Автономная некоммерческая образовательная организация высшего образования «Сколковский институт науки и технологий»

На правах рукописи

Корж Дмитрий Сергеевич

Устойчивость моделей глубокого обучения

Специальность: 1.2.1. Искусственный интеллект и машинное обучение

АВТОРЕФЕРАТ

диссертации на соискание учёной степени кандидата физико-математических наук

Москва — 2025

Работа выполнена в Автономной некоммерческой образовательной организации высшего образования «Сколковский институт науки и технологий».

Научный руководитель: доктор физико-математических наук, профессор
Оселедец Иван Валерьевич

Научный консультант: кандидат физико-математических наук
Рогов Олег Юрьевич

Защита состоится **«день» месяц 2025 года в X часов Y минут** на заседании диссертационного совета **1.2.1.2.**, созданного на базе Автономной некоммерческой образовательной организации высшего образования «Сколковский институт науки и технологий» (Сколтех)

по адресу: Территория Инновационного Центра «Сколково», Большой бульвар д.30, стр.1, Москва 121205, Россия.

С диссертацией можно ознакомиться в библиотеке Сколтеха и на сайте организации https://dissovet.skoltech.ru.

Автореферат разослан «____»________________ 2025 г.

Ученый секретарь
диссертационного совета,
кандидат
физико-математических наук **Копелевич Григорий Александрович**

Autonomous Non-Profit Organization for Higher Education
«Skolkovo Institute of Science and Technology»

As a manuscript

Korzh Dmitrii Sergeevich

Robustness of Deep Learning Models

Speciality: 1.2.1. Artificial Intelligence and Machine Learning

DISSERTATION ABSTRACT

**of the dissertation for the Degree of
Doctor of Philosophy in Physics and Mathematics**

Moscow — 2025

The dissertation was prepared at the Autonomous Non-Profit Organization for Higher Education «Skolkovo Institute of Science and Technology».

Scientific Advisor: Doctor of Physical and Mathematical Siences, Professor
Oseledets Ivan Valerievich

Scientific Consultant: Candidate of Physical and Mathematical Sciences
Rogov Oleg Yurievich

The defense will take place on **Month Day, 2025, at HH:MM p.m.** at the meeting of the Dissertation Council **1.2.1.2.**, based at Skolkovo Institute of Science and Technology

address: The territory of the Skolkovo Innovation Center, Bolshoy Boulevard, 30, p.1, Moscow 121205, Russia.

The dissertation can be found in the library of Skoltech or on the website https://dissovet.skoltech.ru.

The abstract was sent out on «____»________________, 2025.

Academic secretary
of the Dissertation council,
Candidate of Physical and
Mathematical Sciences

Kopelevich Grigoriy Aleksandrovich

General description of work

Background and Relevance of the Work. Deep learning (DL) models have rapidly transformed both academia and real-world applications, such as medical diagnosis [1], biometrics [2], speech processing [3], and generative technologies including Large Language Models (LLMs) [4; 5]. Despite these advances, DL models remain limited in reliability and robustness. Firstly, models are vulnerable to adversarial perturbations, which are minor changes in input that might be insignificant to human perception but can drastically degrade performance [6; 7]. Secondly, DL models can also fail under natural perturbations such as noise, blurring, brightness, or contrast adjustment. Moreover, the spread of generative technologies has created new threats to security and privacy, including realistic voice cloning [8] and multimodal deepfakes. In high-risk applications like medicine, biometrics, or self-driving, such failures can have severe consequences. These risks highlight the importance of investigating vulnerabilities, understanding model limitations, and designing new empirical and certifiable defenses to enhance robustness, trustworthiness, and privacy of DL systems.

Degree of prior research on the topic. Deep learning models are well known to be vulnerable to adversarial perturbations [6]. Formally, for a classifier $f : \mathbb{R}^n \mapsto [0,1]^C$, an adversarial attack aims to change the predicted class

$$\arg\max f(x+\delta) \neq \arg\max f(x), \tag{1}$$

while constraining the perturbation norm $\|\delta\|_p$ (e.g., ℓ_2, ℓ_∞). A classical example is the FGSM attack [7]: $\delta = \varepsilon \, \mathrm{sign}\left[\nabla_x \mathcal{L}(f, x, y)\right]$, where y is a ground truth class, ε is an ℓ_∞ attack level. Numerous attack types exist [9], including white-box [10], black-box [11], patch [12], and physical attacks [13]. To counter them, empirical defenses such as adversarial training were proposed [14], but they often fail against novel attacks [15], leading to a cat-and-mouse game. Certification approaches [16; 17] instead provide provable robustness guarantees within a perturbation set. Among them, randomized smoothing (RS) [18; 19] has become the most widely adopted: a smoothed classifier

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

is shown to be a Lipschitz function, allowing robustness certification due to the property that the model's output perturbation can be limited for a fixed input perturbation level. When $g(x)$ is confident in predicting the correct class i_1 for the input x, $g(x)_{i_1} = p_{i_1} \geq p_{i_2} = \max_{i \neq i_1} g(x)_i$ then it is robust in ℓ_2-ball around x of radius

$$R = \frac{\sigma}{2}\left(\Phi^{-1}(p_{i_1}) - \Phi^{-1}(p_{i_2})\right), \tag{3}$$

$$\forall \delta : \|\delta\|_2 < R \mapsto \operatorname{argmax} g(x) = \operatorname{argmax} g(x + \delta), \tag{4}$$

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

RS has also been applied to semantic transformations [20; 21] and few-shot models [22], though limited to restricted transformation sets and without covering speaker identification tasks.

In parallel, voice anti-spoofing has gained attention with the rise of text-to-speech, voice cloning, and conversion models. Benchmarks and contests such as ASVspoof [23] and ADD [24] drive progress in methods like AASIST [25], which combines spectral and temporal features with graph neural networks. Further improvements are achieved using self-supervised encoders [26; 27], augmentations [28], and specialized losses [29], though performance still degrades in unseen domains, limiting the reliability in applications.

Simultaneously, adversarial methods are also studied for speaker privacy (voice anonymization). Regenerative approaches [30] and universal adversarial perturbations (UAPs) [31] aim to conceal speaker identity while preserving intelligibility. UAPs are lightweight and suitable for real-time applications, but their performance needs enhancement. They struggle with long audio recordings, balancing privacy with speech quality, and can underperform for new speakers and biometrics models, limiting the empirical privacy robustness in applications.

Dissertation objectives. This dissertation investigates methods to improve the robustness, trustworthiness, and privacy of deep learning models, with a particular focus on classification and few-shot classification tasks. **The goal** of this dissertation is to develop novel improved certified and empirical approaches that enhance robustness, reliability, and privacy of deep learning models while preserving their target performance without imposing significant application limitations. To achieve this goal, the following **problems** are addressed:

1. Development of an enhanced certification method for a broad class of resolvable semantic transformations, without underperforming existing methods on the transformations they certify. This requires analyzing transformation-dependent Lipschitz continuity of smoothed classifiers, deriving corresponding robustness certificates, and designing a numerical evaluation scheme.
2. Development of the improved certification method for the prototypical (embedding) few-shot models against norm-bounded perturbations via analysis of scalar mapping from the embedding space and derivation of theoretical robustness guarantees based on its Lipschitz properties.
3. Development of an enhanced UAP-based speaker privacy method to achieve a better balance (trade-off) between fooling rate and preservation of perceptual and speech recognition quality.

4. Development of a novel architecture and training strategy for the voice anti-spoofing models to improve generalization to previously unseen synthetic speech generators.

Scientific novelty.

1. The thesis presents a new universal certification approach against compositions of resolvable semantic transformations based on randomized smoothing, complemented by an efficient numerical verification procedure and yielding state-of-the-art robustness guarantees across multiple datasets. The approach can be applied to any resolvable compositional transformation, in contrast to previous works, and demonstrates state-of-the-art results.
2. A novel randomized smoothing-based certification approach for few-shot embedding models against additive perturbations is proposed. This result also includes the first provable robustness guarantees for the speaker identification task, achieves state-of-the-art results, and establishes the first certification benchmark for this problem.
3. The thesis presents the novel speaker privacy UAP method that achieves state-of-the-art fooling rate and word error rate results, especially for long audio. An application of exponential total variance loss and a length-independent tiling during training is proposed, resulting in superior performance in fooling rate, perceptual quality, and robustness on long-duration audio compared to existing approaches. Additionally, a fair length and noise-agnostic evaluation protocol is proposed.
4. The dissertation also presents the novel voice anti-spoofing architecture enhanced via Kolmogorov–Arnold neural layers, advanced audio pre-processing, augmentations, and a self-supervised backbone to improve overall model generalization.

Theoretical and practical significance. This work presents a theoretical analysis of the transformation-dependent Lipschitz continuity of smoothed image classifiers with respect to transformation parameters and derives corresponding robustness certificates. It provides a theoretically robust guarantee that a smoothed image classifier will be robust against a given resolvable semantic perturbation within the considered parameter set. Also, this work provides new theoretical certification guarantees for the few-shot embedding models based on a scalar mapping from the embedding space. Furthermore, the following methods for robustness evaluation and certification of neural networks, speaker privacy protection, and voice anti-spoofing have been developed:

1. General Lipschitz, theoretical, numerical, and experimental framework for the certification of image classifiers against resolvable semantic transformation and robustness evaluation.

2. ASI Certification, theoretical and experimental framework for the certification of ASI models or other few-shot embedding models, and evaluation of their robustness against bounded additive perturbations.
3. Voice-UAP, a speaker privacy (anonymization) empirical method which is targeted to fail voice biometrics models in a real-time speaker-agnostic scenario, exploiting vulnerabilities of deep learning speaker recognition models.
4. AASIST-3 (KAN-AASIST), a novel voice anti-spoofing model, oriented on the empirically robust detection of artificially generated or modified speech.

Research methodology. The methodology includes methods of machine learning, deep learning, classical algorithms, and data structures. The mathematical methods used in this dissertation include aspects of real analysis, linear algebra, numerical methods (including methods of digital signal processing), probability theory, and statistics.

Propositions for defense.

1. A theoretical certification guarantees of classifiers against resolvable transformations and their compositions, and the corresponding semi-automatic numerical certification procedure.
2. A theoretical certification method for the few-shot embedding models against additive, norm-bounded perturbations, and its particular application for the speaker identification task.
3. A speaker privacy (anonymization) method based on universal adversarial perturbation trained with an application of exponential total variance loss and length-agnostic tiling.
4. Voice anti-spoofing models based on self-supervised audio encoders and the incorporation of Kolmogorov-Arnold neural layers.

Validation of the research results, reliability. The theoretical findings presented in this dissertation are formulated as mathematical statements supported by rigorous proofs. The effectiveness of the proposed theoretical methods for model certification and empirical methods for voice anonymization and voice anti-spoofing has been demonstrated through a broad range of experiments, including comparative evaluations against state-of-the-art approaches, in which the proposed solutions demonstrated superior performance. All developed methods and experimental settings are described in detail in the main text, in the appendix, and in open-sourced code.

Publications. During the PhD studies, 6 works were accepted, 5 of which are already published. The main results of the dissertation are presented in 4 articles: three in CORE A/A* conferences and one in a workshop of CORE A conference. Relevant results of the fifth article are included in the appendix, while the sixth work is not discussed in this dissertation.

The list of the author's **main publications**:

1. **Dmitrii Korzh**, Mikhail Pautov, Olga Tsymboi, and Ivan Oseledets. General Lipschitz: Certified robustness against resolvable semantic transformations via transformation-dependent randomized smoothing //ECAI 2024. – IOS Press, 2024. – C. 1591-1598. (CORE A). [A1].
2. **Dmitrii Korzh**, Elvir Karimov, Mikhail Pautov, Oleg Y. Rogov, and Ivan Oseledets. Certification of speaker recognition models to additive perturbations //Proceedings of the AAAI Conference on Artificial Intelligence. – 2025. – T. 39. – №. 17. – C. 17947-17956. (CORE A*). [A2].
3. Elvir Karimov, Alexander Varlamov, Danil Ivanov, **Dmitrii Korzh**, Oleg Y. Rogov. Novel Loss-Enhanced Universal Adversarial Patches for Sustainable Speaker Privacy //Proc. Interspeech 2025. – 2025. – C. 1513-1517. (CORE A). [A3].
4. Kirill Borodin*, Vasiliy Kudryavtsev*, **Dmitrii Korzh***, Alexey Efimenko*, Grach Mkrtchian, Mikhail Gorodnichev, Oleg Y. Rogov. AASIST3: KAN-enhanced AASIST speech deepfake detection using SSL features and additional regularization for the ASVspoof 2024 Challenge //Proc. ASVspoof 2024. – 2024. – C. 48-55. [A4].

Other publications:

1. Artyom Iudin, **Dmitrii Korzh**, Matvey Skripkin, Oleg Y. Rogov. Clarispeech: LLM-Enhanced Speech Recognition Post-Correction - 2025. This article [A5] was accepted to the Artificial Intelligence and Natural Language (AINL) Conference 2025. The proceedings are under preparation. Several results of this work are discussed in the appendix of the dissertation.
2. Alexey Dontsov, **Dmitrii Korzh**, Alexey Zhavoronkin, Boris Mikheev, Denis Bobkov, Aibek Alanov, Oleg Y. Rogov, Ivan Oseledets, and Elena Tutubalina. Clear: Character unlearning in textual and visual modalities //Findings of the Association for Computational Linguistics: ACL 2025. – 2025. (CORE A*). This work [A6] is not discussed in the dissertation.

Approbation. The author presented the results at several conferences and workshops:

1. ASVspoof 2024 Workshop (Kos Island, Greece; August 2024). Oral presentation. Topic: “AASIST3: KAN-enhanced AASIST speech deepfake detection using SSL features and additional regularization for the ASVspoof 2024 Challenge”.
2. ECAI 2024 (Santiago de Compostela, Spain; October 2024). Oral presentation. Topic: “General Lipschitz: Certified robustness against resolvable semantic transformations via transformation-dependent randomized smoothing”.

3. AINL 2024 (Novosibirsk, Russia; April 2025). Oral presentation. Topic: “Clarispeech: LLM-Enhanced Speech Recognition Post-Correction”.
4. RSI AIRI and SAIL MTUCI Workshop “Cybersecurity and robustness in the AI era” (Moscow, Russia; August 2025). Oral presentation. Topic: “Trusted AI in voice biometrics and multimodal models”.
5. Interspeech 2025 (Rotterdam, the Netherlands; August 2025). Poster presentation. Topic: “Novel Loss-Enhanced Universal Adversarial Patches for Sustainable Speaker Privacy”.

Author’s contribution. The author contributed to the presented works as follows:

1. In the work “General Lipschitz: Certified robustness against resolvable semantic transformations via transformation-dependent randomized smoothing” [A1], the author contributed to the formulation and proof of the theorem and lemmas, namely devoted to the certification condition against resolvable semantic transformations, estimation of logarithm and its derivative of smoothing probability density distribution for resolvable transformations. The author contributed to numerical implementations of the proposed methods, conducted all the extensive experiments, analyzed the results, estimated the error bounds of the numerical scheme, and provided an analytical derivation example. Together with co-authors, the author prepared the manuscript’s text and its revisions.
2. In the work “Certification of speaker recognition models to additive perturbations” [A2], the author proposed the main idea of the article, contributed to the formulation and proof of the theorem (certification of vector function against additive perturbations). The author contributed to the experimental implementation of the proposed methods, designed the evaluation methodology, conducted the main experiments, and analyzed the results. Together with co-authors, the author prepared the manuscript’s text and its revision.
3. In the work “Novel Loss-Enhanced Universal Adversarial Patches for Sustainable Speaker Privacy” [A3], the author proposed the approach, contributed to its experimental development, and to the preparation of the final (best) method. The author significantly participated in the implementation, experiments, and evaluation of the methods. The author prepared several parts of the manuscript.
4. In the work “AASIST3: KAN-enhanced AASIST speech deepfake detection using SSL features and additional regularization for the ASVspoof 2024 Challenge” [A4], the author significantly contributed to the experimental development and to the final proposed method, designed the experimental framework, conducted extensive experiments, and prepared several parts of the manuscript.

Contents of the dissertation

The introduction provides an overview of the research area and prior work, discusses the relevance of the research, defines its goal and objectives, and highlights the scientific novelty and practical significance of the dissertation.

The first chapter is devoted to the certified robustness of image classifiers against semantic perturbations and provides novel state-of-the-art results, including results for the transforms for which certified guarantees were not previously proposed. Section 1.1 provides a general introduction and motivation to this problem, while Section 1.2 describes relevant works and competitors, especially certification methods against semantic transformation.

Section 1.3 is devoted to the proposed certification approach and its theoretical analysis. In subsection 1.3.1, preliminaries are introduced. Suppose a parametric mapping $\phi : X \times \Theta \to X$ corresponds to a semantic perturbation of the input of the classification model, where Θ is the space of parameters of the perturbation. The goal of this paper is to construct a framework to certify that a classifier is robust at $x \in X$ to the transformation $\phi(x, \cdot)$ for some set of parameters $\mathcal{B}(\beta_0)$, where $\phi(x, \beta_0) = x$ for all $x \in X$. A transform $\phi : X \times \Theta \to X$ is called *resolvable* [21] if for any parameter $\alpha \in \Theta$ there exists a continuously differentiable function $\gamma : \Theta \times \Theta \to \Theta$ such that for all $x \in X$ and all $\beta \in \Theta$

$$\phi(\phi(x, \alpha), \beta) = \phi(x, \gamma(\alpha, \beta)). \tag{5}$$

Subsection 1.3.2 defines the smoothed classifier $h(x)$ in the form of expectation over perturbation density $\rho(y|x)$ conditioned on the observed sample x:

$$h(x) = \int_{\Theta} f(\phi(x, \alpha))\rho(\phi(x, \alpha)|x)d\alpha = \int_{\mathbb{R}^n} f(y)\rho(y|x)dy. \tag{6}$$

Moreover, the goal of the study is defined: to present a procedure that guarantees a smoothed model to be robust to semantic perturbations, that is

$$\underset{i \in \{1,2,\ldots C\}}{\operatorname{argmax}} \; h_i(x) = \underset{i \in \{1,2,\ldots C\}}{\operatorname{argmax}} \; h_i(\phi(x, \beta)), \tag{7}$$

for all $\beta \in \mathcal{B}(\beta_0)$, where $\phi(x, \beta_0) = x$. One way to achieve robustness to parametric perturbation is to bound the Lipschitz constant of the classifier from Eq. (6) with respect to the transformation parameters. For this purpose, the perturbation-dependent conditional smoothing density $\rho(y|x)$, which has to be continuously differentiable with respect to perturbation parameters, is introduced in the form:

$$\rho(y|\hat{x}) = \frac{\int_{\Theta} \exp\left\{-\frac{\|y - \phi(\hat{x}, \alpha)\|_2^2}{2\sigma^2}\right\} \tau(\alpha)d\alpha}{(2\pi\sigma^2)^{\frac{n}{2}}}, \tag{8}$$

where $\hat{x} = \phi(x, \beta)$ is a perturbed sample, $y = \phi(\hat{x}, \alpha) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, $\tau(\alpha)$ is the smoothing distribution of the transformation.

Subsection 1.3.3 presents the main theoretical result and a sketch of the proof. Let $x \in X$ be the input object of class c and assume that the smoothed classifier h defined in Eq. (6) correctly classifies x with significant confidence, i.e., $h_c(x) > \frac{1}{2}$. Then, the following result holds.

Theorem 1. *Certification condition. Let $\beta(t) : [0,1] \to \Theta$ be a smooth curve such that $\beta(0) = \beta_0$ and $\beta(1) = \beta$. Then there exist mappings $\xi : [0, 1] \to \mathbb{R}$ and $\hat{g}(\beta) : \Theta \to \mathbb{R}$ such that if $\hat{g}(\beta) < -\xi(1 - h_c(x)) + \xi(1/2)$, then h is robust at x for all $\beta \in \beta(t)$, where $t \in [0, 1]$.*

The Theorem (1) anticipates a numerical procedure to compute certification functions ξ, $\hat{g}$ semi-automatically, based on Lipschitz-continuity analysis of the smoothed classifier via gradient bounding. Subsections 1.3.4-1.3.6 explain the required methods and algorithms in detail.

It is assumed that the input sample x is fixed and the smoothed model h is threat as the function of the perturbation parameter β, namely $h(\phi(x, \beta)) \equiv h(\hat{x}) \equiv h(x,\beta) \equiv h(\beta)$ for simplicity. Within the proposed framework, functions ξ, $\hat{g}$ are derived as the ones bounding the smoothed classifier's directional derivative with respect to the perturbation parameter:

$$\langle \nabla_\beta h(\beta), \beta \rangle \leq \tilde{g}(h(\beta), \beta) \leq p(h) g(\beta), \tag{9}$$

where $\tilde{g}(h(\beta), \beta)$ is an upper bound on the directional derivative. This function is also bounded by the product of a function of h and the function of β. If the functions $p(h)$ and $g(\beta)$ are known, the mappings from Theorem 1 have the following form:

$$\xi(h) = \int \frac{1}{p(h)} dh, \quad \hat{g}(\beta) = \int_0^1 g(\beta(t)) dt. \tag{10}$$

A bound for the directional derivative in the form from Eq. (9) is used to compute functions $\xi, \hat{g}$ from Theorem (1). However, in the case of a complicated form of conditional density from Eq. (8), it may be unfeasible to construct an exact bound. Instead, a numerical procedure to bound directional derivatives of the smoothed model is proposed. The gradient of the smoothed classifier with respect to the parameters of transformation has the following form:

$$\nabla_\beta h = \int f(y) \nabla_\beta \rho(y|\hat{x}) dy = \int f(y) \eta(y, \hat{x}) \rho(y|\hat{x}) dy, \tag{11}$$

where $\eta(y, \hat{x}) = \nabla_\beta \log \rho(y|\hat{x})$. Given fixed β, the problem of bounding the directional derivative $\langle \nabla_\beta h(\beta), \beta \rangle$ is equivalent to the search for the worst base classifier, i.e., the one with the largest bound. The search for the worst classifier

q^* may be formulated as the optimization problem:

$$\begin{aligned} q^* =& \operatorname*{argmax}_{q \in \mathcal{Q}} \int q(y) \eta(y, \hat{x}) \rho(y|\hat{x}) dy, \\ \text{s.t.} \quad h(\hat{x}) =& \int q(y) \rho(y|\hat{x}) dy, \end{aligned} \tag{12}$$

where $\mathcal{Q} = \{q | q : X \to [0, 1]\}$ is the set of all binary classifiers. Under the specific choice of resolvable transform ϕ and perturbation distribution $\tau(\alpha)$, the problem from Eq. (12) admits the analytical solution. In general, if the evaluation of $\rho(y|\hat{x})$ and $\eta(y,\hat{x})$ is available, the solution of the problem from Eq. (12) could be obtained numerically via sampling, and the corresponding procedure is described in the full version of the dissertation.

Since the exact evaluation of the density from Eq. (8) is challenging, it is proposed to emulate sampling from the conditional density $\rho(y|\hat{x})$ by estimating the gradient of the log-density $\eta(y, \hat{x})$ from Eq. (11). While the derivation of the Laplace approximation of the log density admits an arbitrary parametric transform ϕ, for the resolvable one, there exists a closed-form limit when $\sigma \to 0$. The last is summarized in Lemma 1, allowing computation of log-density either analytically or through automatic differentiation tools.

Lemma 1. *Let $\gamma(\alpha, \beta)$ be the resolving function: $\phi(\phi(x, \beta), \alpha) = \phi(x, \gamma(\alpha, \beta))$. Then, the formula for the logarithm of the conditional density from Eq.* (8) *has the limit when $\sigma \to 0$ in the form*

$$\log \rho(y|\hat{x}) = -\frac{1}{2} \log \det J^\top J + \log \tau(\alpha), \quad J = \frac{\partial \phi}{\partial \alpha}. \tag{13}$$

Accordingly, for the resolvable transforms $\rho(y|\hat{x}) = \frac{\tau(\alpha)}{\sqrt{\det J^T J}}$, which, for example, for additive transforms simplifies to the well-known $\tau(\alpha)$. Both density and the logarithm of the density can be evaluated analytically or via automatic differentiation, though numerically it is more precise to do so for the logarithm of the density. The automatic differentiation, however, might require the mapping from normally distributed random variables to variables distributed as $\tau(\alpha)$. Moreover, in the analytical approach, the integration with the more complex (often multidimensional) $\tau(\alpha)$ might still be required. If the log-density $\log \rho(y|\hat{x})$ is known, the expression for $\eta(y, \hat{x}) = \nabla_\beta \log \rho(y|\hat{x})$ is given by the following lemma:

Lemma 2 (Gradient of log-density for resolvable transformations)**.** *Suppose that the log-density $\log \rho(y|\hat{x}) = z(\alpha, \beta) = z(\alpha(\beta), \beta)$ is known. Then*

$$\eta(y|\hat{x}) = \nabla_\beta z = \frac{\partial z}{\partial \beta} - \frac{\partial z}{\partial \alpha} \left(\frac{\partial \gamma}{\partial \alpha} \right)^\dagger \frac{\partial \gamma}{\partial \beta},$$

where γ is a resolving function of the transform: $\phi(\phi(x, \beta), \alpha) = \phi(x, \gamma(\alpha, \beta))$.

The overall numerical procedure is presented in Algorithms 1, 2 in the full version of the dissertation.

Table 1 — Quantitative results on ImageNet dataset. Smoothing distributions and certified robust accuracy are reported for the proposed approach and competitors' methods. The best results are highlighted in **bold**, underlined denotes equivalent performance. Symbol $''-''$ in the table corresponds to the transformation in which a method does not certify the model against the given distribution parameters. The CRA is evaluated in the fixed parameter range $R_l \leq \beta \leq R_r$ for each transformation type. In the parameter column, c, b, γ, (T_x, T_y), r_b represent contrast, brightness, gamma-correction, translations, and Gaussian blur attacks' parameters, respectively. GL stands for the General Lipschitz, the proposed method. The architecture of the base model f is ResNet-50.

Transform	β	R_l	R_r	Distribution	GL	TSS [21]	MP [20]	CGS [32]
Brightness	b	-0.4	0.4	$\mathcal{N}(0, 0.3)$	0.68	0.68	–	0.67
Contrast	c	0.6	1.4	$\mathrm{LogNorm}(0, 0.3)$	0.68	–	0.68	0.67
Blur	r_b	1	4	$\mathrm{Exp}(0.3)$	0.59	0.59	–	0.0
Translation	T_x, T_y	-56	56	$\mathcal{N}(0, 50)$	**0.49**	0.28	–	0.45
Gamma	γ	1.0	2.0	$\mathrm{Rayleigh}(0.1)$	**0.66**	–	0.54	–
Gamma	γ	0.5	1.0	$\mathrm{Rayleigh}(0.1)$	**0.66**	–	0.61	–
Contrast Brightness	c b	0.6 -0.4	1.4 0.4	$\mathrm{LogNorm}(0, 0.6)$ $\mathcal{N}(0, 0.6)$	0.62	0.59	–	0.62
Gamma Contrast	γ c	0.8 0.6	1.4 2.0	$\mathrm{Rayleigh}(0, 0.1)$ $\mathrm{LogNorm}(0, 0.1)$	0.62	–	–	–
Brightness Translation	b T_x, T_y	-0.2 -56	0.2 56	$\mathcal{N}(0, 0.4)$ $\mathcal{N}(0, 30)$	**0.46**	0.02	–	–
Contrast Translation	c T_x, T_y	0.8 -25	1.2 25	$\mathrm{LogNorm}(0, 0.4)$ $\mathcal{N}(0, 30)$	0.09	–	–	–
Contrast Brightness Translation	c b T_x, T_y	0.8 -0.2 -15	1.2 0.2 15	$\mathrm{LogNorm}(0, 0.4)$ $\mathcal{N}(0, 04)$ $\mathcal{N}(0, 15)$	0.06	–	–	–
Translation Blur Brightness Contrast	T_x, T_y r_b b c	-3 1 -0.1 0.95	3 3 0.1 1.05	$\mathcal{N}(0, 10)$ $\mathrm{Rayleigh}(1)$ $\mathcal{N}(0, 0.3)$ $\mathrm{LogNorm}(0, 0.3)$	0.20	–	–	–

The experimental setting and results, including computational complexity estimation, are described in Section 1.4. The certified robust accuracy (CRA) was used for the evaluation, which is a fraction of correctly classified images x_i from the test set on which the certification condition is met.

The proposed approach was evaluated against [20; 21; 32] and the results were presented in Tables 1 and 2. The method achieved state-of-the-art robustness certificates for the majority of transformations and provided the first certified results against such transforms, as Gamma-Contrast and Contrast-Translation. Figure 1 illustrates CRA dependency on the parameter range of the attack.

Table 2 — Certified robust accuracy (CRA) for some attacks on CIFAR-10 and CIFAR-100 datasets. The best results are highlighted in **bold**, underlined denotes equivalent performance. For the contrast transform, the proposed GL method vs. MP has 86.2 vs. 86.2 and 45.6 vs. **46.0** for CIFAR-10 and CIFAR-100, respectively. The architecture of the base model is Resnet-110.

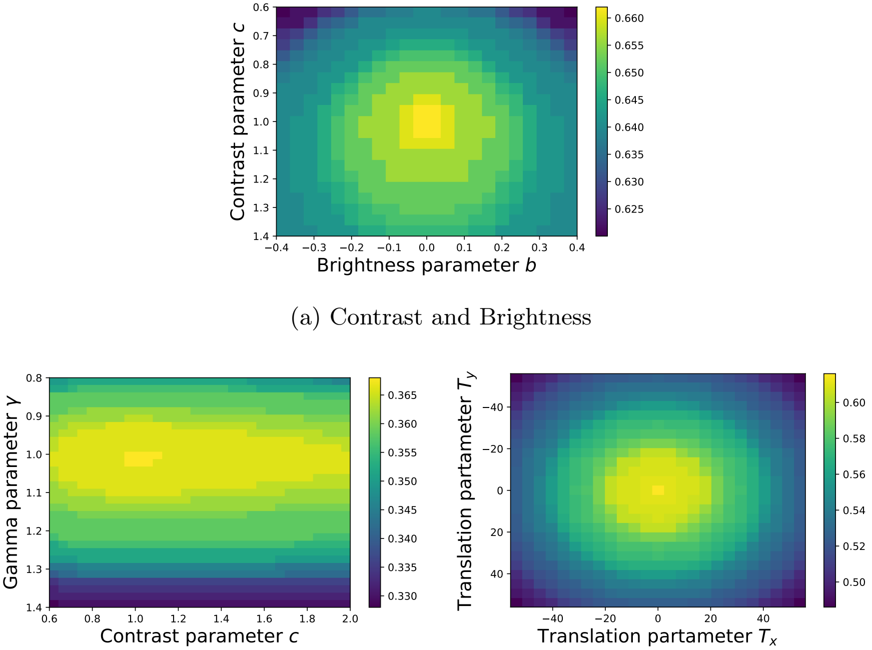
Transform	CIFAR-10			CIFAR-100		
	GL	TSS	GS	GL	TSS	GS
Brightness	**86.8**	86.6	85.6	**45.6**	43.8	43.2
Contrast	**86.2**	–	85.6	**45.6**	–	43.2
Blur	74.2	**75.4**	0.0	39.8	**41.8**	0.0
CB	85.5	83.4	85.5	**41.6**	38.0	41.4

Section 1.5 discusses limitations. For instance, the proposed method is suitable only for the resolvable transformation. It is also worth mentioning that, although the certification theorem provides deterministic guarantees, the smoothed model cannot be evaluated exactly, even for simple transformations, as deep learning models cannot be analytically represented under the integral; thus, Monte-Carlo sampling is used, leading to probabilistic guarantees via confidence estimation. The presented approach is based on the randomized smoothing technique; hence, the certified model can not be evaluated exactly. In the considered experimental setting, for the sample x of class c, the true value of the smoothed classifier $h_c(x)$ is estimated as the lower bound of the Clopper-Pearson confidence interval [33] over $N_{\max}$ samples for some confidence level α^*. Namely, $\hat{h}(x) = B(\alpha^*/2, n, N_{\max} - n + 1)$, where B is Beta distribution, $N_{\max}$ is the sample size and n is the number of perturbations for which $f(\phi(x, \alpha^j)) > \frac{1}{2}$. Thus, the approach produces certificates with probability $p \geq 1 - \alpha^*$, where α^* is the upper bound on the probability to return an overestimated lower bound for the value $h(x)$.

The error analysis of the proposed numeric scheme and its influence on the certification are also discussed. Section 1.6 finishes the first chapter with conclusions and possible future research directions.

The second chapter is devoted to the robustness of few-shot vector (embedding/prototypical) classifiers against ℓ_2-norm bounded additive perturbations and addresses the issues of robustness and privacy in deep learning voice biometrics models, presenting state-of-the-art results. Section 2.1 provides a general introduction, and Section 2.2 describes related works devoted to the speaker recognition tasks, adversarial attacks, empirical and certified defenses in various domains, including few-shot classification.

Section 2.3 defines the problem statement, provides an overview of the techniques used, and describes the proposed method and the proof of the main theoretical result for certifying embedding models against norm-bounded additive perturbations. Consider $f : \mathbb{R}^n \to \mathbb{R}^d$ as the base model that maps



(a) Contrast and Brightness

(b) Gamma and Contrast

(c) Translation

Figure 1 — CRA visualization for different transformations, ImageNet dataset. By design of the proposed approach, if the classifier is certified at the input point x for the parameter value β, it is certified for all parameters $\beta^* \in [\beta_0, \beta]$. The CRA values are presented by color bars. Remark: the CRA against the given transform in Table 1 is the infimum of CRAs on the corresponding plot.

input audios to normalized embeddings, where $\|f(\cdot)\|_2 = 1$, n is an input dimension, and d is an embedding dimension. After training the embedding model, new speakers, whom one wants to authorize later in the biometrics system, are required to enroll. 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. For simplicity, $M(k)$ is fixed in the experiments $M(k) = M$ for the fair comparison.

$$c_k = \frac{1}{M} \sum_{x \in S_k^e} f(x), \quad \|c_k\|_2 = 1, \tag{14}$$

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:

$$i_1 = \underset{k \in [1,\ldots,K]}{\operatorname{argmin}} \; \rho(f(x), c_k). \tag{15}$$

The automatic speaker identification (ASI) and few-shot models are equated in this chapter to emphasize that the proposed method is also applicable to other few-shot scenarios.

Suppose that $f : \mathbb{R}^n \to \mathbb{R}^d$ is the base vector (embedding) model. The smoothed vector model $g : \mathbb{R}^n \to \mathbb{R}^d$ might be obtained as the ordinary smoothed classifier following the Eq. (2) and the corresponding certification goal against additive perturbations can be described in a similar way, but instead of logits one should use some distance metric. 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. A scalar mapping is introduced $\phi : \mathbb{R}^d \to [0,1]$ in the form

$$\phi = \phi(g(x), c_{i_1}, c_{i_2}) = \frac{\langle g(x), c_{i_1} - c_{i_2} \rangle}{2\|c_{i_1} - c_{i_2}\|_2} + \frac{1}{2}, \tag{16}$$

then the following robustness guarantee holds:

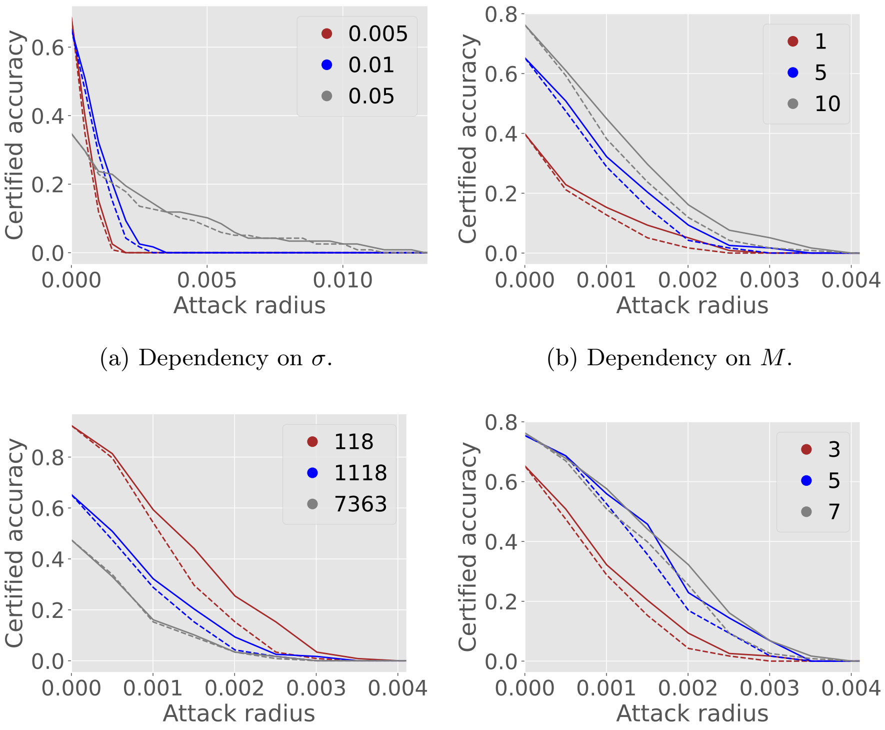
Theorem 2 (Main result). *For all additive perturbations $\delta : \|\delta\|_2 \le R(\phi, \sigma) = \sigma \Phi^{-1}(\phi)$*

$$\underset{k \in [1,\ldots K]}{\operatorname{argmin}} \|g(x) - c_k\|_2 = \underset{k \in [1,\ldots K]}{\operatorname{argmin}} \|g(x + \delta) - c_k\|_2, \tag{17}$$

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

The proof of this theorem is presented in the full version of the dissertation in Subsection 2.3.2.

Section 2.4 describes the corresponding numeric scheme and probabilistic guarantees. The reason is similar to that in the previous chapter, the base deep learning model cannot be “smoothed” analytically to obtain $g(x)$, the sample-mean estimation $\hat{g}(x)$ via Monte-Carlo sampling is used. However, in the few-shot setting, the model is not a classifier that returns logits, but an embedder, thus a different statistical evaluation for confidence intervals is



(a) Dependency on σ.

(b) Dependency on M.

(c) Dependency on K.

(d) Dependency on length (s).

Figure 2 — Pyannote model. Few-shot setting. Dependency of certified accuracy on the variance σ of the additive noise, on the number M of audios of a single speaker, the number of enrolled speakers K, and the audio length in seconds. The dashed lines represent results for SE, while the solid lines correspond to the presented method.

used, which collects distances from the sample-mean embedding to the nearest speaker centroids c_i, in contrast to those described in the previous chapter.

Section 2.5 describes experimental details, default parameters, considered common datasets and speaker recognition models, and presents the comparison with the competitor's approach Smoothed Embeddings (SE) [22]. Evaluation procedure implied considering K enrolled speakers and, for each of them, creating $c_k \in S^c$ of M randomly sampled speakers' enrollment audios, which are presented in S^e. To test the models, inference audios $x \in S^i$, $S^e \cap S^i = \emptyset$ were provided, where the number of unique test speakers in S^i is fixed and equal to 118. The CA was reported 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

$$i_1(x) = \underset{k\in\{1,\dots,K\}}{\operatorname{argmin}} \rho\left(g(x), c_k\right), \tag{18}$$

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

$$CA(S^c, S^i, \varepsilon) = \frac{|(x,y) \in S^i : R(x) > \varepsilon \ \wedge \ i_1(x) = y|}{|S^i|}, \tag{19}$$

where $R(x)$ is the certified radius from Theorem 2. In addition, empirical robust accuracy (ERA) was used, which is the fraction of correctly recognized perturbed audios $x + \delta$ for all sampled perturbations $\delta \leq l$, where l is the current attack level.

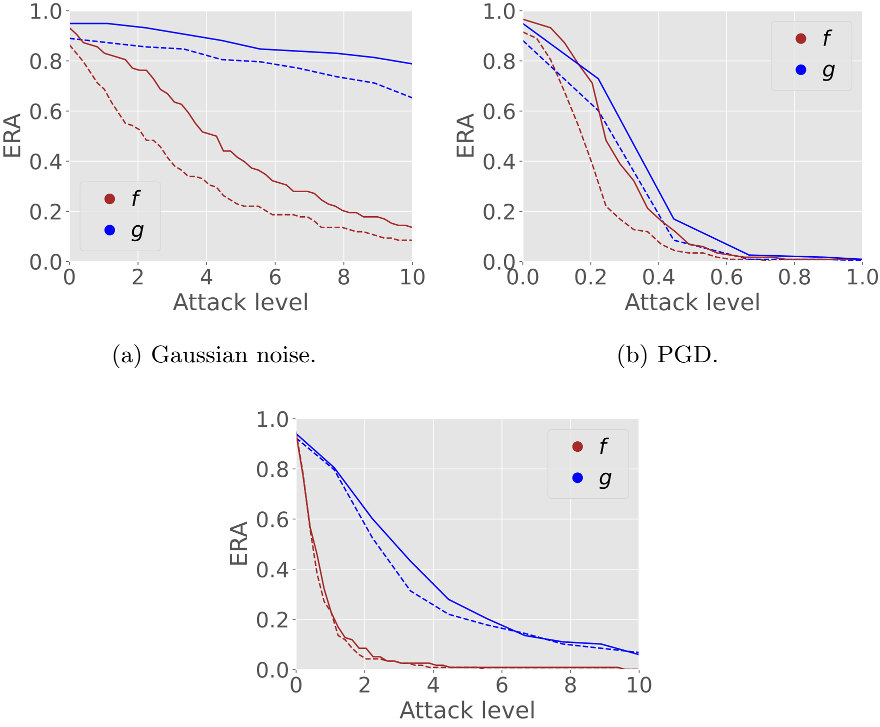
Section 2.6 presents and discusses the results, such as Figure 2, which illustrates the effects of varying a single parameter while keeping all others at their default values for the SE and the presented approaches. Several observations were obtained from these results:

- The proposed method demonstrates a marginal improvement across all scenarios compared to the SE approach;
- σ significantly impacts the certification system (the proposed, SE, and RS). Higher values lead to a more robust system, which comes at the expense of reduced accuracy (robustness-accuracy trade-off);
- Confidence level 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;
- An increase of the number of noise samples $N_{\max}$ enhances the certification process, while classification difficulty rises as the number of enrolled speakers K increases.

Section 2.6 continues the discussion that the empirical robustness of g and f is significantly better than the certification results of g, explaining it with the fact that the presented attacks do not necessarily convey the worst certification result, as stronger attacks exist, and the worst-case ERA might be closer to CA. Additionally, dependency on inference audio length, transferability to other tasks and types of attacks, and comparison with the RS over class probabilities are discussed. Section 2.7 concludes the study and provides possible future research directions.

The third chapter discusses two problems, devoted to the empirical robustness of speech models. In section 3.1, voice anonymization results are presented, while section 3.2 discusses voice antispoofing enhancement.

Subsection 3.1.1 introduces the speaker privacy (anonymization) systems (SPS), which are designed to confuse ASI models while maintaining the intelligibility and naturalness of modified audio. The focus of the study is an



(a) Gaussian noise.

(b) PGD.

(c) UAP.

Figure 3 — Pyannote model. Few-shot setting. Empirically Robust Accuracy (ERA) of functions f and g under different perturbations: Gaussian noise, PGD adversarial attack, speaker anonymization, and Universal Adversarial Patch (UAP). The dashed lines represent results for inference audio length $= 3$ seconds, while the solid lines correspond to the 5 seconds.

improvement of SPS based on additive adversarial attacks, specifically, on universal adversarial patches (UAP). Let f be a speaker recognition model that maps an input audio to a speaker embedding space $f : \mathbb{R}^n \to \mathbb{R}^d$. Given an audio sample $x \in \mathbb{R}^n$ and the set of enrolled speakers representations $e_i \in \mathbb{R}^d$, one can introduce a UAP $\hat{\delta} \in \mathbb{R}^l$, that aims to mislead the ASI model f for any speaker: $\operatorname{argmax}_k \rho(f(x+\delta), e_k) \neq \operatorname{argmax}_k \rho(f(x), e_k)$, where $\delta \in \mathbb{R}^n$ is a repeated UAP to suit the length of x. Simultaneously, audio distortion should be minimized to ensure the perturbation remains imperceptible, enabling practical real-time applications such as emitting the described repeated additive noise over-the-air.

Subsection 3.1.2 provides a literature review, providing an overview of voice biometrics and SPS methods, and emphasizing the drawbacks of existing solutions. Subsection 3.1.3 provides relevant explanations, including why UAPs work, and the proposed methodology, describing the designed loss functions, preprocessing techniques, and an optimized training procedure. Namely, the well-known cosine similarity loss that ensures that the adversarial perturbation effectively alters the speaker representation $L_{\text{fooling}} = \rho(f(x), f(x+\delta))$. To compete with previous studies which incorporate ℓ_2 regularization ($L_{\ell_2} = \frac{||\hat{\delta}||_2}{l}$), novel Exponential Total Variance loss, which is inspired by the total variation (TV) loss commonly used in the image domain, is described to preserve the imperceptibility of the UAP:

$$L_{\text{Exp TV}} = \frac{1}{l}\sum_{i=0}^{l-1} \phi(\hat{\delta}_i, \hat{\delta}_{i+1}), \tag{20}$$

where δ_i UAP's amplitudes and

$$\phi(x, y) = \begin{cases} \exp\{(|y| - |x|)\} - 1 & \text{if } |y| > |x|, \\ \exp\{(|y|)\} - 1 & \text{elif } \operatorname{sign}(y) \neq \operatorname{sign}(x), \\ 0 & \text{else.} \end{cases} \tag{21}$$

Here, $l = 3200$ corresponds to the UAP $\hat{\delta}$ length (0.2 seconds at a 16 kHz sampling rate). The goal of this loss is to penalize only when the absolute values of the amplitudes increase, while avoiding penalization when they decrease, allowing the perturbation to remain minimal in less critical regions, while adapting to higher magnitudes where necessary. This point-wise approach distinguishes the proposed loss function from traditional losses, enabling a more targeted and adaptive perturbation strategy.

Table 3 — Comparison of the proposed methods with the state-of-the-art. ECAPA-TDNN model. VoxCeleb2 dataset. 20s-length audio samples.

Loss Function	FR (%)	SNR	PESQ	WER (%)
L_{ℓ_2}	74.80	17.54	2.48	94.1
$L_{\text{Exp TV}}$	75.70	19.18	2.68	52.6
Hanina S. et al.[31]	67.94	19.37	2.74	73.2

Subsection 3.1.4 describes the experimental setup, which is partially the same as that discussed in Chapter 2, including common voice biometrics models and the dataset. The training strategy proposes the use of longer audio than in previous studies and loudness normalization. The primary evaluation metric is Fooling Rate (FR), which is the percentage of audio samples for which the speaker recognition model produces incorrect predictions after applying

Table 4 — Comparison of the proposed method with the state-of-the-art on 4 speaker recognition models. LibriSpeech dataset. 10s-length audio samples. All methods (UAPs) were trained using the ECAPA-TDNN model on the LibriSpeech train dataset. “ZP” stands for zero-padding, while “RP” – for repeat padding.

ASI Model	Loss	Volume	FR (ZP)	FR (RP)	WER (%)	SNR	PESQ
ECAPA-TDNN	$L_{\text{Exp TV}}$	-23	83.9	67.1	22.2	18.52	2.29
WavLM			92.5	85.4			
ResNet			31.6	0.05			
XVecSincNet			64.2	59.5			
ECAPA-TDNN	[31]	-23	80.0	45.5	79.2	23.2	2.91
WavLM			89.2	82.6			
ResNet			32.1	0.01			
XVecSincNet			57.9	53.9			
ECAPA-TDNN	[31]	-27	89.9	73.7	94.3	19.2	2.34
WavLM			92.0	85.9			
ResNet			38.3	0.04			
XVecSincNet			62.7	56.4			

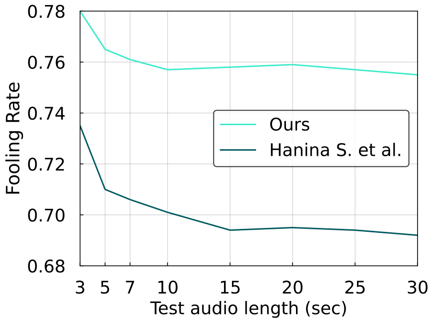


Figure 4 — Comparison of the proposed UAP (with $L_{\text{Exp TV}}$) approach performance with that of [31] across different test audio lengths.

the adversarial perturbation. Also, perceptual (PESQ, SNR) and performance metrics (Word Error Rate, WER) are evaluated on anonymized audio.

Subsection 3.1.5 compares the proposed method with the previous state-of-the-art approach and discusses the results. The main result include Tables 3, 4 and Figure 4. The obtained results demonstrate that the Exponential TV loss effectively enhances the trade-off between fooling rate and imperceptibility. It is also shown that the proposed method outperforms the competitor for various audio lengths. Embeddings Similarity Analysis of similarity between vectors of attacked and initial audio is also presented in the subsection 3.1.5. Subsection 3.1.6 concludes the presented study and proposes future research directions.

Subsection 3.2.1 briefly introduces the voice antispoofing problem as an audio binary classification, which should distinguish genuine speech (bona fide) from artificially generated (spoofs, deepfakes). This subsection also describes related works, highlighting weak generalization to rapidly developing voice generative models. The distinguishable features of the presented models, AASIST3 (or KAN-AASIST), are summarized, which is based on a spatial-temporal graph attention network AASIST [25], Kolmogorov-Arnold Networks (KAN) [34], and strong self-supervised encoders [26].

Table 5 — Final evaluation results of the submitted predictions (AASIST results were evaluated later) on the evaluation (test) set of the ASVspoof5.

condition	model	EER
closed	AASIST	27.49
closed	AASIST3	22.67
open	$\tilde{f}$	4.89

Table 6 — Additional experiments on extended and custom datasets. All values are rounded to 4 significant digits after the decimal point. EER (Equal Error Rate), ROC AUC (Area Under the Receiver Operating Characteristic Curve), Acc (accuracy), BF (bonafide). Bonafide is considered a positive class.

Model	**Train**	**Notes**	**EER**	**Precision**	**Recall**	**F1**	**AUC-ROC**	**Spoof Acc**	**BF Acc**
AASIST	№2	–	0.1073	0.8124	0.8927	0.8506	0.9346	0.9194	0.8151
W2V2-AASIST	№2	–	0.0579	0.8944	0.9421	0.9177	0.9716	0.9405	0.9526
W2V2-AASIST3	№2	–	0.0537	0.9017	0.9463	0.9235	0.9803	0.9413	0.9649
W2V2-AASIST3	№1	–	0.0218	0.9538	0.9832	0.9683	0.9942	0.9752	0.9832
W2V2-AASIST3	№1	Averaging	**0.0199**	**0.9583**	**0.9847**	**0.9713**	**0.9955**	**0.9777**	**0.9847**

Subsection 3.2.2 provides preliminary information about the Kolmogorov–Arnold theorem and KAN layers. Subsection 3.2.3 describes the models' main blocks, including feature extractors, encoders, homogeneous and heterogeneous graph attention layers, graph pooling, stacking, and classification operations.

Subsection 3.2.4 describes experimental settings, the utilized datasets, augmentations, and loss functions. A pre-emphasis, which is a filtering operation that attenuates low-frequency components while amplifying high-frequency ones, is discussed as helpful. It was applied to the raw audio amplitudes of initial audio as $x_l = x_l - 0.97 \cdot x_{l-1}$, where $l \in \{1, 2, \ldots, L\}$ and L denotes the total length of the audio signal. The submission details, training, and inference tricks for the ASVspoof5 [23] contest are discussed. Two conditions (contest's tracks) were considered. In a closed condition, it is prohibited to use pre-trained models or external data, in contrast to the open condition. The corresponding submission results and several ablation results are presented in Tables 5, 6. The primary metric is equal-error-rate (EER), which is a decision threshold regarding the ASI or classification task for which the model's false acceptance and false rejection rates are equal. $\tilde{f}$ averages predictions of W2V2-AASIST3, and W2V2-AASIST-KAN, that utilize Wav2Vec2 features for the frontend.

Subsection 3.2.5 discusses limitations, including a general problem in voice antispoofing that, although the proposed models generalize better than a strong competitor, they still fail a lot. Despite new, stronger models, loss functions, or augmentations, a straightforward dataset expansion and re-training are also required, which is illustrated in Table 6. Train set 1 extends train set 2, with several additional datasets, and the models were evaluated on an internal test set. One can observe that, having the same augmentations and loss functions, the main contributions to the improved performance come from SSL encoders and more data. Subsection 3.2.5 concludes the presented study and proposes future work directions.

Conclusion

This dissertation investigated robustness, trustworthiness, and privacy methods for deep learning image and audio classification models. The primary outcome is the development of novel certified and empirical approaches that enhance the robustness and reliability of the models. The following results were obtained:

1. A universal certification approach against resolvable compositions of semantic transformations was proposed with a corresponding semi-automatic numeric procedure, enabling efficient robustness verification and achieving state-of-the-art guarantees across multiple datasets.
2. An improved certification method for few-shot embedding models against additive norm-bounded perturbations was developed, providing the first provable robustness guarantees for speaker identification and surpassing existing baselines.
3. A novel universal adversarial perturbation framework for speaker privacy was introduced, suitable for long audio and based on incorporating the exponential total variance loss during training and length-agnostic tiling during training and inference. The method outperformed the competitor in terms of fooling and word error rates, while demonstrating comparable perceptual metrics.
4. An enhanced voice anti-spoofing architecture was designed using Kolmogorov–Arnold network layers, advanced audio pre-processing and augmentations, and a self-supervised backbone, resulting in a more robust performance for new test data.

Publications of the author on the subject of the dissertation

A1 General Lipschitz: Certified robustness against resolvable semantic transformations via transformation-dependent randomized smoothing [Text] / D. Korzh [et al.] // ECAI 2024. — IOS Press, 2024. — P. 1591—1598.

A2 Certification of speaker recognition models to additive perturbations [Text] / D. Korzh [et al.] // Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 39. — 2025. — P. 17947—17956.

A3 Novel Loss-Enhanced Universal Adversarial Patches for Sustainable Speaker Privacy [Text] / E. Karimov [et al.] // Interspeech 2025. — 2025. — P. 1513—1517.

A4 AASIST3: KAN-enhanced AASIST speech deepfake detection using SSL features and additional regularization for the ASVspoof 2024 Challenge [Text] / K. Borodin [et al.] // Proc. ASVspoof 2024. — 2024. — P. 48—55.

A5 Clarispeech: LLM-Enhanced Speech Recognition Post-Correction [Text] / A. Iudin [et al.] // AINL 2025. — 2025.

A6 CLEAR: Character Unlearning in Textual and Visual Modalities [Text] / A. Dontsov [et al.] // Findings of the Association for Computational Linguistics. - 2025. — P. 20582—20603.

References

1. *Khan*, *A. I.* CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images [Text] / A. I. Khan, J. L. Shah, M. M. Bhat // Computer methods and programs in biomedicine. — 2020. — Vol. 196. — P. 105581.

2. *Desplanques*, *B.* Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification [Text] / B. Desplanques, J. Thienpondt, K. Demuynck // arXiv preprint arXiv:2005.07143. — 2020.

3. Conformer: Convolution-augmented transformer for speech recognition [Text] / A. Gulati [et al.] // arXiv preprint arXiv:2005.08100. — 2020.

4. Analyzing and improving the image quality of stylegan [Text] / T. Karras [et al.] // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2020. — P. 8110—8119.

5. The llama 3 herd of models [Text] / A. Dubey [et al.] // arXiv preprint arXiv:2407.21783. — 2024.

6. *Szegedy*, *C.* Intriguing properties of neural networks [Text] / C. Szegedy // arXiv preprint arXiv:1312.6199. — 2013.

7. *Goodfellow*, *I. J.* Explaining and harnessing adversarial examples [Text] / I. J. Goodfellow, J. Shlens, C. Szegedy // arXiv preprint arXiv:1412.6572. — 2014.
8. Xtts: a massively multilingual zero-shot text-to-speech model [Text] / E. Casanova [et al.] // arXiv preprint arXiv:2406.04904. — 2024.
9. A survey on adversarial attacks and defences [Text] / A. Chakraborty [et al.] // CAAI Transactions on Intelligence Technology. — 2021. — Vol. 6, no. 1. — P. 25—45.
10. *Madry*, *A.* Towards deep learning models resistant to adversarial attacks [Text] / A. Madry // arXiv preprint arXiv:1706.06083. — 2017.
11. Square attack: a query-efficient black-box adversarial attack via random search [Text] / M. Andriushchenko [et al.] // European conference on computer vision. — Springer. 2020. — P. 484—501.
12. Dpatch: An adversarial patch attack on object detectors [Text] / X. Liu [et al.] // arXiv preprint arXiv:1806.02299. — 2018.
13. Physical Adversarial Camouflage through Gradient Calibration and Regularization [Text] / J. Liang [et al.] // arXiv preprint arXiv:2508.05414. — 2025.
14. *Wong*, *E.* Fast is better than free: Revisiting adversarial training [Text] / E. Wong, L. Rice, J. Z. Kolter // arXiv preprint arXiv:2001.03994. — 2020.
15. *Athalye*, *A.* Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples [Text] / A. Athalye, N. Carlini, D. Wagner // International conference on machine learning. — PMLR. 2018. — P. 274—283.
16. On the effectiveness of interval bound propagation for training verifiably robust models [Text] / S. Gowal [et al.] // arXiv preprint arXiv:1810.12715. — 2018.
17. *Li*, *L.* Sok: Certified robustness for deep neural networks [Text] / L. Li, T. Xie, B. Li // 2023 IEEE symposium on security and privacy (SP). — IEEE. 2023. — P. 1289—1310.
18. *Cohen*, *J.* Certified adversarial robustness via randomized smoothing [Text] / J. Cohen, E. Rosenfeld, Z. Kolter // International Conference on Machine Learning. — PMLR. 2019. — P. 1310—1320.
19. Provably robust deep learning via adversarially trained smoothed classifiers [Text] / H. Salman [et al.] // Advances in Neural Information Processing Systems. — 2019. — Vol. 32.
20. *Muravev*, *N.* Certified robustness via randomized smoothing over multiplicative parameters of input transformations [Text] / N. Muravev, A. Petiushko // arXiv preprint arXiv:2106.14432. — 2021.

21. Tss: Transformation-specific smoothing for robustness certification [Text] / L. Li [et al.] // Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. — 2021. — P. 535—557.

22. Smoothed embeddings for certified few-shot learning [Text] / M. Pautov [et al.] // Advances in Neural Information Processing Systems. — 2022. — Vol. 35. — P. 24367—24379.

23. ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale [Text] / X. Wang [et al.] // arXiv preprint arXiv:2408.08739. — 2024.

24. Add 2023: the second audio deepfake detection challenge [Text] / J. Yi [et al.] // arXiv preprint arXiv:2305.13774. — 2023.

25. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks [Text] / J.-w. Jung [et al.] // ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). — IEEE. 2022. — P. 6367—6371.

26. wav2vec 2.0: A framework for self-supervised learning of speech representations [Text] / A. Baevski [et al.] // Advances in neural information processing systems. — 2020. — Vol. 33. — P. 12449—12460.

27. Wavlm: Large-scale self-supervised pre-training for full stack speech processing [Text] / S. Chen [et al.] // IEEE Journal of Selected Topics in Signal Processing. — 2022. — Vol. 16, no. 6. — P. 1505—1518.

28. Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing [Text] / H. Tak [et al.] // ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — IEEE. 2022. — P. 6382—6386.

29. *Zhang*, *Y.* One-class learning towards synthetic voice spoofing detection [Text] / Y. Zhang, F. Jiang, Z. Duan // IEEE Signal Processing Letters. — 2021. — Vol. 28. — P. 937—941.

30. {V-Cloak}: Intelligibility-, Naturalness-& {Timbre-Preserving}{Real-Time} Voice Anonymization [Text] / J. Deng [et al.] // 32nd USENIX Security Symposium (USENIX Security 23). — 2023. — P. 5181—5198.

31. Universal adversarial attack against speaker recognition models [Text] / S. Hanina [et al.] // ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — IEEE. 2024. — P. 4860—4864.

32. GSmooth: Certified Robustness against Semantic Transformations via Generalized Randomized Smoothing [Text] / Z. Hao [et al.] // International Conference on Machine Learning. — PMLR. 2022. — P. 8465—8483.

33. *Clopper*, *C. J.* The use of confidence or fiducial limits illustrated in the case of the binomial [Text] / C. J. Clopper, E. S. Pearson // Biometrika. — 1934. — Vol. 26, no. 4. — P. 404—413.

34. KAN: Kolmogorov-Arnold Networks [Text] / Z. Liu [et al.]. — 2024. — arXiv: `2404.19756 [cs.LG]`. — URL: `https://arxiv.org/abs/2404.19756`.

Корж Дмитрий Сергеевич

Устойчивость моделей глубокого обучения

Автореф. дис. на соискание ученой степени канд. физ-мат. наук

Подписано в печать ___.___._____. Заказ № ____________
Формат 60×90/16. Усл. печ. л. 1. Тираж 50 экз.
Типография _______________________________________