https://doi.org/10.31891/2219-9365-2025-82-5
UDC 004.056

VYNOGRADOV Ivan
State University of Intelligent Technologies and Telecommunications
https://orcid.org/0009-0000-9901-7811
e-mail: ipvinner@gmail.com

VOICE FAKE DETECTION: MODERN TECHNIQUES AND APPLICATIONS FOR UKRAINIAN LANGUAGE

The subject matter of this article is the detection of fake voices generated by text-to-speech (TTS) synthesis and voice conversion (VC) technologies, with a focus on their application to the Ukrainian language. The goal is to analyze modern datasets, competitions (ASVspoof, ADD Challenge), and detection algorithms to assess the feasibility of integrating Ukrainian data into international frameworks or developing a dedicated dataset. This approach addresses not only the shortage of Ukrainian-language recordings in widely used repositories—many of which are limited to English or Chinese—but also the unique phonetic structures, diverse accents, and morphological complexities inherent to Ukrainian. By comparing performance across multiple spoofing scenarios, researchers can more accurately quantify how language-specific features influence classification accuracy, ultimately informing more robust detection frameworks. The tasks solved in the article: to examine existing datasets and their suitability for Ukrainian, evaluate the performance of fake voice detection systems using Equal Error Rate (EER), Weighted EER (WEER), and Detection Success Rate (DSR), and determine the best approach—expanding ASVspoof or creating a new resource. The methods used include systematic analysis, dataset comparison, and performance evaluation of modern synthesis systems like ElevenLabs, Assembly AI, and Tacotron. The results show that adapting fake voice detection systems to the Ukrainian language enhances accuracy and robustness. Moreover, targeted inclusion of different regional dialects and speaker profiles emerges as a key factor in maintaining high Detection Success Rate (DSR) values. The findings highlight that advanced neural vocoders, which replicate fine-grained prosodic and timbral nuances, necessitate specialized countermeasures able to discern subtle synthetic artifacts. Consequently, the study underscores the importance of iterative dataset refinement, periodic algorithmic updates, and cross-lingual benchmarking to sustain robust performance against evolving voice spoofing threats. Conclusions. The study confirms that integrating Ukrainian-language data into international datasets or developing a specialized dataset significantly improves detection reliability. The scientific novelty lies in: 1) the first systematic analysis of Ukrainian fake voice detection; 2) identification of key factors affecting detection performance; 3) recommendations for improving dataset structures and algorithm adaptation for Ukrainian speech.

Keywords: fake voice detection, speech synthesis, voice conversion, Ukrainian language, ASVspoof, datasets, evaluation metrics, EER, WEER, DSR.

ВИНОГРАДОВ Іван
Державний університет інтелектуальних технологій і зв’язку

ВИЯВЛЕННЯ ГОЛОСОВОГО ФЕЙКУ: СУЧАСНІ ТЕХНІКИ ТА ЗАСТОСУВАННЯ ДЛЯ УКРАЇНСЬКОЇ МОВИ

Предметом дослідження статті є методи розпізнавання підроблених голосів, створених за допомогою технологій синтезу мовлення (TTS) та перетворення голосу (VC), з акцентом на їх адаптацію для української мови. Метою є аналіз сучасних наборів даних, змагань (ASVspoof, ADD Challenge) та алгоритмів детекції для оцінки можливості інтеграції українських мовних ресурсів у міжнародні стандарти або створення спеціалізованого набору даних. Цей підхід спрямований не лише на вирішення проблеми обмеженого обсягу україномовних аудіозаписів у широко використовуваних репозиторіях (значна частина яких охоплює переважно англійську або китайську мови), а й на врахування унікальних фонетичних ознак, різноманітних акцентів і морфологічної складності, притаманних українській мові. Завдяки порівнянню ефективності систем у різних сценаріях підробок дослідники можуть точніше оцінити, як мовноспецифічні чинники впливають на точність класифікації, що зрештою сприятиме розробленню більш надійних механізмів виявлення фальшивих голосів.Завдання: дослідити існуючі набори даних та їхню відповідність українській мові, оцінити ефективність систем виявлення підроблених голосів за допомогою таких метрик, як Equal Error Rate (EER), Weighted EER (WEER) та Detection Success Rate (DSR), а також визначити оптимальний підхід—розширення ASVspoof чи розробка нового мовного ресурсу. Методи дослідження включають систематичний аналіз, порівняння наборів даних та оцінку ефективності сучасних систем синтезу мовлення, таких як ElevenLabs, Assembly AI та Tacotron. Результати свідчать, що адаптація алгоритмів виявлення фальшивих голосів до особливостей української мови підвищує точність та надійність їхньої роботи. Висновки. Дослідження підтверджує, що інтеграція українських мовних даних у міжнародні набори або створення окремого спеціалізованого ресурсу суттєво покращує якість детекції. Крім того, цілеспрямоване залучення різних регіональних діалектів і мовленнєвих профілів виявляється вирішальним чинником для збереження високих значень показника Detection Success Rate (DSR). Результати дослідження вказують, що передові нейронні вокодери, які відтворюють тонкі просодичні й темброві нюанси, потребують спеціалізованих контрзаходів, здатних розпізнавати ледь помітні синтетичні артефакти. Як наслідок, у цій роботі підкреслюється важливість багатоетапного вдосконалення наборів даних, періодичних оновлень алгоритмів і міжмовного бенчмаркінгу для підтримання надійної ефективності за умови появи нових загроз спуфінгу голосу. Наукова новизна отриманих результатів полягає у наступному: 1) проведено перший системний аналіз методів виявлення підроблених голосів для української мови; 2) визначено ключові фактори, що впливають на ефективність розпізнавання голосових фальсифікацій; 3) розроблено рекомендації щодо покращення структури наборів даних та адаптації алгоритмів для українського мовлення.

Ключові слова: виявлення підроблених голосів, синтез мовлення, перетворення голосу, українська мова, ASVspoof, набори даних, оцінювальні метрики, EER, WEER, DSR.

Стаття надійшла до редакції / Received 26.04.2025
Прийнята до друку / Accepted 14.05.2025

1. INTRODUCTION

In recent years, text-to-speech (TTS) and voice conversion (VC) technologies powered by deep learning have significantly improved [1]. These technologies can generate human-like speech that is difficult to distinguish from real audio. OpenAI recently introduced a neural network [2] capable of cloning a voice from a 15-second recording, but due to security concerns, its public release was delayed.

1.1. Motivation

The growing availability of advanced synthesis tools has led to an increase in security threats, including voice spoofing and deepfake fraud. Competitions [3] demonstrate that in 2018, the best systems achieved an average naturalness score of 4.1, with 80% of samples closely matching speaker identity. By 2023, deepfake voices had reached near-human naturalness, though imperfections remained in complex tasks. According to Sumsub’s identity verification platform [4], deepfake-related fraud in the financial sector surged by 700% in 2023 compared to 2022, with the first quarter alone surpassing the total number of incidents recorded in the previous year.

These trends highlight the urgent need for advanced fake voice detection systems. While significant progress has been made in developing detection algorithms, most efforts focus on English and Chinese, leaving other languages, including Ukrainian, underrepresented in research. Given the increasing digitalization in Ukraine addressing this gap is crucial for cybersecurity and speech authentication systems.

1.2. State of the art

In International competitions such as ASVspoof and ADD Challenge [3] have accelerated the development of voice anti-spoofing methods. However, Ukrainian-language voice spoofing remains largely unexplored. Existing datasets predominantly cover English, requiring either adaptation or expansion to accommodate Ukrainian phonetic characteristics.

Fake voice detection performance is typically assessed using Equal Error Rate (EER), Weighted EER (WEER), and Detection Success Rate (DSR), which serve as key benchmarks in evaluating algorithm effectiveness. However, the absence of standardized Ukrainian-language datasets limits the applicability of current approaches and hinders model optimization for local contexts.

1.3. Objectives and tasks

This study examines modern methods for detecting fake voices, evaluates the feasibility of integrating Ukrainian-language data into international frameworks, and explores the development of a dedicated Ukrainian dataset. The research applies comparative analysis of existing datasets, performance evaluation of detection models, and experimental adaptation of TTS/VC synthesis systems such as ElevenLabs, Assembly AI, and Tacotron.

The object of this research is the detection and prevention of fake voices generated using TTS and VC technologies.

The subject of the research is the adaptation and optimization of detection models to improve accuracy for the Ukrainian language.

The purpose of this research is to enhance the effectiveness of fake voice detection systems by incorporating Ukrainian data, identifying key challenges, and proposing solutions for dataset development and algorithm adaptation.

Paper structure. Section 1 reviews existing fake voice detection methods and global benchmarks. Section 2 examines Ukrainian-language datasets and evaluates their suitability for voice spoofing research. Section 3 presents experimental results using standard performance metrics. Section 4 discusses practical implications and challenges in adapting detection systems for Ukrainian. The paper concludes with key findings and recommendations for future research.

2. CURRENT RESEARCH ANALYSIS

Several recent works [1, 5, 8] emphasize that the problem of detecting fake voices (voice fakes) generated using speech synthesis (TTS) and voice conversion (VC) technologies is rapidly gaining relevance. This is primarily due to the rapid development of neural networks, which allow achieving high quality of the speech signal and realistic imitation of the voice of a particular person. Works [2, 3] emphasize that the level of naturalness of synthesized speech has almost approached human, and this creates additional challenges for speaker authentication and fraud protection.

Considerable attention is focused on improving specialized datasets and methods for their evaluation. Studies [5, 6] present the results of competitions (ASVspoof, ADD Challenge), where various attack scenarios (TTS, VC, replay attacks, etc.) are proposed, which complicates the task of recognizing fakes. At the same time, works [8, 9] consider new approaches to building multilingual corpora that include proper nouns, dialect features, and emotionally colored vocabulary. According to these studies, adding fragments of the Ukrainian language to international datasets (for example, within ASVspoof) or creating a separate specialized Ukrainian corpus contributes to a more precise tuning of voice spoofing detection algorithms.

According to the authors [4, 7], one of the decisive factors is the development of extended performance indicators, in particular Weighted EER (WEER) and Pairwise Cross-Model EER. They more fully take into account the multitasking and the ability of the algorithm to generalize to unknown types of attacks. Also, studies [1, 3] draw attention to the use of machine learning methods (for example, deep convolutional and recurrent neural networks), which allow analysing large volumes of speech data and better detecting specific features of synthesized or converted speech. It is important to emphasize that the inclusion of various attack scenarios in the model (TTS with different architectures, VC with different voice characteristics, etc.) increases the system's resistance to new threats.

Thus, modern research indicates the high efficiency of a comprehensive approach: the development of multilingual datasets, the use of extended evaluation indicators (EER, WEER, DSR, Pairwise Cross-Model EER) and the adaptation of algorithms to specific language features. In the context of the Ukrainian language, the development or expansion of corpora taking into account the unique phonetic characteristics and regional differences of speakers is of particular importance. This will significantly improve the detection results and contribute to increasing the level of protection of information systems where voice authentication is used.

3. REVIEW OF MODERN FAKE VOICE DETECTION METHODS

3.1. Classification of voice fake types

The field of audio deepfake detection is rapidly evolving due to advances in deepfake technologies, the increasing number of competitions, the emergence of new datasets, improved evaluation metrics, and enhanced detection methods. These developments enable researchers and specialists to more accurately identify fake audio, which is particularly relevant in the areas of security, financial services, and data protection.

There are several types of voice fake, including text-based speech synthesis, voice alteration to mimic another person, and voice cloning based on a short audio sample. These technologies allow for the creation of synthetic speech as well as the modification or copying of voices, providing a high degree of realism. Table 1 presents these types.

Table 1

Classification of voice fake types

Type	Description	Main characteristics
Text-to-Speech Synthesis (TTS)[6]	Synthesizing voice from text using machine learning models.	The voice and intonation can be artificially generated but sound natural. The speaker's identity is faked.
Voice Conversion (VC)[7]	Replacing one person's voice with another while retaining the speech content.	Timbre and intonation are altered to imitate a specific person's voice. The speaker's identity is faked.
Emotion Fake	Altering the emotional tone of speech while keeping the content and speaker identity unchanged.	The speaker's emotions are modified (e.g., from sadness to joy) while maintaining the same content and voice.
Scene Fake	Changing the background environment or scene while keeping the speech content and voice unchanged.	Background acoustics are altered (e.g., adding echo or noise) to simulate a different location, preserving the same voice.

3.2. Voice fake competition

Over the past few years, a number of competitions have played a key role in accelerating the development of audio deepfake detection technologies. Table 2 presents [8] the characteristics and baseline models of these representative competitions.

Table 2

Comparative analysis of the parameters of the ASVspoof, ADD, DFDC, VCC, and ComParE competitions

Parameter	ADD 2023	ASVspoof 2021	DFDC	VCC	ComParE
Year	2023	2021	2020	2023	2023
Language	English, multilingual data	English	English	Multilingual (voice data)	Task-dependent
Fake type	Audio deepfakes, voice fakes	Audio deepfakes, voice spoofing	Video and audio deepfakes	Voice conversion	Various, including fakes, emotions
Conditions	Complex, multiple rounds	ASV systems protection	Deepfake detection in video	Voice transformation	Paralinguistic analysis
Format	Multi-round competition	Competition	Competition	Competition	Competition
Audio frequency	16-24 kHz	16 kHz	16 kHz	24 kHz	8 to 16 kHz
Genuine utterances	14,907	222,617	Task-dependent	Task-dependent	Task-dependent
Fake utterances	95,383	589,212	Task-dependent	Task-dependent	Task-dependent
Real speakers	>500	48	Task-dependent	Task-dependent	Task-dependent
Fake speakers	>500	48	Task-dependent	Task-dependent	Task-dependent
Availability	Limited to participants	Participant access	Open to general public	Participant access	Participant access
Protection technologies	Neural networks, ML	Specialized ASV models	Neural network models	Various voice conversion models	Machine learning methods
Number of participants	More than 30 teams	More than 15 teams	Over 2,000 participants	Over 20 teams	Over 30 teams
Evaluation metrics	EER, WEER	EER	Accuracy, Precision, Recall	Voice naturalness, similarity	RMSE, Precision, Recall

3.3. Datasets and their characteristics

For successful detection of fake voice, high-quality and diverse datasets are required, including various types of fake recordings and recording conditions. Several specialized datasets have been created in this research field to evaluate the effectiveness of detection algorithms. Table 3 presents the main characteristics of these datasets.

Table 3

Overview of Datasets Used for Voice Fake Detection

Dataset name	Year	Language	Types of fake voice	Size	Purpose and features
ASVspoof	2015-2024	English	Speech synthesis, voice conversion, replay	~200,000 recordings	Designed for anti-spoofing tasks in automatic verification. Includes diverse fake types.
WaveFake	2021	English	Speech synthesis	200 hours	Contains fake recordings generated by modern TTS models. Focuses on complex synthesized audio.
FoR (Fake or Real)	2020	English	Speech synthesis	150 hours	Includes TTS-generated recordings. Intended for testing detector robustness against TTS attacks.
In-the-Wild	2022	English	Various, real audio alterations	1,000+ recordings	Recordings from open sources, including social media. Tests detection in real-world conditions.

Most existing datasets are aimed at detecting entirely fake voices and are limited to a single language (most often English).

There are only a limited number of specialized datasets available for developing models to detect fake audio in the Ukrainian language. However, some existing datasets designed for speech synthesis or recognition tasks can be adapted for this purpose, table 4.

Table 4

Ukrainian Datasets for Fake Voice Detection and Adaptation

Dataset name	Year	Purpose	Size	Description and features
Ukrainian TTS Datasets	2021	Speech synthesis (TTS)	~30 hours of audio	A dataset for creating and testing TTS models. Includes speech recordings with transcriptions.
Common Voice (Ukrainian)	2022	Speech recognition (STT)	122 hours of verified data	An open dataset from Mozilla with Ukrainian audio. Suitable for creating fake audio using TTS.
Lingua Libre (Ukrainian)	2020	Speech recognition (STT)	20+ hours of audio	A Wikimedia project with Ukrainian speech. Contains various types of recordings that are usable for TTS/STT.
Multilingual LibriSpeech (MLS)	2020	Speech recognition (STT)	60 hours of Ukrainian audio	A multilingual dataset for training speech recognition systems. The Ukrainian portion contains audiobooks.
Ukrainian Speech Recognition Dataset	2021	Speech recognition (STT)	200 hours of audio	Includes Ukrainian speech recordings with transcriptions, collected for speech recognition tasks.

To develop a high-quality voice fake detection model for the Ukrainian language, a premium dataset that accounts for phonetic nuances and linguistic specificity is essential. Creating such a dataset is a complex and challenging task.

One approach is to utilize existing Ukrainian-language datasets, such as Common Voice, Lingua Libre, and Ukrainian TTS Datasets, supplemented with generated fakes. This can be done using the services and libraries listed in Table 5.

3.4. Evaluating metrics

One approach is to utilize existing Ukrainian-language datasets, such as Common Voice, Lingua Libre, and Ukrainian TTS Datasets, supplemented with generated fakes. This can be done using the services and libraries listed in Table 5.

Table 5

Tools and Platforms for Generating Synthetic and Fake Audio

Service name	Type	Ukrainian language support	Features
Google Text-to-Speech API	Speech synthesis (TTS)	Yes	High-quality speech synthesis with customizable intonation and accents.
Microsoft Azure Speech	Speech synthesis (TTS)	Yes	Multilingual support; flexible style and speech rate customization.
ElevenLabs	Voice generation and cloning	Partial (depends on data)	Precise voice generation with emotional tone customization.
Resemble AI	Voice generation and cloning	Partial	Custom voice creation, API integration support.
Tacotron + WaveNet	Speech synthesis (TTS)	Yes (during training)	An open library for custom speech synthesis, suitable for Ukrainian.

The Equal Error Rate [20] metric is widely applied to measure classifier performance. It represents the percentage of errors where the proportion of false positives (Pfa) equals the proportion of false negatives (Pmiss). It is calculated as follows:

$$Pfa(\theta) = \frac{\#\{forged\ recordings\ exceeding\ the\ threshold\ \theta\}}{\#\{total\ number\ of\ forged\ recordings\}} \tag{1}$$

$$Pmiss(\theta) = \frac{\#\{genuine\ recordings\ not\ exceeding\ the\ threshold\ \theta\}}{\#\{total\ number\ of\ genuine\ recordings\}} \tag{2}$$

EER corresponds to the value where:

$$Pfa(\theta_{EER}) = Pmiss(\theta_{EER}) \tag{3}$$

This is an extension of the EER [22] metric for multi-round evaluations. The final result is computed as a weighted sum of errors from different rounds:

$$WEER = \alpha \cdot EER_{R1} + \beta \cdot EER_{R2} \tag{4}$$

Detection Success Rate (DSR). This metric is used to evaluate the success of classifying synthesized audio. DSR measures the proportion of correctly classified attack data relative to the total number of attacks. The metric is useful in scenarios requiring an assessment of a model's specific vulnerabilities[22].

Pairwise Cross-Model EER. This metric assesses a model's ability to generalize across different domains (e.g., datasets with various speech synthesizers). It measures EER on test sets significantly different from the training data, revealing how well the model handles unknown types of attacks[22].

4. RESULTS AND DISCUSSION

Effective voice fake detection model development for Ukrainian requires careful analysis of data creation and utilization strategies. Two main approaches are considered: extending the existing ASVspoof dataset by means of including Ukrainian data or creating a new specialized dataset inspired by ASVspoof. Both approaches offer benefits[9] and limitations, necessitating a choice based on criteria like time, resources, linguistic adaptation, and usage flexibility.

Extending ASVspoof offers a practical and cost-effective approach, integrating Ukrainian data into an established framework. This solution enables ready-made analysis protocols, standard metrics (EER, WEER, and DSR), and model testing in a multilingual environment. Adding a Ukrainian subset into the existing dataset simplifies model performance comparisons, providing a unified analytical foundation. However, this approach risks Ukrainian data structure mismatches with ASVspoof's requirements, potentially requiring extra resources for data alignment to a single format.

At the same time, creating a specialized dataset offers more customization possibilities for Ukrainian phonetic and linguistic features. This approach makes it possible to build structural flexibility, create unique attack types, and adapt to Ukrainian realities. However, it demands significant time and resources for design, data collection, and validation. The benefit of this approach lies in full independence from existing solutions, enabling the developing and testing models optimized for specific tasks.

Choosing between these approaches depends on research priorities and available resources. Extending ASVspoof is expedient for the main task of integrating Ukrainian data into global standards and evaluating models in a multilingual environment. Conversely, creating a new dataset is preferable when the key priority is maximum data adaptation to the Ukrainian language and its unique traits.

5. CONCLUSIONS

This article analyzes modern methods and approaches for verifying fake voices (voice fakes), emphasizing the need to develop solutions for the Ukrainian language. It discusses the advantages and limitations of two main approaches: expanding the existing ASVspoof dataset and creating a new specialized resource adapted for Ukrainian speech. A systematic review of contemporary datasets, evaluation metrics, and synthetic data generation tools, such as ElevenLabs, Assembly AI, and Tacotron, is presented, which play a key role in creating fake audio files for testing purposes.

The scientific novelty of this work lies in the proposed adaptation of existing dataset structures for the Ukrainian language, an area previously unexplored. For the first time, the use of standard metrics, including EER, WEER, and DSR, is considered in the context of the unique features of Ukrainian speech. The analysis demonstrates that high-quality integration of Ukrainian data, whether through an extended ASVspoof dataset or the creation of a new one, can significantly improve the effectiveness of voice fake detection models and enhance their practical applicability in cybersecurity contexts.

Future research should focus on implementing the proposed approaches, including the creation of experimental Ukrainian data subsets and performance testing of existing algorithms. This will not only advance the development of voice fake detection technologies but also strengthen the security of speech data processing systems.threats and minimize financial losses.

References

1. Boháček, M., & Farid, H. Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. Proceedings of the National Academy of Sciences,2022. Available at: https://www.semanticscholar.org/paper/Protecting-world-leaders-against-deep-fakes-using-Boh%C3%A1%C4%8Dek-Farid/b76f88db1d94a1b4cc003f56c578d37f1dae9906 (accessed 26.03.2025).
2. OpenAI. Navigating the Challenges and Opportunities of Synthetic Voices. Available at: https://openai.com/index/navigating-the-challenges-and-opportunities-of-synthetic-voices/ (accessed 26.04.2025).
3. Vocal Tract Length Challenge. Vocal Tract Length Challenge. Available at: http://vc-challenge.org/ (accessed 26.04.2025).
4. Sumsub. Sumsub Launches Advanced Deepfakes Detector. Available at: https://sumsub.com/newsroom/sumsub-launches-advanced-deepfakes-detector/ (accessed 26.04.2025).
5. Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., Liu, X., Lee, K. A., Kinnunen, T., & Evans, N. ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection. The ASVspoof 2021 Workshop, 2021. DOI: 10.21437/ASVSPOOF.2021-8.
6. Zhou, Z. Analysis of the Survey of Voice Synthesis Technology. Applied and Computational Engineering, 2023, vol. 4, pp. 490-496. DOI: 10.54254/2755-2721/4/2023310.
7. Schröder, M. Emotional Speech Synthesis: A Review. Proc. 7th European Conference on Speech Communication and Technology (EUROSPEECH), 2001, pp. 561-564. DOI: 10.21437/Eurospeech.2001-150.
8. Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C. Y., & Zhao, Y. Audio Deepfake Detection: A Survey. Journal of LaTeX Class Files, 2023, vol. 14, no. 8. DOI: 10.48550/arXiv.2308.14970.
9. Müller, N. M., Kawa, P., Choong, W. H., Casanova, E., Gölge, E., Müller, T., Syga, P., Sperl, P., & Böttinger, K. MLAAD: The Multi-Language Audio Anti-Spoofing Dataset. arXiv preprint, 2024. DOI: 10.48550/arXiv.2401.09512.
10. Gaydhani, A., Doma, V., Kendre, S., & Bhagwat, L. Detecting Hate Speech and Offensive Language on Twitter Using Machine Learning: An N-gram and TF-IDF-Based Approach. arXiv preprint, 2018, ID 1809.08651. DOI: 10.48550/arXiv.1809.08651
11. Suzuki, M., Itoh, N., Nagano, T., & Kurata, G. Improvements to the N-gram Language Model Using Text Generated from the Neural Language Model. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 7245-7249. DOI: 10.1109/ICASSP.2019.8683481.
12. Shadiev, R., Hwang, W.-Y., Chen, N.-S., & Huang, Y.-M. Review of Speech-to-Text Recognition Technology for Enhancing Learning. Educational Technology & Society, 2014, vol. 17, pp. 65-84. Available at: https://www.airitilibrary.com/Article/Detail/P20221223002-N202302240012-00005 (accessed 26.04.2025).
13. Jones, D. A., Wolf, F., Gibson, E., Williams, E., & Fedorenko, E. Measuring the Readability of Automatic Speech-to-Text Transcripts. Proc. 8th European Conference on Speech Communication and Technology (EUROSPEECH), 2003, pp. 1585-1588. DOI: 10.21437/Eurospeech.2003-463.
14. Juang, B. H., & Rabiner, L. R. Hidden Markov Models for Speech Recognition. Technometrics, 1991, vol. 33, pp. 251-272. DOI: 10.2307/1268779.
15. Picone, J. Continuous Speech Recognition Using Hidden Markov Models. IEEE Signal Processing Magazine, 1990, vol. 7, no. 3, pp. 26-41. DOI: 10.1109/53.54527.
16. A Deep Learning Approach in Text-to-Speech System: A Systematic Review and Recent Research Perspective. Multimedia Tools and Applications, 2022, vol. 82, pp. 15171-15197. DOI: 10.1007/s11042-022-13943-4.
17. Kumar, Y., Koul, A., & Singh, C. A Deep Learning Approach in Text-to-Speech System: A Systematic Review and Recent Research Perspective. Multimedia Tools and Applications, 2023, vol. 82, pp. 15171-15197. DOI: 10.1007/s11042-022-13943-4.
18. Sphinx. Sphinx Sources and Docs. Available at: https://www.sphinx-doc.org/en/master/ (accessed 26.04.2025).
19. Rabiner, L., & Juang, B.-H. Fundamentals of Speech Recognition. Prentice Hall, 1993. Available at: https://www.academia.edu/4924307/Fundamental_of_Speech_Recognition_Lawrence_Rabiner_Biing_Hwang_Juang (accessed 26.04.2025).
20. Zhang, L., Wang, X., Cooper, E., Evans, N., & Yamagishi, J. Range-Based Equal Error Rate for Spoof Localization. Proceedings of INTERSPEECH 2023, pp. 3212-3216. DOI: 10.21437/Interspeech.2023-1214
21. Yi, J., Zhang, C. Y., Zhang, J., Wang, C., Yan, X., Ren, Y., Gu, H., & Zhou, J. ADD 2023: Towards Audio Deepfake Detection and Analysis in the Wild. arXiv preprint, 2024. DOI: 10.48550/arXiv.2408.04967.
22. Yi, J., Fu, R., Tao, J., Liang, S., Lian, Z., Nie, S., et al. ADD 2022: The First Audio Deep Synthesis Detection Challenge. arXiv preprint, 2022. DOI: 10.48550/arXiv.2202.08433.