UDC 004.056.55
DOI: 10.15587/2706-5448.2026.352336

Ivan Vynogradov

ASSESSMENT OF QUALITY DEGRADATION IN MODERN VOICE DEEPFAKE DETECTORS UNDER CROSS-LINGUISTIC SHIFT FROM ENGLISH TO UKRAINIAN

The research object is the processes and algorithms of automated discrimination between real and synthesized speech (anti-spoofing systems) when they function in conditions of a pronounced linguistic shift. The research solves the scientific and practical issue of quantitatively assessing the critical degradation of the precision of modern neurolinguistic detectors, using the example of the AASIST architecture with graph attention, when they encounter high-quality voice attacks in the Ukrainian. Special attention was paid to attacks formed using industrial new-generation neural vocoders, which are practically not represented in classic English training samples.

The essence of the obtained results lies in establishing and mathematically confirming the existence of a significant "generalization gap" in cross-language testing. It was experimentally proven that the transition from the English acoustic domain into the Ukrainian domain causes the growth of the equal error rate (EER) coefficient by 2.5–3.5 times. In the most advanced synthesis systems, the EER reached a critical threshold of 25.64%, which indicates the loss of the system's protective capabilities in this language domain.

These results were obtained through the usage of an experimental stand, which unites the AASIST model and closed-end commercial APIs of neural speech cloning. Unlike standard tests on archive databases, the suggested approach, using a specially EXT dataset that includes five independent attack groups, allowed for modeling real cyber threat scenarios.

In practice, these results can be used in the design of voice biometric authentication systems in the banking and governmental sectors of Ukraine and justify the mandatory necessity of linguistic adaptation and deep fine-tuning of classifiers using localized datasets to achieve the required level of information security.

Keywords: *anti-spoofing, voice deepfakes, voice cloning, linguistic shift, biometric authentication.*

Received: 22.11.2025
Received in revised form: 21.01.2026
Accepted: 11.02.2026
Published: 28.02.2026

How to cite

Vynogradov, I. (2026). Assessment of quality degradation in modern voice deepfake detectors under cross-linguistic shift from English to Ukrainian. Technology Audit and Production Reserves, 1 (2 (87)), 6–10. https://doi.org/10.15587/2706-5448.2026.352336

1. Introduction

Over the past years, the text-to-speech (TTS) and voice conversion (VC) technologies, based on deep learning, approached natural human speech in their quality. Broad availability of these tools caused the growth of abuses, ranging from synthetic identity fraud to large-scale financial fraud. The situation is aggravated by the vulnerability of modern detectors to new-generation neural network-based attacks, such as diffusion models and transformers [1]. In these cases, anti-spoofing becomes a critically important element of informational security systems, ranging from banking authentication to protection of government communication channels.

Despite significant progress, the overwhelming majority of studies only focus on the English language, which creates a gap between laboratory studies and real applications in a multilingual environment. The issue of generalization of detectors to other languages, particularly Ukrainian, is not sufficiently studied, which causes hidden risks for local cybersecurity systems [2].

International competitions, such as ASVspoof and ADD Challenge, defined test protocols and base architectures, where the AASIST model is separately noteworthy. Modern solutions are typically based on a combination of spectral signs and end-to-end architectures. However, the dominance of English-language data during the model training creates a linguistic bias: there is an open question of how these models behave when the phonetics and prosody of the Ukrainian language change. A number of studies show a quality drop in the domain shift. However, the absence of rigid benchmarks for the Ukrainian language does not allow for assessing the real scale of that vulnerability [3, 4]. Recent studies emphasize that the lack of localized datasets leads to a "phonetic bias" which significantly degrades the performance of detectors when applied to another language structures [5]. Furthermore, advancements in latent diffusion vocoders have demonstrated the ability to bypass traditional anti-spoofing filters, need the evaluation of architectures like AASIST in cross-lingual scenarios [6]. Current research highlights that integrating self-supervised learning (SSL) front-ends could potentially mitigate these domain shifts, though it's good for the Ukrainian language remains an open scientific challenge [7].

The research object is the processes and algorithms of automated discrimination between real and synthesized speech (anti-spoofing systems) under conditions of a linguistic shift. *The aim of the research* is to quantitatively assess the degradation of the quality of modern voice deepfake detectors when the language is changed from English to Ukrainian and to evaluate their robustness against attacks generated using state-of-the-art (SOTA) voice synthesis systems. To achieve this aim, the following research objectives were set:

– to construct the experimental EXT UA dataset based on real Ukrainian speech and synthetic voice attacks generated using commercial systems such as ElevenLabs and Resemble AI;

– to formalize a mathematical apparatus for evaluating the detector's reliability based on the statistical distribution of FAR/FRR and the equal error rate (*EER*) under domain shift conditions;

– to analyze changes in the equal error rate (*EER*) metric and the FAR/FRR probabilities when transitioning from the English to the Ukrainian acoustic domain.

2. Materials and Methods

The modern state of the anti-spoofing industry can be characterized by the development of several technological lines, each demonstrating high efficacy in the closed English-language benchmarks. The analysis of the scientific periodicals of the last five years (Scopus, ScienceDirect resources) allows it to define seven dominating directions of the detector's development, the characteristics of which are provided in Table 1. The first direction is based on classic spectral descriptors, such as LFCC and CQCC [8], combined with compact LCNN architectures [9]. The second direction encompasses end-to-end architectures with learnable front-ends, operating on waveform-level representations or spectro-temporal graphs, such as RawNet2 [10] and RawGAT-ST [11]. The third direction is linked to the use of integrated AASIST systems [12] and multilingual SSL encoders XLS-R [13, 14].

Despite the availability of representative lines data, a number of significant gaps are found in the analyzed samples. Firstly, the issue of linguistic universality remains unsolved: the overwhelming majority of models are optimized only for English-language phonetics. Secondly, there is a deficit of studies on the robustness of academic systems in the conditions of dynamically developing commercial SOTA generators. Table 2 shows a comparative analysis of five leading industrial solutions for voice synthesis and cloning that support the Ukrainian language [15–21].

The reason for the lack of such data is the deficit of profile Ukrainian datasets and the complexity of the integration of the closed-end commercial APIs into the research pipelines. Systematization of these factors makes it possible to formulate the open scientific issue: the absence of qualitative assessments of the robustness of AASIST-like architectures in the conditions of simultaneous influence of linguistic shift and new-generation attacks.

To address the issue in this research, it is possible to select two models from the available variety (Tables 1 and 2). The selection represents polar approaches (LFCC + AASIST and XLS-R + AASIST) and two attack sources (ElevenLabs and Resemble AI). This creates conditions for a controlled stress test, which are as close to real threats in the Ukrainian cyber space segment as possible.

The formulated issue and the choice of tools define the research objective, which is described in detail in the next chapter of research: making a quantitative analysis of the degradation of the selected detectors when they encounter synthetic attacks in the Ukrainian-language domain.

To ensure the validity and exclude the effect of "vendor-specific" artifacts, all experiments were conducted in a single environment, using two independent speech synthesis platforms [22, 23]. Calculations, including pre-processing and inference of the neural network detector, were conducted on a server powered by Linux OS (Ubuntu), equipped with NVIDIA RTX 4000 SFF Ada Generation video card. The software environment is based on Python 3.10 programming language and the PyTorch framework. Librosa and SoundFile libraries were used for Digital Signal Processing (DSP).

The standard implementation of the AASIST architecture was used as a basic detector [24] that demonstrates the SOTA results on standard benchmarks but requires verification of robustness in case of new kinds of attacks [25]. The model was run in eval mode with weights pretrained on the ASVspoof 2019 Logical Access (LA) dataset. To match the frequency characteristics of modern generators (44.1 kHz) and detectors (16 kHz), a strict normalization pipeline was used that includes the forced re-discretization by the algorithm with anti-aliasing and tensor length fixation on the level of 64,600 samples (~4.03 s), which is critically important for detection accuracy [26].

Table 1

Comparative characteristics of the studied anti-spoofing architectures

ID	Model	Family	Country	Representativeness	Reproducibility	License
D1	LFCC + GMM	Signature baseline	France/UK	Historic reference point ASVspoof	High	Permissive
D2	CQCC + GMM	Signature baseline	France	Classic feature-based approach	High	Permissive
D3	LFCC + LCNN	Signs + CNN	USA/China	Light DNN reference point based on signs	High	Permissive
D4	RawNet2	End-to-end (raw wave)	South Korea	Basic e2e reference point	Medium	Permissive
D5	RawGAT-ST	End-to-end (GAT wave)	Japan/China	Strong e2e reference	Medium	Permissive
D6	AASIST	Integrated GAT on signs	France/South Korea	De facto the standard of open anti-spoofing	High	Permissive
D7	XLS-R/wav2vec 2.0 + AASIST	SSL frontend + detector	USA	Multilingual frontend, upper limit of tolerance	Medium-high	Permissive

Table 2

Comparative analysis of commercial systems of voice synthesis for stress-testing

ID	System	UA language support	Cloning	Streaming/RT	API/SDK	Reasons for inclusion
G1	ElevenLabs	Yes	Yes (zero/few-shot)	Yes	Yes	Industrial quality, multilingual support, stable API
G2	Resemble AI	Yes	Yes	Yes	Yes	Cross-lingual localization, independence from the academic sets
G3	Microsoft Azure Neural TTS	Yes (uk-UA)	Custom Neural Voice	Yes	Yes	Corporate standard, containers/on-prem
G4	Google cloud TTS	Yes (uk-UA)	Custom voices (limited)	Yes	Yes	Cloud standard, rich formats
G5	Speechify TTS API	Yes	Yes	Yes	Yes	Alternative commercial API

3. Results and Discussion

3.1. Construction of the EXT UA dataset for Ukrainian speech anti-spoofing

The experiment scheme is designed to estimate the robustness of a detector to language domain shift and check the hypothesis of the ability of modern Ukrainian-language synthesis methods to bypass the protection, which was trained on English-language datasets [27, 28]. Scripts interacting with APIs of ElevenLabs (model eleven_multilingual_v2) and Resemble AI (Neural Audio Engine) platforms were developed to generate attacks. The use of two different architectures was necessary to prove the systemic nature of the Generalization Gap issue.

The structure of the formed textual set is presented in Table 3. To exclude the retraining effect based on the text, all groups used the single phrase corpus. The real speech set EXT_REAL_UA is formed in the FLAC 16 kHz (mono) format to avoid the compression losses that distort metrics [29]. The authentic speech set (EXT_REAL_UA) consists of 58 recordings collected from 4 different speakers (2 males and 2 females) to ensure gender balance and minimize speaker-dependent bias.

The attack generation process in the ELV and RES categories was done using modern proprietary algorithms based on transformers and diffusion models that minimize the number of spectral artifacts. For the ElevenLabs platform, it is possible to use the v2 multilingual model that works with a fixed voice profile, with the levels of parameters Stability set to 0.5 and Similarity Boost set to 0.75. These settings provide the balance between the naturalness of intonations and a high level of acoustic similarity to the original. In parallel, it is possible to use the zero-shot cloning method for Resemble AI datasets, when the neuron driver of the synthesis builds its acoustic representation of voice on a short reference fragment of a real utterance.

Using two principally different cloud architectures makes it possible to prove the systemic nature of detector degradation, eliminating the possibility of accidental incompatibility with a specific vendor. All generated audio materials passed through the stage of automated volume normalization and forced re-discretization to 16 kHz, which ensured their compatibility with the input format of the AASIST detector.

To provide the representativeness of the sample and to exclude the semantic context, a unified textual corpus was developed that includes four stylistic categories. The corpus includes everyday phrases and greetings, specific terminology of banking identification systems, narrative sentences for assessing prosodic stability, and short commands for the voice command systems. Phrase adaptation for Ukrainian and English language groups was made with the preservation of identical meaning and emotional hue, which allows making the correct comparative analysis of the influence of linguistic shift.

Table 3

Experimental results of detector robustness evaluation under linguistic domain shift

Dataset	Language/Generator	Volume	Mean Pspoof (%)	Interpretation
REAL UA	Real/UA	58	31.2%	Model suspects real voice (ideally would be <5%)
ELV_UA_SPOOF	ElevenLabs/UA	100	45.5%	Peak is lower than the threshold (50%), the attack successfully gets through
RES_UA_SPOOF	Resemble AI/UA	99	63.5%	Peak is higher than the threshold, the attack is less successful
ELV_EN_SPOOF	ElevenLabs/EN	103	38.1%	Model lets fakes through even in the native EN language
RES_EN_SPOOF	Resemble AI/EN	106	60.9%	Artifacts are more evident than in ElevenLabs

3.2. Mathematical apparatus of evaluation

A posteriori probability of the fact that an incoming audio signal is an attack was used as a primary metric. This value is calculated based on the output logits of the model (values of the output layer before classification) using the Softmax function, and is defined by the following expression

$$P_{spoof}\left(x\right)=\frac{e^{s_{fake}}}{e^{s_{bonafide}}+e^{s_{fake}}}\cdot 100\%, \qquad (1)$$

where $P_{spoof}\left(x\right)$ – a posteriori probability of the fact that the audio signal is synthetic; s_{fake} – the output logit of the model for the spoof ("attack") class; $s_{bonafide}$ – the output logit of the model for the real speech ("bona fide") class. For the integral evaluation of the system under conditions of intersection of distributions, the coefficient of equal error rate (*EER*) is calculated, which corresponds to the point where the probability of the false acceptance rate (*FAR*) becomes equal to the false rejection rate (*FRR*)

$$EER=FAR=FRR. \qquad (2)$$

Classification of the incoming signal is based on the underlying rule

$$y\hat{}=\{1,\text{ if }P_{spoof}\left(x\right)\geq\theta;\ 0,\text{ if }P_{spoof}\left(x\right)<\theta, \qquad (3)$$

where $\hat{y}=1$ corresponds to the attack detection, and is the θ classification threshold (within this research, $\theta=0.5$).

3.3. Analysis of EER and FAR/FRR metrics transition from the English to the Ukrainian acoustic domain

The comparative assessment of the precision of the AASIST detector based on five formulated datasets was made during the test. Key indicators of the system efficacy, expressed through the coefficient of equal error rate (*EER*), are systematized in Table 4.

Table 4

Results of the AASIST detector test based on the EXT datasets

Dataset identifier	Language	Generator	*EER*, %
ELV_EN_SPOOF	EN	ElevenLabs	7.18
RES_EN_SPOOF	EN	Resemble AI	6.42
ELV_UA_SPOOF	UA	ElevenLabs	31.88
RES_UA_SPOOF	UA	Resemble AI	31.17

Data analysis shows critical degradation of the detector's precision when shifting from English to the Ukrainian language. For the *ELV_UA_SPOOF* dataset, the *EER* value reached 31.88%, which is more than four times the indicator of the basic English dataset. Analogical tendency is observed also for the Resemble AI platform, where the volume of errors grew from 6.42% to 31.17%.

Visualization of the density of a posteriori probability distribution $P_{spoof}\left(x\right)$ (Fig. 1) explains the physical nature of such degradation. The graphs of the Ukrainian domain show a significant shift and the widening of the "tails" of distributions, both for real speech and attacks. This causes the loss of the clear divisibility of X_{real} and X_{spoof} classes and the formation of a wide uncertainty area.

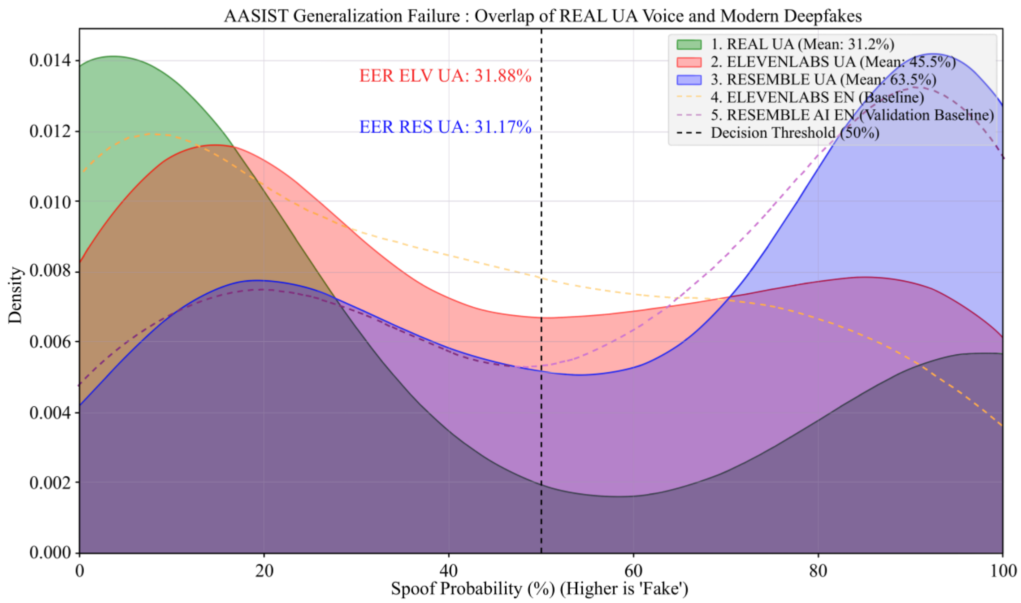

Fig. 1. Score distributions of the AASIST model for genuine Ukrainian speech and synthetic attacks generated by ElevenLabs and Resemble AI

The phenomenon of "failure" of the model in the Ukrainian language using ElevenLabs (the worst result, 31.89%) proves the hypothesis that modern multilingual synthesis models effectively imitate the prosodic peculiarities of the Ukrainian language, which a detector trained on English phonetics erroneously interprets as natural. Thus, the existence of a significant generalization gap is proven, caused by a linguistic shift.

Visual representation in Fig. 1 proves that acoustic signs of Ukrainian phonemes, when they are processed by the model optimized for the English domain, lead to significant overlap in definitions of real and synthesized speech. This overlap demonstrates the zone of statistical uncertainty, where the AASIST model is incapable of making the correct decision, which directly leads to the observed *EER* degradation.

3.4. Limitations and future directions for the research

Practical meaning: The received results become a quantitative ground for a technological audit of the existing systems of voice biometry used in the Ukrainian IT market segment. The estimated error threshold (31.88% for ElevenLabs) is a strict reason for the need for mandatory linguistic fine-tuning of anti-spoofing software. These data can be directly used by the developers of banking systems of authentication and state digital platforms to improve the security protocols to fight high-quality voice cloning.

Research limits: The main limit of this research is the focus on only one, although the most advanced, AASIST architecture. Despite the fact that the AASIST architecture is the reference industry standard, the obtained results can somewhat differ from other end-to-end models. Also, high-quality (made in a studio environment) voice uttering was used in the research, so the influence of the background surrounding noise and channel distortions (such as telephone lines) on the cross-language robustness requires additional verification.

Perspectives of future researches: The future researches shall be focused on the development of methods of "cross-language fine-tuning" of models, where the detector is additionally trained on small localized batches of data to minimize the generalization gap. Another perspective direction is the research of hybrid architectures that unite the analysis of acoustic attacks with linguistic and semantic inspection to improve the detection stability independently of the language domain.

4. Conclusions

1. The specialized experimental dataset EXT UA was formed, which includes the specimens of authentic Ukrainian language and synthesized attacks. Usage of ElevenLabs and Resemble AI advanced commercial platforms made it possible to create a representative sample of deepfakes necessary for a valid assessment of the influence of phonetic and prosodic features of the Ukrainian language on the functioning of detectors.
2. A mathematical apparatus for evaluating the detector's reliability was formalized, based on the statistical distribution of *FAR*/*FRR* and the equal error rate (*EER*). This theoretical foundation allowed for a precise calculation of the *EER* shift and provided a methodology for analyzing the degradation of class separability under domain mismatch conditions.
3. The quantitative analysis of the *EER* metric and *FAR*/*FRR* probabilities during the shift to the Ukrainian acoustic domain was made. Error rate level critical increase was established: the *EER* value reached its peak of 31.88% for attacks in ElevenLabs and 31.17% in Resemble AI, which is more than three times the level of values on the basic English language. Analysis of the distribution density graphs confirmed the loss of separability of classes, which supports the necessity of localized fine-tuning of anti-spoofing systems to ensure the cybersecurity of the Ukrainian IT market segment.

Conflict of interest

The author declares that he has no conflict of interest regarding this research, including financial, personal, authorship, or other, that could influence the research and its results presented in this article.

Financing

The research was conducted without financial support.

Data availability

All key findings of research (calculated Spoof Probability points and *EER* metrics) are included in the additional materials of the research. To ensure the full reproducibility of the experiment, the entire software code used for the generation of synthetic datasets (via API of ElevenLabs and Resemble AI), calculation of *EER* metrics, and visualization of results ($P_{spoof}(x)$ distribution graph) is in open access of the GitHub repository [30].

Note: Raw audio files, generated by commercial platforms (ElevenLabs, Resemble AI) and real voice records (REAL_UA) cannot be provided in open access due to restrictions of license agreements and data confidentiality.

Use of artificial intelligence

The author confirms that in order to ensure the transparency of the experimental part of the research, the Google Gemini generative artificial intelligence was used. AI assistance was received in Chapter 3 Results and Discussion to create and debug the Python scripts (.py). These scripts were necessary for the automation of interaction with APIs of commercial platforms of speech synthesis (ElevenLabs and Resemble AI) with the aim of generating test audio files and their further preparation for scanning by the AASIST detector. The results obtained using these scripts were validated by the author by checking the functionality and output data formats. Using AI did not affect the scientific conclusions of the research since the interpretation and analysis of the final metrics (*EER*/*FAR*) were made by the author independently.

Authors' contributions

The sole author, ***Ivan Vynogradov***, is responsible for the entire content of the manuscript, including the conceptualization, methodology design, data collection, experimental execution, formal analysis, results interpretation, writing the original draft, and final manuscript review and editing.

References

1. Rabhi, M., Bakiras, S., Di Pietro, R. (2024). Audio-deepfake detection: Adversarial attacks and countermeasures. *Expert Systems with Applications, 250,* 123941. https://doi.org/10.1016/j.eswa.2024.123941
2. Vynogradov, I. (2025). Voice fake detection: modern techniques and applications for Ukrainian language. *Measuring and computing devices in technological processes, 82 (2),* 31–36. https://doi.org/10.31891/2219-9365-2025-82-5
3. Marek, B., Kawa, P., Syga, P. (2024). *Are audio DeepFake detection models polyglots?* arXiv preprint. https://doi.org/10.48550/arXiv.2412.17924
4. Liu, T., Kukanov, I., Pan, Z., Wang, Q., Sailor, H. B., Lee, K. A. (2024). Towards Quantifying and Reducing Language Mismatch Effects in Cross-Lingual Speech Anti-Spoofing. *2024 IEEE Spoken Language Technology Workshop (SLT),* 1185–1192. https://doi.org/10.1109/slt61566.2024.10832142
5. Moreno, V., Lima, J., Simões, F., Violato, R., Neto, M. U., Runstein, F., Costa, P. (2025). Revealing Cross-Lingual Bias in Synthetic Speech Detection under Controlled Conditions. *5th Symposium on Security and Privacy in Speech Communication,* 1–7. https://doi.org/10.21437/spsc.2025-1
6. Kong, J., Kim, J., Bae, J. H. (2020). HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *Advances in Neural Information Processing Systems (NeurIPS),* 33. Available at: https://doi.org/arXiv:2010.05646
7. Wang, X., Delgado, H., Tak, H., Jung, J., Shim, H., Todisco, M. et al. (2024). ASVspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale. *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024),* 1–8. https://doi.org/10.21437/asvspoof.2024-1
8. Delgado, H., Evans, N., Kinnunen, T., Lee, K. A., Liu, X., Nautsch, A. et al. (2021). *ASVspoof 2021 Evaluation Plan.* arXiv preprint. Available at: https://www.asvspoof.org/asvspoof2021/asvspoof2021_evaluation_plan.pdf
9. Todisco, M., Delgado, H., Evans, N. (2017). Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language, 45,* 516–535. https://doi.org/10.1016/j.csl.2017.01.001
10. Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N., Larcher, A. (2021). End-to-End anti-spoofing with RawNet2. *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 6369–6373. https://doi.org/10.1109/icassp39728.2021.9414234
11. Tak, H., Jung, J., Patino, J., Kamble, M., Todisco, M., Evans, N. (2021). End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge,* 1–8. https://doi.org/10.21437/asvspoof.2021-1
12. Jung, J., Heo, H.-S., Tak, H., Shim, H., Chung, J. S., Lee, B.-J. et al. (2022). AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks. *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 6367–6371. https://doi.org/10.1109/icassp43922.2022.9747766
13. Tak, H., Todisco, M., Wang, X., Jung, J., Yamagishi, J., Evans, N. (2022). Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation. *The Speaker and Language Recognition Workshop (Odyssey 2022),* 112–119. https://doi.org/10.21437/odyssey.2022-16
14. Zhang, Q., Wen, S., Hu, T. (2024). Audio Deepfake Detection with Self-Supervised XLS-R and SLS Classifier. *Proceedings of the 32nd ACM International Conference on Multimedia,* 6765–6773. https://doi.org/10.1145/3664647.3681345
15. Models. *ElevenLabs.* Available at: https://elevenlabs.io/docs/models
16. Dubbing. *ElevenLabs.* Available at: https://elevenlabs.io/docs/capabilities/dubbing
17. Realtime Text-to-Speech AI Voice Generator built for Voice Agents. *Resemble AI.* Available at: https://www.resemble.ai/text-to-speech-converter/
18. Ukrainian Text-to-Speech and AI Voice Generator. *Resemble AI.* Available at: https://www.resemble.ai/ukrainian-tts/
19. What's new in Azure Speech in Foundry Tools? *Microsoft Learn.* Available at: https://learn.microsoft.com/azure/ai-services/speech-service/releasenotes
20. Gemini-TTS. Chirp 3 HD – Supported languages (uk-UA). *Google Cloud.* Available at: https://cloud.google.com/text-to-speech/docs/gemini-tts
21. Language Support – Languages supported by Speechify Text-to-Speech API. *Speechify.* Available at: https://docs.sws.speechify.com/docs/features/language-support
22. Bringing technology to life. *ElevenLabs.* Available at: https://elevenlabs.io
23. Bringing technology to life. *Resemble AI.* Available at: https://www.resemble.ai
24. Kinnunen, T., Lee, K. A., Delgado, H., Evans, N., Todisco, M., Sahidullah, M. et al. (2018). t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification. *The Speaker and Language Recognition Workshop (Odyssey 2018),* 312–319. https://doi.org/10.21437/odyssey.2018-44
25. Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N. et al. (2020). ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language, 64,* 101114. https://doi.org/10.1016/j.csl.2020.101114
26. Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C. Y., Zhao, Y. (2023). Audio Deepfake Detection: A Survey. *Journal of Latex Class Files, 14 (8).* https://doi.org/10.48550/arXiv.2308.14970
27. Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A. et al. (2021). ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge,* 47–54. https://doi.org/10.21437/asvspoof.2021-8
28. Müller, N. M., Czempin, P., Dieckmann, J., Froghyar, A., Böttinger, K. (2022). *Does Audio Deepfake Detection Generalize?* https://doi.org/10.48550/arXiv.2203.16263
29. Liu, X., Wang, X., Sahidullah, M., Patino, J., Delgado, H., Kinnunen, T. et al. (2023). ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31,* 2507–2522. https://doi.org/10.1109/taslp.2023.3285283
30. Voicefakedetector. *GitHub repository.* Available at: https://github.com/ipvinner/voicefakedetector Last accessed: 22.12.2025

Ivan Vynogradov, *PhD Student, Department of Cyber Security and Technical Information Protection, State University of Intelligent Technologies and Telecommunications, Odesa, Ukraine, e-mail: ipvinner@gmail.com, ORCID: https://orcid.org/0009-0000-9901-7811*