $N \leftarrow N + N_0$

Hoeffding Confidence Interval and Error Probability

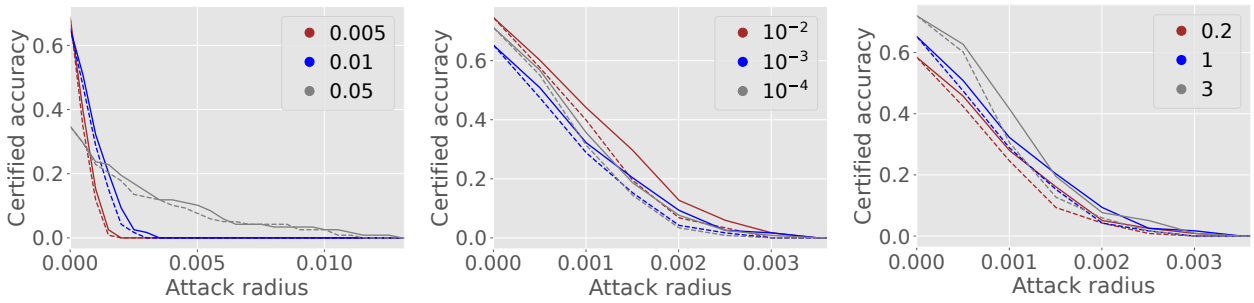
Hoeffding inequality (Hoeffding 1994) bounds the probability of a large deviation of a sample mean from the population mean, namely

$$\mathbb{P}(|\overline{X} - \mathbb{E}(X)| \geq t) \leq 2 \exp\left(-\frac{2t^2 N^2}{\sum_{i=1}^{N} (b_i - a_i)^2}\right), \quad (12)$$

where $\overline{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$, and X_i are i.i.d. random variables such that $\mathbb{P}(X_i \in (a_i, b_i) = 1)$.

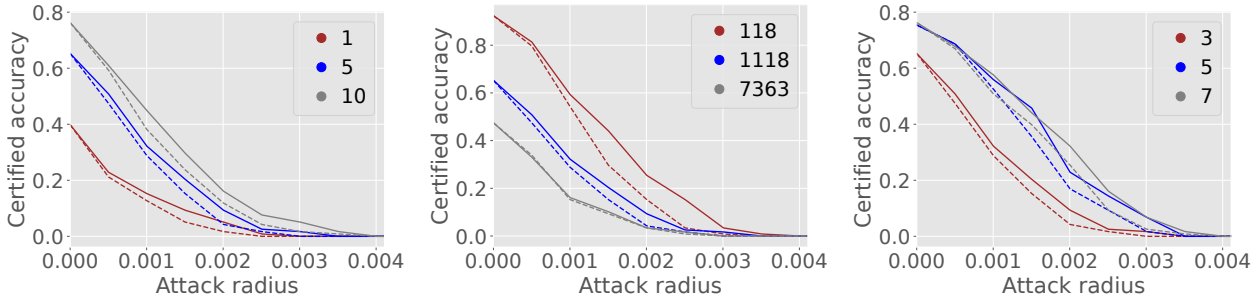
Distances to the Centroids. An estimation of distance between the smoothed embedding $g(x)$ from Eq. (8) and the speaker centroid c_i from Eq. (1) may be derived from an estimation of the dot product $\langle \hat{g}_1(x) - c_i, \hat{g}_2(x) - c_i \rangle$, where

$$\begin{aligned} \hat{g}_1(x) &= \frac{1}{N} \sum_{i=1}^{N} f(x + \varepsilon_i), \\ \hat{g}_2(x) &= \frac{1}{N} \sum_{j=1}^{N} f(x + \varepsilon_j) \end{aligned} \quad (13)$$



(a) Dependency on σ. (b) Dependency on α. (c) Dependency on $N_{\max}$ ($\times 10^5$).

Figure 2: Pyannote model. Few-shot setting. Dependency of certified accuracy on the variance σ of the additive noise, confidence level α, and the maximum number of noise samples $N_{\max}$. The dashed lines represent results for SE, while the solid lines correspond to our method.



(a) Dependency on M. (b) Dependency on number of speakers. (c) Dependency on the audio length (s).

Figure 3: Pyannote model. Few-shot setting. Dependency of certified accuracy on number M of audios of a single speaker, number of enrolled speakers K, and the audio length in seconds. The dashed lines represent results for SE, while the solid lines correspond to our method.

are two independent unbiased estimates of $g(x)$. Once computed, confidence interval (l_i^2, u_i^2) for the expression $\langle \hat{g}_1(x) - c_i, \hat{g}_2(x) - c_i \rangle$ implies confidence interval (l_i, u_i) of interest. The work of (Pautov et al. 2022) provides a detailed derivation of the confidence interval.

Estimation of $\hat{\phi}$. Hoeffding inequality is also used to compute confidence intervals for the value ϕ from Theorem 1. Namely, given

$$\tilde{\phi} - \frac{1}{2} = \frac{\left\langle \frac{\hat{g}_1(x)+\hat{g}_2(x)}{2}, c_{i_1} - c_{i_2} \right\rangle}{2\|c_{i_1} - c_{i_2}\|_2}, \tag{14}$$

as an estimation of $\phi - \frac{1}{2}$ over $2N$ samples ξ_j in the form

$$\xi_j = \frac{\langle f(x+\varepsilon_j), c_{i_1} - c_{i_2} \rangle}{2\|c_{i_1} - c_{i_2}\|_2}, \tag{15}$$

such that $\xi_j \in [-\frac{1}{2}, \frac{1}{2}]$, we compute lower bound $\hat{\phi} - \frac{1}{2}$ of $\phi - \frac{1}{2}$ in the form

$$\hat{\phi} - \frac{1}{2} = \tilde{\phi} - \frac{1}{2} - \sqrt{\frac{\ln \frac{2}{\alpha}}{4N}}. \tag{16}$$

Note that α in Eq. (16) is the upper bound for the error probability. In other words,

$$\mathbb{P}(\phi < \hat{\phi}) < \alpha. \tag{17}$$

All the procedures are combined in the numerical pipeline presented in the Algorithm 1 and schematically in Fig. 1.

Error Probability of Algorithm 1. Since the procedure in Algorithm 1 is not deterministic (as it depends on the computation of confidence intervals), it is important to estimate its failure probability. First, estimating the two closest centroids is statistically sound only if all the distances between smoothed embedding and the centroids are within corresponding confidence intervals. In contrast, if at least one of the distances

$$\|g(x) - c_1\|_2, \ldots, \|g(x) - c_K\|_2 \tag{18}$$

is not within the corresponding interval, it is impossible to guarantee that the two closest centroids are correctly determined. Thus, all the respective Hoeffding inequalities have to hold. It happens with the probability $p_1 = (1-\alpha)^K$, where K is the number of classes. Secondly, note that the lower confidence bound for ϕ from Theorem 1 is correct with probability $p_2 = (1-\alpha)$.

Thereby, the probability of the correct output of Algorithm 1 is $p_1 p_2 = (1-\alpha)^{K+1}$ what leads to the error probability $q = 1 - (1-\alpha)^{K+1}$.

Experiments

Datasets

For our experiments, we used the VoxCeleb1 (Nagrani, Chung, and Zisserman 2017) and VoxCeleb2 (Chung, Nagrani, and Zisserman 2018) datasets, which are standard for speaker recognition and verification tasks. VoxCeleb1 comprises 1211 development speakers and 40 test speakers, with over 150000 utterances spanning 350 hours. VoxCeleb2 includes 5994 development speakers and 118 test speakers, totaling about 2400 hours and 1.1 million utterances. These multilingual datasets feature speakers from over 140 nationalities, covering various accents and ages. We evaluated our method by varying the number of enrolled speakers from 118 (VoxCeleb2 test set) to nearly all available speakers (7323), excluding the VoxCeleb1 test set.

Evaluation Protocol

We evaluate the methods in several settings. Experiments were conducted using various backbone embedding models: ECAPA-TDNN (Desplanques, Thienpondt, and Demuynck 2020) from the Speechbrain framework (Ravanelli et al. 2021) that utilizes Mel-Spectrogram for the frontend; the Pyannote framework (Bredin et al. 2020), which focuses on speaker diarization and utilizes the raw-waveform frontend SincNet. These models transform speech into vector representations of dimensions $d = 192$ and 512 correspondingly. For the ECAPA-TDNN-based f, plain accuracy is $\mathrm{Acc} = 95.0\%$, equal-error-rate $\mathrm{EER}(f) = 0.34\%$, $\mathrm{EER}(g) = 0.89\%$. For the Pyannote-based f, $\mathrm{Acc} = 88.3\%$, $\mathrm{EER}(f) = 1.17\%$, $\mathrm{EER}(g) = 1.40\%$. EER is a decision threshold regarding the ASV or classification task for which the model's false acceptance and false rejection rates are equal. We conducted experiments in an ASI setting.

For the certification procedure in Algorithm 1, the default parameters are the following: standard deviation of additive noise used for smoothing $\sigma = 10^{-2}$, the maximum number of samples to construct $\hat{g}$ is set to be $N_{\max} = 10^5$, the confidence level $\alpha = 10^{-3}$, number of enrolled speakers is $K = 1118$, number of random audios used to create the speaker enrollment vector $M = 5$, and length of given audios is set to be 3s with sampling rate 16 kHz, number of speakers in the test set S^i is 118 (VoxCeleb2 Test).

For the evaluation, we considered K enrolled speakers and, for each of them, created $c_k \in S^c$ of M randomly sampled speaker's enrollment audios, which are presented in S^e. We tested our models, providing inference audios $x \in S^i$, $S^e \cap S^i = \emptyset$, where number of unique test speakers in S^i is fixed and equal to 118 (VoxCele2 test). We report certified accuracy (CA) for each method on the S^c centroids and S^i test audios. Certified accuracy represents the proportion of correctly matched samples from S^i to the corresponding centroids in S^c for which the smoothed model has a certified radius exceeding the given attack magnitude. Specifically, given the recognition rule

$$i_1(x) = \underset{k \in \{1,\dots,K\}}{\operatorname{argmin}} \rho\left(g(x), c_k\right), \tag{19}$$

and the norm of perturbation ε, the certified accuracy is computed as follows:

$$CA(S^c, S^i, \varepsilon) = \frac{|(x,y) \in S^i : R(x) > \varepsilon \ \wedge \ i_1(x) = y|}{|S^i|}, \tag{20}$$

where $R(x)$ is the certified radius from Theorem 1.

We compared our approach to the work of (Pautov et al. 2022), where authors propose the method called *Smoothed Embeddings* (*SE*) to certify prototypical networks. The certified radius $R^{SE}(x)$ produced by *SE* has the form

$$R^{SE}(x) = \sqrt{\frac{\pi\sigma^2}{2}} \frac{\|c_{i_2} - g(x)\|_2^2 - \|c_{i_1} - g(x)\|_2^2}{2\|c_{i_1} - c_{i_2}\|_2^2}. \tag{21}$$

in our notation. In contrast to our work, they perform a geometrical analysis of Lipschitz properties of the smoothed model, whereas we study the properties of the scalar mapping from the embedding space.

We also provided results (see Appendix of the full version of the manuscript) based on vanilla RS (Cohen, Rosenfeld, and Kolter 2019; Salman et al. 2019) in the classification setting (6) of a fixed number K of speakers in S^c. Default parameters were the same as in our approach. Since an exact estimation of $g_{\text{clf}}(x)$ (4) is impossible, a similar sample-mean is utilized with the Clopper-Pearson (Clopper and Pearson 1934) test for estimation of $\hat{p}_{i_1}$ which is the lower confidence bound of p_{i_1}. In a nutshell, this is a Binomial proportion test confidence interval of top class v.s. the rest. This requires the correct class to be predicted in more than half of samples $\hat{p}_{i_1} > \frac{1}{2}$ for the certification.

The computational time is approximately 30 seconds for the Pyannote model and 120 seconds for ECAPA-TDNN.

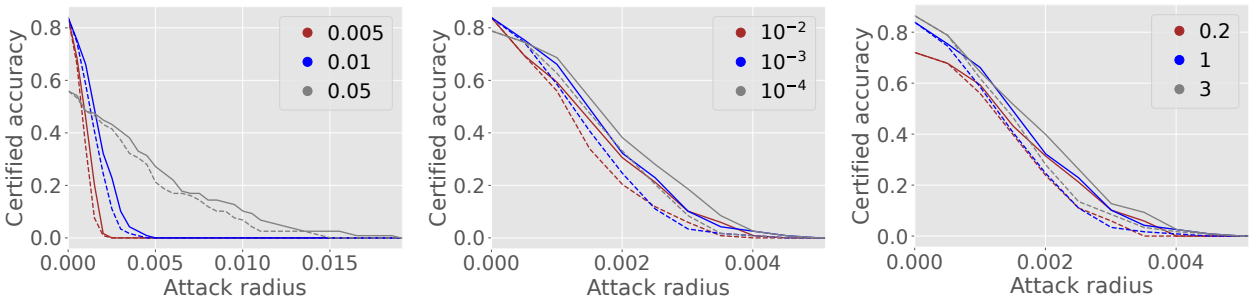
Results and Discussion

In Figures 2, 3 and 4, 5 we present results that illustrate the effects of varying a single parameter while keeping all other at their default values for the SE and our approaches for two backbone models. Several observations can be obtained from these results:

- σ significantly impacts the certification system (ours, SE, and RS). Higher values lead to a more robust system, which comes at the expense of reduced accuracy (robustness-accuracy trade-off);
- α does not affect the certification significantly;
- There are threshold values for the number of speaker enrollment audios M and audio length beyond which the results remain nearly unchanged;
- Evidently, an increase of $N_{\max}$ parameter enhances the certification process, while classification difficulty rises as the number of enrolled speakers K increases.

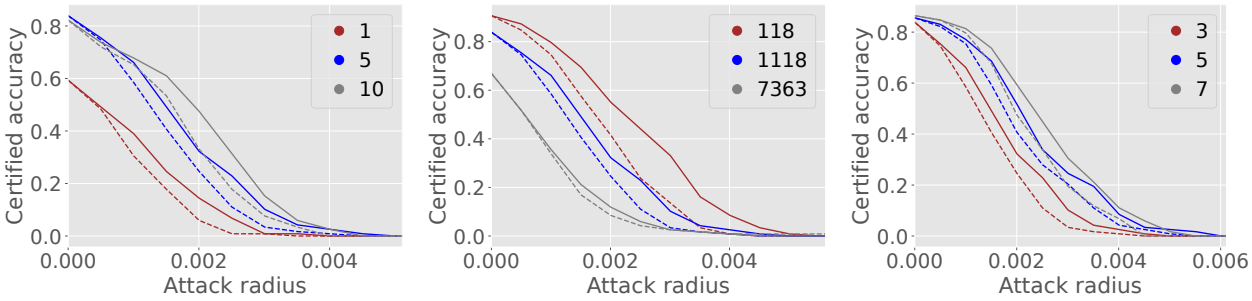
Additionally, our method demonstrates a marginal improvement across all scenarios compared to the SE approach. Figures 2 - 5 illustrate that our method achieves enhanced certified accuracy for the same attack levels.

In Figure 6, we demonstrate empirical robust accuracy (ERA) – the fraction of correctly recognized perturbed audios $x + \delta$ for all sampled perturbations $\delta \leq l$, where l is a current attack level. g was estimated as in Eq. (11) without



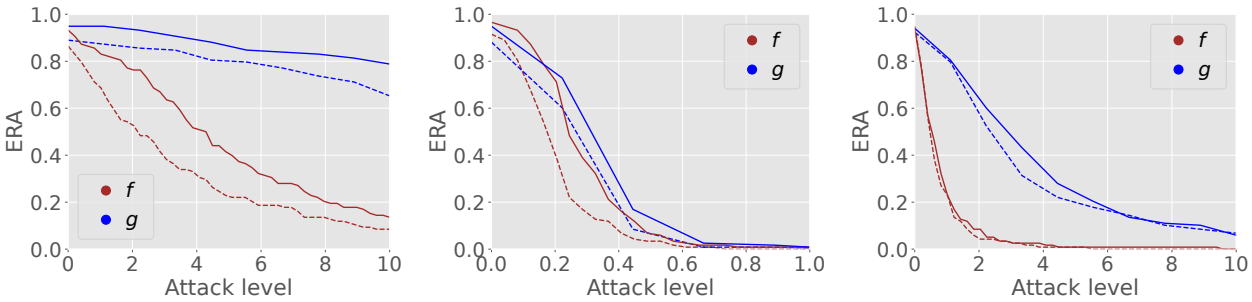
(a) Dependency on σ. (b) Dependency on α. (c) Dependency on N_{max} ($\times 10^5$).

Figure 4: ECAPA-TDNN model. Few-shot setting. Dependency of certified accuracy on the variance σ of the additive noise, confidence level α, and the maximum number of noise samples N_{max}. The dashed lines represent results for SE, while the solid lines correspond to our method.



(a) Dependency on M. (b) Dependency on number of speakers. (c) Dependency on the audio length (s).

Figure 5: ECAPA-TDNN model. Few-shot setting. Dependency of certified accuracy on number M of audios of a single speaker, number of enrolled speakers K, and the audio length in seconds. The dashed lines represent results for SE, while the solid lines correspond to our method.



(a) Gaussian noise. (b) PGD. (c) UAP.

Figure 6: Empirically Robust Accuracy (ERA) of functions f and g in the Pyannote model under different perturbations: Gaussian noise, PGD adversarial attack, speaker anonymization, and Universal Adversarial Patch (UAP) (Liu et al. 2024). The dashed lines represent results for audio length = 3 seconds, while the solid lines correspond to audio length = 5 seconds. This evaluation follows a few-shot learning setting.

certification criteria. Projected Gradient Descent (Madry et al. 2018) is selected as it is considered a standard adversarial attack to evaluate models' robustness. One can notice that the empirical robustness of g and f is significantly better than the certification results of g. Nonetheless, presented attacks do not necessarily convey the worst certification result, as stronger attacks exist, and the worst-case ERA might be closer to CA.

The certification condition in Theorem 1 does not depend on audio length explicitly. For a given sample $x \in \mathbb{R}^n$, it

yields the certified radius $R(x)$ – a lower bound on the l_2-norm of a perturbation δ that can change a smoothed model's prediction. However, the relative distortion (e.g., signal-to-noise ratio) differs for various n. One can notice from the (Fig. 3c and Fig. 5c) that the longer the audio sample is, the smaller the relative distortion is and consequently, certification results are better. Additionally, achieving audio-length independent certification against $l_\infty-$norm bounded perturbations seems unsolvable (Hayes 2020). The Theorem 1 is still valid if l_2-norm as the distance function is replaced by the negative cosine distance.

Our method is evaluated for the speaker identification task only. The method can be transferred to the speaker diarization task but cannot be applied directly in an ASV scenario.

It is feasible to extend our certification procedure to multiplicative and semantic transformations (Muravev and Petiushko 2022; Li et al. 2021) by applying different mappings and smoothing distributions. Nonetheless, the method certifies the model only against additive perturbations for the fixed voiceprint x, but these guarantees do not apply a priori to the new voiceprint x_1 even if it is a genuine speech of the same speaker: $\forall \delta : \|\delta\|_2 \leq R(x) \mapsto \operatorname{argmin}_k \rho(g(x+\delta), c_k) = i_1$, where i_1 is a correct class, but $\exists \delta_1 : \|\delta_1\|_2 \leq R(x)$, but $\operatorname{argmin}_k \rho(g(x_1+\delta_1), c_k) \neq i_1$. Additionally, current methods cannot help to certify SR models against rapidly evolving deepfakes (Yamagishi et al. 2021; Wang et al. 2024).

Although RS over class probabilities provides better certification radii (see Appendix of the full version of the manuscript) compared to our approach, our method does not imply knowledge of the class probabilities that may be more suitable for the metric learning tasks.

Conclusion

In this work, we presented a new approach to certify speaker identification models that map input audios to normalized embeddings against norm-bounded additive perturbations. We introduced scalar mapping from the embedding space and derived theoretical robustness guarantees based on its Lipschitz properties. We experimentally evaluated our approach against the concurrent method and achieved state-of-the-art certification results in a few-shot setting. In addition, our method can be applied to the certification of other metric learning tasks, such as face biometrics.

In summary, we expect this work to highlight the issue of certified robustness in biometrics systems, particularly in speaker identification, and improve AI safety. Future developments in this topic might be devoted to improving empirical and certified guarantees and developing certification against other types of attacks, including non-additive ones such as deepfakes.

Acknowledgements

The authors acknowledge the support from the Russian Science Foundation grant No. 25-41-00091. The authors are grateful to Olesya Kuznetsova for valuable discussions during the preparation of this paper.

References

Andriushchenko, M.; and Flammarion, N. 2020. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33: 16048–16059.

Apostolidis, K. D.; and Papakostas, G. A. 2021. A survey on adversarial deep learning robustness in medical image analysis. *Electronics*, 10(17): 2132.

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, 274–283. PMLR.

Bredin, H.; Yin, R.; Coria, J. M.; Gelly, G.; Korshunov, P.; Lavechin, M.; Fustes, D.; Titeux, H.; Bouaziz, W.; and Gill, M.-P. 2020. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7124–7128. IEEE.

Castan, D.; McLaren, M.; Ferrer, L.; Lawson, A.; and Lozano-Diez, A. 2017. Improving Robustness of Speaker Recognition to New Conditions Using Unlabeled Data. In *Interspeech*, 3737–3741.

Cheng, C.-H.; Nührenberg, G.; and Ruess, H. 2017. Maximum resilience of artificial neural networks. In *International Symposium on Automated Technology for Verification and Analysis*, 251–268. Springer.

Chung, J. S.; Nagrani, A.; and Zisserman, A. 2018. VoxCeleb2: Deep Speaker Recognition. In *19th Annual Conference of the International Speech Communication Association, Interspeech 2018, Hyderabad, India, September 2-6, 2018*, 1086–1090. ISCA.

Clopper, C. J.; and Pearson, E. S. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4): 404–413.

Cohen, J.; Rosenfeld, E.; and Kolter, Z. 2019. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, 1310–1320. PMLR.

Deng, J.; et al. 2023. V-Cloak: Intelligibility-, Naturalness-Timbre-Preserving Real-Time Voice Anonymization. In *32nd USENIX Security Symposium (USENIX Security 23)*, 5181–5198.

Deng, Y.; Zheng, X.; Zhang, T.; Chen, C.; Lou, G.; and Kim, M. 2020. An analysis of adversarial attacks and defenses on autonomous driving models. In *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 1–10. IEEE.

Desplanques, B.; Thienpondt, J.; and Demuynck, K. 2020. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, 3830–3834. ISCA.

Fan, M.; Chen, C.; Wang, C.; Zhou, W.; and Huang, J. 2023. On the Robustness of Split Learning Against Adversarial Attacks. In *ECAI 2023 - 26th European Conference on Artificial Intelligence*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, 668–675. IOS Press.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR*.

Gowal, S.; Dvijotham, K. D.; Stanforth, R.; Bunel, R.; Qin, C.; Uesato, J.; Arandjelovic, R.; Mann, T.; and Kohli, P. 2019. Scalable verified training for provably robust image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4842–4851.

Hao, Z.; Ying, C.; Dong, Y.; Su, H.; Song, J.; and Zhu, J. 2022. GSmooth: Certified Robustness against Semantic Transformations via Generalized Randomized Smoothing. In *International Conference on Machine Learning*, 8465–8483. PMLR.

Hayes, J. 2020. Extensions and limitations of randomized smoothing for robustness guarantees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Hoeffding, W. 1994. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, 409–426. Springer.

Inkawhich, N.; Wen, W.; Li, H. H.; and Chen, Y. 2019. Feature space perturbations yield more transferable adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7066–7074.

Jeong, J.; and Shin, J. 2020. Consistency regularization for certified robustness of smoothed classifiers. *Advances in Neural Information Processing Systems*, 33: 10558–10570.

Kaviani, S.; Han, K. J.; and Sohn, I. 2022. Adversarial attacks and defenses on AI in medical imaging informatics: A survey. *Expert Systems with Applications*, 198: 116815.

Kaziakhmedov, E.; Kireev, K.; Melnikov, G.; Pautov, M.; and Petiushko, A. 2019. Real-world attack on MTCNN face detection system. In *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, 0422–0427. IEEE.

Khrulkov, V.; and Oseledets, I. 2018. Art of singular vectors and universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8562–8570.

Koch, G.; Zemel, R.; Salakhutdinov, R.; et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2, 1–30. Lille.

Komkov, S.; and Petiushko, A. 2021. Advhat: Real-world adversarial attack on arcface face id system. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 819–826. IEEE.

Lan, J.; Zhang, R.; Yan, Z.; Wang, J.; Chen, Y.; and Hou, R. 2022. Adversarial attacks and defenses in Speaker Recognition Systems: A survey. *Journal of Systems Architecture*, 127: 102526.

Lecuyer, M.; Atlidakis, V.; Geambasu, R.; Hsu, D.; and Jana, S. 2019. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, 656–672. IEEE.

Li, L.; Xie, T.; and Li, B. 2023. Sok: Certified robustness for deep neural networks. In *2023 IEEE Symposium on Security and Privacy (SP)*, 1289–1310. IEEE.

Li, L.; et al. 2021. Tss: Transformation-specific smoothing for robustness certification. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 535–557.

Li, Z.; et al. 2020. Practical adversarial attacks against speaker recognition systems. In *Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications*, 9–14.

Liu, X.; Tan, H.; Zhang, J.; Li, A.; and Gu, Z. 2024. Transferable universal adversarial perturbations against speaker recognition systems. *World Wide Web*, 27(3): 33.

Lyu, Z.; Ko, C.-Y.; Kong, Z.; Wong, N.; Lin, D.; and Daniel, L. 2020. Fastened crown: Tightened neural network robustness certificates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5037–5044.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *6th International Conference on Learning Representations, ICLR*.

Meng, Q.; Zhao, S.; Huang, Z.; and Zhou, F. 2021. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14225–14234.

Muravev, N.; and Petiushko, A. 2022. Certified Robustness via Randomized Smoothing over Multiplicative Parameters of Input Transformations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 3366–3372.

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. VoxCeleb: A Large-Scale Speaker Identification Dataset. In *Interspeech*, 2616–2620. ISCA.

Olivier, R.; and Raj, B. 2021. Sequential Randomized Smoothing for Adversarially Robust Speech Recognition. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Pautov, M.; Kuznetsova, O.; Tursynbek, N.; Petiushko, A.; and Oseledets, I. 2022. Smoothed embeddings for certified few-shot learning. *Advances in Neural Information Processing Systems*, 35: 24367–24379.

Pulina, L.; and Tacchella, A. 2010. An abstraction-refinement approach to verification of artificial neural networks. In *International Conference on Computer Aided Verification*, 243–257. Springer.

Qin, Z.; Zhao, W.; Yu, X.; and Sun, X. 2023. OpenVoice: Versatile Instant Voice Cloning. *arXiv preprint arXiv:2312.01479*.

Ravanelli, M.; et al. 2021. SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Salman, H.; Li, J.; Razenshteyn, I.; Zhang, P.; Zhang, H.; Bubeck, S.; and Yang, G. 2019. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in Neural Information Processing Systems*, 32.

Singla, S.; and Feizi, S. 2020. Second-order provable defenses against adversarial attacks. In *International Conference on Machine Learning*, 8981–8991. PMLR.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30.

Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; and Khudanpur, S. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5329–5333. IEEE.

Su, J.; Vargas, D. V.; and Sakurai, K. 2019. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5): 828–841.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR*.

Wan, L.; Wang, Q.; Papir, A.; and Moreno, I. L. 2018. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4879–4883. IEEE.

Wang, F.; and Liu, H. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2495–2504.

Wang, H.; Zheng, S.; Chen, Y.; Cheng, L.; and Chen, Q. 2023a. CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking. In *Interspeech*, 5301–5305. ISCA.

Wang, H.; et al. 2023b. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Wang, X.; et al. 2024. ASVspoof 5: Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale. In *ASVspoof Workshop 2024*.

Wang, Y.; Sun, T.; Li, S.; Yuan, X.; Ni, W.; Hossain, E.; and Poor, H. V. 2023c. Adversarial attacks and defenses in machine learning-empowered communication systems and networks: A contemporary survey. *IEEE Communications Surveys & Tutorials*.

Wu, H.; et al. 2021. Improving the adversarial robustness for speaker verification by self-supervised learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 202–217.

Yamagishi, J.; et al. 2021. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*.

Yang, G.; Duan, T.; Hu, J. E.; Salman, H.; Razenshteyn, I.; and Li, J. 2020. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, 10693–10705. PMLR.

Yu, Y.-Q.; and Li, W.-J. 2020. Densely Connected Time Delay Neural Network for Speaker Verification. In *Interspeech*, 921–925.

Yuan, Z.; Zhang, J.; Jia, Y.; Tan, C.; Xue, T.; and Shan, S. 2021. Meta Gradient Adversarial Attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7748–7757.

Zhang, X.; Zhang, X.; Sun, M.; Zou, X.; Chen, K.; and Yu, N. 2023. Imperceptible black-box waveform-level adversarial attack towards automatic speaker recognition. *Complex & Intelligent Systems*, 9(1): 65–79.

Zhou, Z.; Chen, J.; Wang, N.; Li, L.; and Wang, D. 2023. Adversarial data augmentation for robust speaker verification. In *Proceedings of the 2023 9th International Conference on Communication and Information Processing*, 226–230.