

Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

**Кафедра комп'ютерно-інтегрованих технологій, автоматизації, робототехніки та
безпекової інженерії**

**X Міжнародна Конференція
ВИРОБНИЦТВО
&
МЕХАТРОННІ СИСТЕМИ 2026**



**X International Conference
MANUFACTURING
&
MECHATRONIC SYSTEMS 2026**

M&MS

2026

**X International Conference
22-23 October
Kharkiv**

M&MS 2026, 22-23 October, Kharkiv, UKRAINE

УДК: 005:004.896:62-65:338.3

Виробництво & Мехатронні Системи 2026: матеріали X-ої Міжнародної конференції, Харків, 22-23 жовтня 2026 р.: тези доповідей / [редкол. І.Ш. Невлюдов (відповідальний редактор)].-Харків: [електронний друк], 2026. – 87 с.

У збірник включені тези доповідей, які присвячені сучасним тенденціям розвитку технологій та засобів виробництва та мехатронних систем, передовому досвіду та впровадженню їх в галузях систем промислової автоматизації та керування виробництвом; системній інженерії; CAD/CAM/CAE системах; мехатроніці (електро-механічних системах, електронних інструментах систем керування, механічних CAD системах); робототехніці та засобах інтелектуалізації; MEMS (сучасних матеріалів та технологіях виготовлення MEMS) та компонентах і технологіях автоматизації видобутку, переробки та транспортування нафти та газу.

Редакційна колегія: І.Ш. Невлюдов, В.В. Євсєєв.

Manufacturing & Mechatronic Systems 2026: Proceedings of X st International Conference, Kharkiv, October 22-23, 2026: Thesises of Reports / [Ed. I.Sh. Nevlyudov (chief editor).] .- Kharkiv .: [electronic version], 2026. - 87 p.

The collection includes the theses of reports on modern trends in the development of technologies and means of production and mechatronic systems, top experience and implementation of them in fields of: industrial automation and production management systems; systems engineering; CAD/CAM/CAE systems; mechatronics (electrical and mechanical systems, electronic control tools, mechanical CAD systems); robotics and intellectual tools; MEMS (modern materials and manufacturing technologies MEMS) and components and technologies for the automation of oil, gas and oil extraction, processing and transportation.

Editorial board: Igor.Sh. Nevlyudov, Vladyslav.V. Yevsieiev

© Кафедра комп'ютерно-інтегрованих технологій, автоматизації, робототехніки та безпекової інженерії (КІТАРБІ), ХНУРЕ, 2026

Міністерство освіти і науки України (МОНУ)
Харківський національний університет радіоелектроніки (ХНУРЕ)
Варшавський університет сільського господарства (WULS - SGGW)
Азербайджанський державний університет нафти і промисловості
Національний університет «Львівська політехніка»
Festo Didactic Україна
Jabil Circuit Ukraine Limited
ТОВ «Науково-виробниче підприємство «УКРІНТЕХ»»
Факультет автоматизації, комп'ютерно-інтегрованих технологій та систем (АКТС)
Кафедра комп'ютерно-інтегрованих технологій, автоматизації, робототехніки та безпекової інженерії (КІТАРБІ),
Державне підприємство «Харківський науково-дослідний інститут технології машинобудування»
Державне підприємство «Південний державний проектно-конструкторський та науково-дослідний інститут авіаційної промисловості»

МАТЕРІАЛИ

Х-ої Міжнародної Конференції

ВИРОБНИЦТВО & МЕХАТРОННІ СИСТЕМИ 2026

(22-23 жовтня 2026)

Харків, Україна

ЗМІСТ

Інна Хондак

Інтелектуальні робототехнічні системи як засіб підвищення безпеки життєдіяльності та ефективності цивільного захисту	10
--	----

Vladyslav Yevsieiv

Intelligent Method of Adaptive Neuro-PWM Speed Control for a Mobile Robot in a Constrained Workspace	13
--	----

I. B. Жарікова

Контактні пристрої на основі гнучких комутаційних структур для електричного контролю електронних компонентів.....	17
---	----

Anton Andreiev, Svetlana Sotnik

Architecture of an Automatic Speech-Based Production Task Generation Service for Flexible Robotic Systems	20
---	----

Y. V. Shalko

Conceptual Approaches to Enhancing the Intelligence of Navigation Methods for Mobile Warehouse Robots in Dynamic Environments	25
---	----

Alina Konieva, Svetlana Sotnik

Methods of Terrain Traversability Assessment Based on Image Analysis for Mobile Robots	29
--	----

Дмитро Гурін

Методи ідентифікації оператора у робочій зоні колаборативного робота	34
--	----

Олександр Мамонтов

Вдосконалення методу CLM з урахуванням системного характеру небезпек виробничого процесу	37
--	----

Diana Sukhomlinova, Svetlana Sotnik

Overview of Online Remote Control Architectures for Unmanned Aerial Vehicles	41
--	----

Kyrylo Horovyi

Comparative Analysis of Wedge Bonding Methods and Factors Affecting the Quality of Contact Interconnections in Microelectronics	46
---	----

Architecture of an Automatic Speech-Based Production Task Generation Service for Flexible Robotic Systems

Anton Andreiev, Svetlana Sotnik

Department CITARSE, Kharkiv National University of Radio Electronics, Ukraine,
Kharkiv, av. Nauki. 14., email: anton.andreiev@nure.ua

Annotation: This work focuses on the architecture of a service that uses language models to convert recognized speech into formalized production tasks for flexible robotic systems. The service integrates voice activity detection, speaker diarization, automatic speech recognition, and a large language model into a single pipeline. The NVIDIA NeMo framework is justified: its pretrained models separate speech fragments from non-speech segments and generate segment labels required for diarization. At the same time, its customization capabilities allow fine-tuning speaker separation accuracy for shop-floor acoustic conditions. Two processing pipelines are described. The first serves a single operator's utterances, while the second handles multi-participant discussions, where the composition of speakers is reconstructed using diarization. A six-step task generation procedure is proposed, encompassing normalization, correction of part and tooling names against a manufacturing operations directory, recognition of the required task, creation of a structured JSON object, its validation, and speaker role verification. It is shown that the task queue in the database becomes the single point of integration between humans and automated equipment. The described approach automates the manual chain of discussion, formulation, and task assignment currently existing in manufacturing environments.

Key words: speech recognition, Large language models, speaker diarization, voice activity detection, flexible robotic systems.

I. INTRODUCTION

Flexible robotic systems have become the foundation of modern discrete manufacturing [1, 2], as they combine the productivity of automated lines with the ability to quickly reconfigure for a new product.

However, a significant portion of the information controlling these systems is still generated through speech: a foreman formulates a task aloud, a process engineer clarifies the sequence of operations, and only then does someone manually enter the result into the information system or forward it to the next shop.

Global industrial digitalization processes have radically changed the speed requirements for such information exchange [3, 4]. The manual chain of "discussion – decision – formulation – delivery to the executor" remains the slowest and least reliable element of production management. Its vulnerabilities are obvious: part of the agreements gets lost, the task formulation becomes distorted during transmission, the authorship of a decision is not recorded, and the shop floor response time depends entirely on the workload of a specific person.

At the same time, speech technologies and language models have reached a level where this chain can be automated [5, 6].

Modern recognition models maintain acceptable quality on noisy speech [7], diarization models restore the list of

speakers in a group conversation, and large language models convert unstructured text into strictly structured objects [8].

Combining these three classes of models transforms an ordinary shop floor conversation into an input signal for the control system.

The purpose of the work is to develop a service architecture that automatically converts speech into formalized production tasks – verified against the technology operation catalog and speaker access rights – and transmits them to a shared queue accessible to both humans and automated systems of flexible robotic complexes.

The relevance of such a solution lies in the fact that it automates the existing organizational logic of production rather than introducing another user interface to it [9].

II. THE ROLE OF SPEECH TECHNOLOGIES IN DEFINING PRODUCTION TASKS

Converting a statement into a task is not performed by a single model. It is a sequence of layers, each reducing uncertainty for the next. Table 1 shows the role that each system component plays in this sequence.

Table 1. The role of components in formulating production tasks

System component	Role in generating production assignments
1	2
Voice Activity Detection (vad_marblenet)	Detection of speech and non-speech segments, as well as removal of pauses and background noise.
Speaker Diarization (diar_msdd_telephonic)	Segmentation of group audio recordings into sections corresponding to individual speakers.
Speaker Identification (titanet-l)	Matching voice embeddings with profiles of registered employees.
Speech Recognition (Whisper)	Converts speech segments into text.
Translation (Whisper)	Translates recognized speech into English, which serves as the working language of the service.

Continue table 1

1	2
Language Model (Gemma 4)	Separates commands from casual conversation and forms a structured work order object.
Infrastructure (FastAPI, Docker, Database)	Provides API endpoints, enables local deployment, and manages the directory, task queue, and user roles.

Each stage processes the result of the previous one; therefore, the quality of the final task largely depends on the quality of each link in the pipeline.

Voice activity detection performs a dual function: it separates speech segments from pauses and non-speech noise areas, and forms segment labels upon which diarization is built; thus, the quality of subsequent processing heavily depends on this stage.

Diarization partitions the group recording into segments of individual speakers, while their identification links each statement to a specific employee and, consequently, to their role.

The Whisper model converts speech segments into text and, if necessary, translates it into English – the working language of the service – so that subsequent stages operate on text in a single language.

The language model acts as an interpretation layer, converting the recognized text into a structured production task, while the infrastructure layer enables running all models locally, ensuring that audio from production meetings does not leave the enterprise [11].

III. ARCHITECTURAL FEATURES OF A SERVICE WITH DUAL AUDIO PROCESSING PIPELINES

The system is implemented in Python as a service based on FastAPI [12]. The service provides a set of auxiliary endpoints for managing the reference book, users, and queue; however, audio processing is concentrated in two main use cases, which differ by the number of speakers in the recording [13].

The general structure of both pipelines is shown in Fig. 1.

The first scenario handles a situation where a single employee interacts directly with the system: they utter a request, and the service must understand which task needs to be generated from it. Here, the audio undergoes Voice Activity Detection (VAD), after which a speech recognition model generates the text of the utterance, and a language model forms the task.

The second scenario serves group discussions, where several people at a workplace coordinate further workflow steps. In this case, speaker diarization is added to the pipeline, allowing the conversation to be segmented by individual speakers. Subsequently, the presence of a target request is determined for each segment, and an appropriate task is formed if needed.

Models from the NVIDIA NeMo framework were used for the acoustic portion of the system [14].

The choice of this framework is attributed to its wide range of ready-made components and the depth of their configuration: a single toolkit provides voice activity detection (vad_marblenet), speaker embedding extraction (titanet-l), and a multi-scale diarization model (diar_msdd_telephonic) [15], while model and algorithm parameters can be tuned according to the acoustic conditions of the industrial environment.

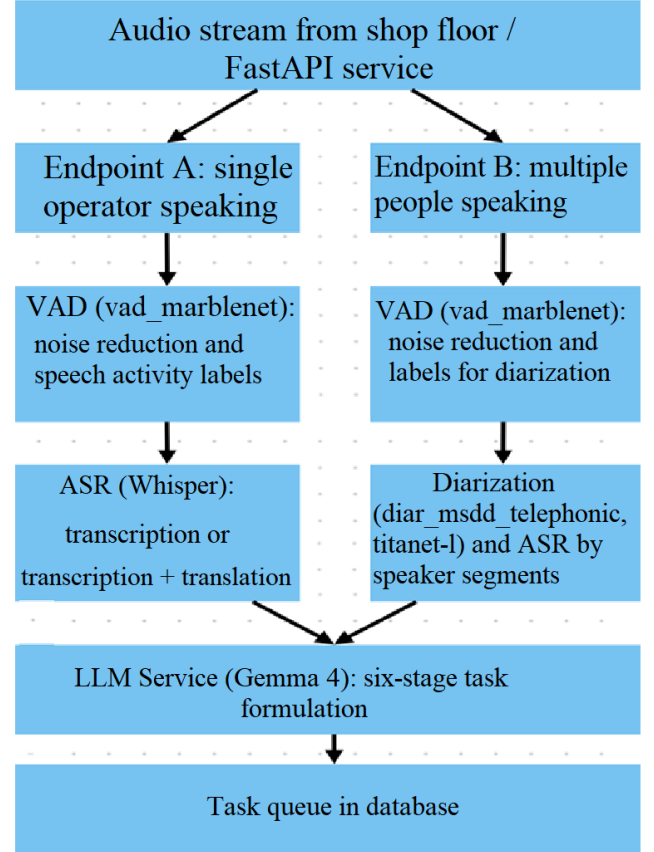


Fig. 1. Two audio processing pipelines in the service

TitaNet-L [16] embeddings are additionally used for registering employee voices, allowing a recording segment to be linked to a specific individual – and consequently to their role – rather than an abstract speaker.

Speech recognition is performed by the Whisper model, which operates in two modes: standard transcription, used when the spoken language matches the service's working language, and transcription with translation (translate mode), used when the speech needs to be converted into the working language.

The working language of the service is English, as this is the target language Whisper translates into. In both cases, the language model receives text in a single language, keeping prompts and the dictionary aligned.

Subsequent processing is identical for both pipelines and is handled by a component designated as the LLM Service. Built as a six-stage chain (shown in Fig. 2), each stage resolves a single specific task instead of delegating the entire transformation to a single model call.

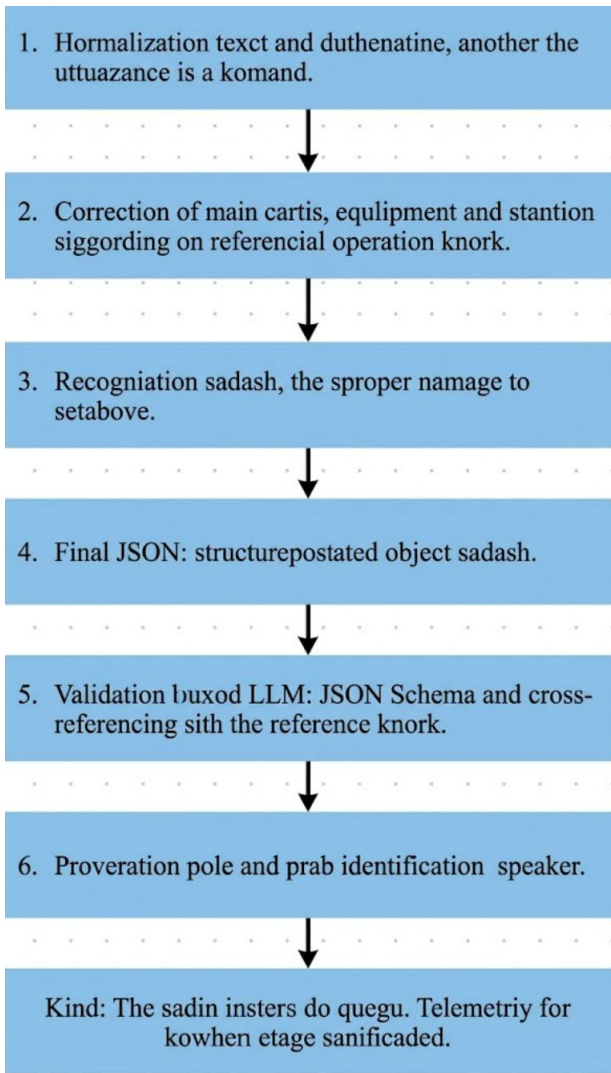


Fig. 2. Six stages of task formulation in an LLM service

1. Normalization and command identification. The recognized text is cleaned of speech artifacts, and the model determines whether the utterance is actually an action request. Most fragments of shop-floor conversation are not commands, and filtering out such fragments at the first stage prevents the creation of false tasks.

2. Terminology correction. Names of parts, tooling, and workstation numbers are matched against the directory of technological operations stored in the database. The directory is maintained in English and contains synonyms as well as abbreviated and distorted variants of names for each item, including those resulting from translation. Spoken language contains abbreviations and distorted names, so this stage translates them into designations that actually exist at the enterprise.

3. Target task recognition. The model identifies which of the known task types the utterance corresponds to and extracts its parameters. Since the directory of available tasks is stored in the database, expanding the set of supported tasks does not require code changes: updating the directory is sufficient.

4. Final JSON generation. The result is assembled into a structured object containing the task type, its parameters, the

author's identifier, a link to the recording fragment, and timestamps.

5. Model output validation. The resulting object is verified deterministically: first against the JSON schema, and then against the directory to ensure that every value of every field corresponds to a real-world entity. An object that fails the check is returned for re-generation or rejected.

6. Role verification. The speaker's permissions are verified before accepting the task [17]; this stage is discussed separately in the following section.

An end-to-end example of a single utterance passing through all stages is shown in Fig. 3.

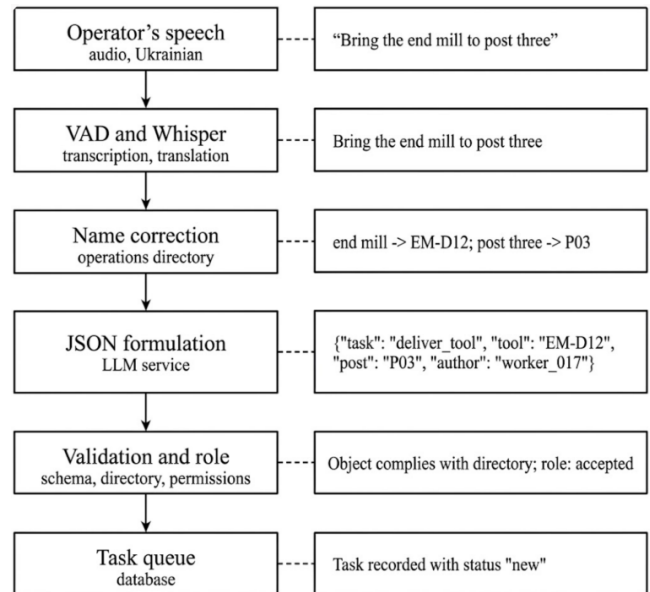


Fig. 3. Example of converting a statement into a production task (single-operator conveyor)

Figure 3 illustrates the processing of a single operator utterance through the first pipeline. After filtering out non-speech fragments, the phrase "Bring the end mill to post three" is recognized by the Whisper model and translated into English. During the terminology correction stage, the expressions "end mill" and "post three" are matched against the reference directory and replaced with identifiers that actually exist at the enterprise (in this example, EM-D12 and P03).

Next, a JSON object is generated; for clarity, the figure displays only the task type, its parameters, and the author, whereas the complete object also contains a link to the audio recording segment and timestamps.

Following validation against the schema and the reference directory, the speaker's role is verified, and the accepted task is written to the queue.

Note: The designations and task type in this example are tentative.

IV. ACCESS CONTROL, TASK QUEUE, AND BUILT-IN METRICS

Permission checks have been deliberately moved outside the language model and implemented in code. The employee's role is passed along with the request to the endpoint, and for group pipelines, it is additionally verified

by matching the voice embedding with a registered profile. The check has three possible outcomes: the task is accepted, rejected as exceeding the speaker's permissions, or routed for confirmation to an employee with a higher role. Thus, the language model influences the content of the task, but never the permission to issue it.

All tasks that pass verification are logged to the database [18, 19]. This queue serves as the single point of integration between the system and the production environment [20, 21], as shown in Fig. 4.

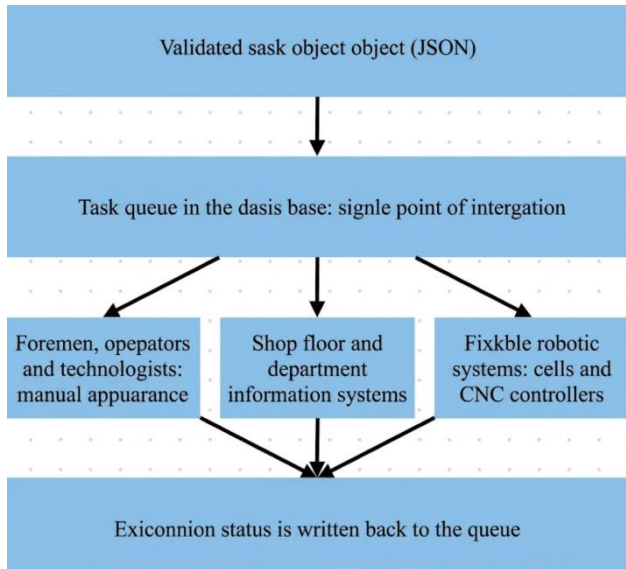


Fig. 4. Job queue as an integration point with the production environment

The queue is read in two ways. Masters, operators, and technologists from other shops and departments see tasks addressed to them and accept them manually; at the same time, automated shop systems and controllers of flexible robotic complexes poll the same queue programmatically and retrieve tasks belonging to them [22, 23]. Execution status is written back, closing the control loop and making the status of each request observable.

An important property of the described architecture is its built-in observability. The service measures the resources consumed at each stage of the chain, which is necessary both for operational latency control and for the economic evaluation of automation. Table 2 lists the recorded indicators.

Table 2. Indicators recorded by the service

Processing stage	Recorded indicators
1	2
Audio intake and VAD	Recording duration, speech ratio in the recording, stage time.
Diarization	Number of detected speakers, stage time, matching confidence with registered profiles.
Recognition and translation	Stage time, volume of extracted text, segment confidence score.
Language model stages	Number of input and output tokens, duration and cost of each call, model and prompt versions.

Continue table 2

1	2
Validation and role checking	Proportion of objects rejected by schema or dictionary, proportion of requests rejected by permissions, number of retry attempts.
Queue entry	Total time from the end of utterance to task appearance, author and task IDs.

The service records the indicators specified in Table 2 for each request and stores them alongside the task itself, enabling the evaluation of the automated chain's duration and cost, as well as the comparison of model configurations. Numerical values accumulated during operation constitute the subject of a separate experimental study and are not presented in this paper.

The main limitations of the approach should also be noted. The quality of task generation depends on the completeness of the manufacturing operations directory, making its maintenance an organizational requirement. Acoustic conditions on the shop floor remain challenging for diarization, making the registration of workers' voice profiles virtually mandatory. Finally, local execution of models requires on-premises computational resources – this is the price paid to guarantee that production audio never leaves the enterprise perimeter.

V. CONCLUSIONS

The paper proposes a service architecture in which language models convert recognized speech into formalized production tasks for flexible robotic systems. The main results are as follows:

1. The composition of the system's acoustic part was defined. The use of models from the NVIDIA NeMo framework is motivated by the broad array of off-the-shelf components and the depth of their customization. It is shown that voice activity detection ensures both the separation of speech fragments from non-speech regions and the generation of segment labels for diarization, while titanet-l embeddings allow linking an audio segment to a specific employee.

2. A service architecture with two processing pipelines was developed. Pipelines for a single operator and for a group discussion are described. They differ by the presence of diarization but converge at a common task formulation component, enabling the service to handle both direct commands and standard conversational discussions on the shop floor.

3. The task formulation process was formalized. A six-stage chain was developed that separates text normalization and command identification, terminology correction using a reference dictionary, identification of the required task, creation of a structured object, its validation, and role verification. Deterministic checks and access control logic were moved out of the language model directly into code.

4. A task queue was proposed as a unified integration point. It is shown that a single queue within a database serves both humans and the automated systems of shop floors and robotic complexes simultaneously, while returning execution status closes the control loop.

The practical value of the obtained results lies in the ability to:

Automate the chain of discussing, formulating, and transferring production tasks, which is currently performed manually;

Preserve the authorship of each request and verify speaker permissions before a task is accepted;

Expand the set of supported tasks by updating the reference dictionary without making code changes;

Estimate the execution time and cost of the automated chain thanks to built-in performance tracking at each stage.

Further development of this work should focus on the experimental evaluation of pipeline accuracy under real shop-floor acoustic conditions [24], investigating the impact of diarization configuration on the correctness of request attribution, and expanding the system to directly generate control programs for robotic cells.

LIST OF REFERENCES

- [1] С. Сотник, "Розробка автоматизованої інформаційно-пошукової системи вибору маніпулятора промислових роботів," Електромеханічні і енергозберігаючі системи, 2025, 1 (68), pp. 52-58
- [2] O. R. Kolbasa, et al., "The significance and necessity of automating the selection of sensors and actuators," «Computer-integrated technologies, automation and robotics» CITAR-2025, 2025, pp. 63-67
- [3] M. S. Achkan, et al., "Integration of cloud technologies into modern SCADA systems: prospects and challenges," «Computer-integrated technologies, automation and robotics» CITAR-2025, 2025, pp. 26-29
- [4] K. A. Polikanov, et al., "Overview of modern technologies for residential automation," «Computer-integrated technologies, automation and robotics» CITAR-2025, 2025, pp. 85-89
- [5] A. Tverdokhlib, et al., "Intelligent tools for optimizing information and search engines," Manufacturing & Mechatronic Systems 2024: Proceedings of VIII st International Conference, Kharkiv, October 25-26, 2024, pp. 28-31
- [6] A. Konieva, et al., "Main trends in the development of automated image processing systems," «Computer-integrated technologies, automation and robotics» CITAR-2025, 2025, pp. 68-71
- [7] A. Radford, et al., "Robust speech recognition via large-scale weak supervision," Proceedings of the 40th International Conference on Machine Learning (ICML), 2023, pp. 28492-28518
- [8] Gemma Team, Google DeepMind, "Gemma 4 technical report," arXiv preprint arXiv:2503.19786, 2025
- [9] A. S. Andreiev, et al., "Computer games and Web design," Proceedings of the XVII International scientific and practical conference «Information technologies and automation – 2024», 2024, pp. 712-714
- [10] F. Jia, S. Majumdar, B. Ginsburg, "MarbleNet: deep 1D time-channel separable convolutional neural network for voice activity detection," Proceedings of ICASSP 2021, 2021, pp. 6818-6822
- [11] S. V. Sotnik, "Analysis of Personal Information Security Issues in Peacetime and Wartime," International Journal of Academic Engineering Research (IJAER), 2024, Vol. 8, Issue 10, pp. 108-113
- [12] S. Sotnik et al., "Development Features Web-Applications," International Journal of Academic and Applied Research (IJAAR), 2023, Vol. 7, Issue 1, pp. 79-85
- [13] S. V. Sotnik, "Features of using REST architecture for development of ARS for information systems," Міжнародна науково-практична конференція «Інформаційні системи в управлінні проектами та програмами», Коблево, 9–13 вересня 2024 р. Збірник праць. – Харків: ХНУРЕ, 2024, pp. 42–45
- [14] O. Kuchaiev, et al., "NeMo: a toolkit for building AI applications using neural modules," arXiv preprint arXiv:1909.09577, 2019
- [15] T. J. Park, et al., "Multi-scale speaker diarization with dynamic scale weighting," Proceedings of Interspeech 2022, 2022, pp. 5080-5084
- [16] N. R. Koluguri, T. Park, B. Ginsburg, "TitaNet: neural model for speaker representation with 1D depth-wise separable convolutions and global context," Proceedings of ICASSP 2022, 2022, pp. 8102-8106
- [17] R. S. Sandhu, et al., "Role-based access control models," IEEE Computer, 1996, Vol. 29, Issue 2, pp. 38-47
- [18] M. Rudenko, et al., "Overview of approaches to scaling relational databases in development and adaptation of web applications," Сучасні проблеми і досягнення в галузі радіотехніки, телекомунікацій та інформаційних технологій: Тези доповідей XII Міжнародної науково-практичної конференції (10-12 грудня 2024 р., м. Запоріжжя). – Запоріжжя: НУ «Запорізька політехніка», 2024, pp. 398-402
- [19] S. Sotnik, M. Rudenko, "Evaluating relational database scaling strategies in web engineering," International Conference on Advanced Trends In Radioelectronics and Infocommunications (ATRICE-2025) (May 21–22, 2025), Lviv Polytechnic Publishing House, Lviv, Ukraine, 2025, pp. 224-228
- [20] Y. I. Khalimonov, et al., "Integration of IoT into security systems: opportunities and risks," Комп'ютерно-інтегровані технології автоматизації технологічних процесів на транспорті та у виробництві : матеріали всеукр. наук.-практ. конф. здобувачів вищ. освіти і молодих учених, 20 листоп. 2024 р., 2024, pp. 117-121
- [21] S. Sotnik, "Integration of IoT into security systems: opportunities and risks," International Journal of Academic Engineering Research (IJAER), 2024, 8 (11), pp. 56-61
- [22] R. V. Marunich, et al., "Features of IoT application in the security sector," «Computer-integrated technologies, automation and robotics» CITAR-2025, 2025, pp. 80-84
- [23] R. V. Marunich, et al., "Modern IoT technologies for creating automated access systems," Sustainable smart cities and communities: business and innovation solutions 2025: Proceedings of I st I International Conference, Kharkiv, April 21, 2025: Theses of Reports, 2025, pp. 38-39
- [24] S. V. Sotnik, et al., "Analysis of searching methods for explosive objects using information technology and computer modeling," Стан, досягнення та перспективи інформаційних систем і технологій / Матеріали XXIV Всеукраїнської науково-технічної конференції молодих вчених, аспірантів та студентів. Одеса, 18-19 квітня 2024 р. – Одеса, Видавництво ОНТУ, 2024, pp. 20-22