ADVANCES IN CYBER-PHYSICAL SYSTEMS

Vol. 10, No. 1, 2025

IoT SYSTEM FOR REAL-TIME AUDIO INFORMATION PROCESSING

Oleh Osadchuk, Igor Olenych

Ivan Franko National University of Lviv, 50, Drahomanova str., Lviv, 79005, Ukraine
Authors' e-mails: *oleh.osadchuk@lnu.edu.ua, igor.olenych@lnu.edu.ua*

https://doi.org/10.23939/acps2025.01.016

Submitted on 12.04.2025

***Abstract*: This paper presents the development and investigation of a speech-to-text conversion and speaker identification system based on a Raspberry Pi microcomputer, designed for local audio data processing in environments with limited network connectivity. The system integrates Silero and WebRTC models for voice activity detection, SpeechBrain for speaker identification, and the Whisper family of models for speech recognition. In particular, a comparative analysis has been conducted on the efficiency of local speech processing using Whisper Tiny and Whisper Large 2 models versus cloud-based processing through the Whisper-1 and Whisper-1-en APIs (the latter applied exclusively to English-language speech). The study evaluates the impact of sentence length, processing time, memory consumption, and recognition accuracy on system performance. The advantages and resource-related limitations of the models in local and cloud-based IoT environments has been analyzed, and the feasibility of their application in real-time and data privacy contexts has been determined. Performance metrics of the models under various conditions has been used for the analysis.**

***Index terms*: Raspberry Pi, IoT, speech-to-text conversion, speaker identification, Whisper models, SpeechBrain.**

I. INTRODUCTION

The current state of development in electronics and information technologies enables significant advancements in modeling natural perception processes such as vision, hearing, and speech processing. Intelligent systems for sound and natural language recognition take a central place in the tasks of automation, analytics, and big data processing. Their key applications include security systems [1], smart homes [2], voice-controlled devices, automated assistants, as well as applications in healthcare, education, and adaptive technologies [3, 4]. The demand for accuracy and operational speed of such systems continues to grow, which defines the relevance of research focused on optimizing speech recognition algorithms.

In most contemporary solutions, speech processing is performed via cloud-based platforms. On the one hand, this approach makes it possible to leverage powerful computational resources for analyzing large volumes of speech data. On the other hand, the use of cloud services has several limitations, including data transmission delays, risks of privacy violations [5], dependence on stable Internet connectivity, and increasing costs related to data transmission and storage [6]. Local speech processing (edge computing) performed directly on IoT devices addresses these challenges. This approach not only reduces data transmission latency but also minimizes energy consumption and enhances data confidentiality [7]. In this context, an important task is evaluating the performance of modern speech models within resource-constrained IoT devices such as the Raspberry Pi.

II. LITERATURE REVIEW AND PROBLEM STATEMENT

Recent speech recognition models based on deep learning, particularly the Whisper family by OpenAI [8], have established themselves as effective solutions for multilingual audio processing, even in complex acoustic conditions. These models offer flexible integration opportunities due to their scalable architecture. Whisper Tiny is designed for low-power, resource-limited devices, such as single-board computers like the Raspberry Pi, while Whisper Large 2 is intended for high-accuracy processing because of its significantly larger model size. Whisper-1 and Whisper-1-en, developed for cloud-based processing via API, provide additional resources for multilingual processing (Whisper-1) and specialized tasks limited to the English language (Whisper-1-en) [9,10]. Performance analysis of these models enables the determination of which configuration is the most effective, depending on the specific task.

The necessity to process large volumes of voice commands or speech data in real time requires a detailed investigation of such parameters as processing time, memory consumption, and text recognition accuracy, depending on sentence length, lexical complexity, and pronunciation variability. Particular interest is also directed toward the influence of local and cloud-based processing architectures, as each of them is suited for different types of IoT tasks and scalable voice processing systems [11, 12].

III. SCOPE OF WORK AND OBJECTIVES

Within the context of the IoT, a constant trade-off exists between device resource consumption, task execution speed, and result accuracy, which requires an objective comparison of the efficiency of models of different scales. Therefore, this study is focused on comparing the performance of local speech processing (Whisper Tiny and Whisper Large 2) with cloud-based processing via the Whisper-1 and Whisper-1-en APIs to identify

optimal approaches for implementing real-time IoT systems. In particular, the study provides a systematic assessment of key performance metrics (processing time, memory consumption, and recognition accuracy) for local and cloud-based computations, offering a foundation for designing IoT solutions.

IV. RESEARCH METHODS

Two audio processing architectures were developed: local and cloud-based. Each architecture implements its approach to real-time speech analysis, offering different usage conditions based on device resources, required accuracy, and processing speed.

Local audio data processing was performed using the Whisper Tiny and Whisper Large 2 models, which were implemented directly on a Raspberry Pi 4 device (Fig. 1). The audio data was processed in real time through voice activity detection (VAD) algorithms, namely Silero and WebRTC. When speech was detected, the audio batch was sent to the model for transcription, while speaker identification was simultaneously performed using the pre-trained SpeechBrain ECAPA-TDNN model.

A key feature of the local architecture is that all operations are executed without transmitting data over the network. This ensures high processing speed and data confidentiality. The main limitations of this approach include the dependence on computational resources and the high memory requirements for large machine learning models.

Cloud-based audio processing was implemented using the Whisper-1 and Whisper-1-en models. The audio signal was stored as a .wav file, and a segment was then sent via the OpenAI API for speech-to-text transcription (Fig. 2). Transcription results were supplemented with speaker information determined using SpeechBrain, as in the case of local processing.

The cloud-based architecture enables the use of powerful language models that provide high accuracy even for complex texts. A distinctive feature of this approach is the dependence on internet connection quality and data transmission delays, which can negatively affect processing speed.

Four Whisper models differing in architectural features, parameter count, and integration method were selected for the study. Table 1 demonstrates key characteristics of the models, such as the number of parameters, language support, and data processing type.

Table 1

Characteristics of Whisper-based Models

Models	Model size	Language support	Data processing
Whisper Tiny	37.8 M	English	Local
Whisper Large 2	1.54 B	Multilingual	Local
Whisper-1	1.54 B	Multilingual	Cloud-based
Whisper-1-en	1.54 B	English	Cloud-based

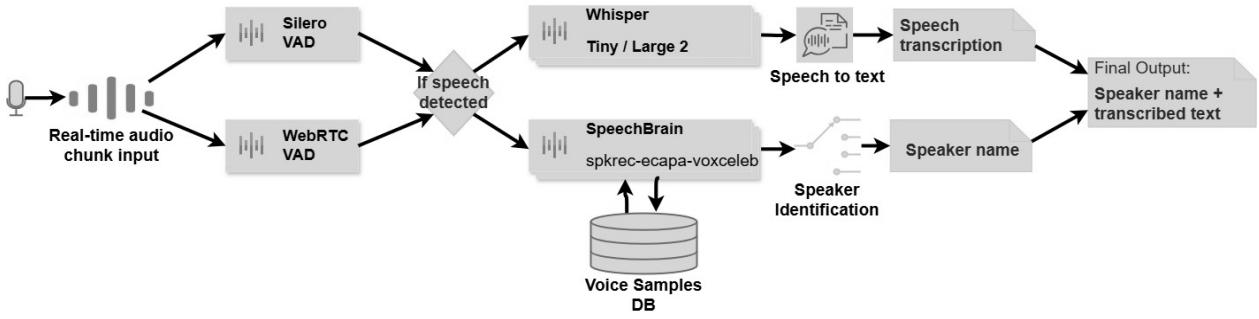


Fig. 1. Local real-time speech processing architecture using Whisper Tiny and Whisper Large 2 models

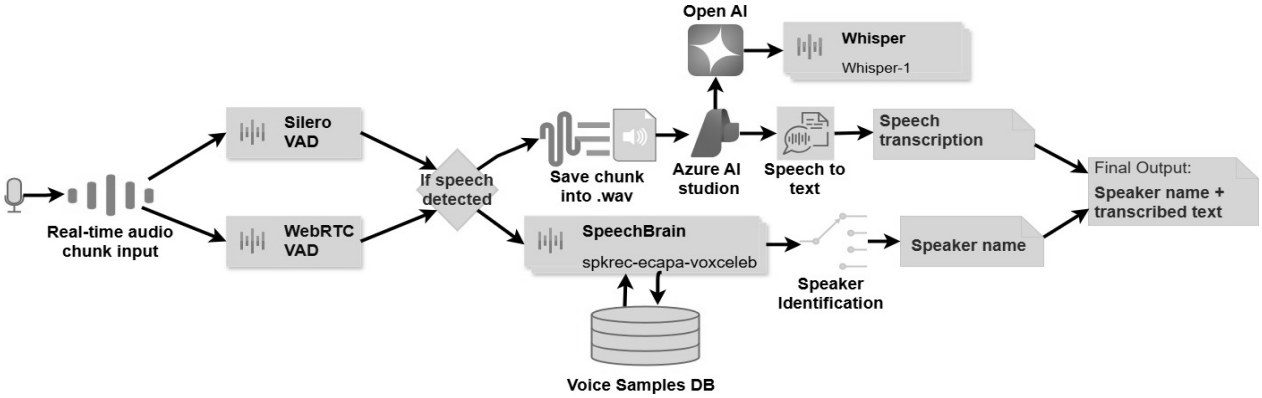


Fig. 2. Cloud-based speech processing architecture via the Whisper-1 and Whisper-1-en API using .wav file format

The Whisper Tiny and Whisper Large 2 models are integrated locally and executed directly on the device, reducing dependence on network stability but limiting processing power due to hardware constraints. Whisper-1 and Whisper-1-en are cloud-oriented, utilizing an API that offers stable performance for large text volumes by leveraging server-side hardware.

Three main types of text, namely everyday vocabulary, technical vocabulary, and literary vocabulary, were used to evaluate the performance of systems. The everyday vocabulary simulated simple voice commands to test basic recognition capabilities in daily scenarios. Technical vocabulary involved texts with complex terminology and multi-level structures, allowing an assessment of the system's ability to understand context and specific terms. Literary vocabulary tested the ability of models to correctly interpret complex syntactic constructions, figurative expressions, and lexical nuances. Sentence length varied from 10 to 101 words, allowing an examination of the impact of text volume on accuracy, processing time, and memory usage. Special emphasis was placed on tests involving accented speech. The data included recordings with regional English accents and pronunciations by non-native English speakers. This enabled testing of the models' adaptability to multicultural environments and real-world language variations. So, the created text sets covered a wide range of use cases, ensuring representativeness and thorough analysis of Whisper model performance in both local and cloud-based processing contexts.

V. RESULTS AND DISCUSSION

The experimental sample for testing each of the four language models consisted of 100 audio fragments. This approach ensured reproducibility of experimental conditions and enabled a comparative analysis of transcription results based on a single dataset. The sample was representative and accounted for several critically important parameters: varying audio fragment lengths (from 10 to 101 words), thematic diversity (everyday, technical, and artistic vocabulary), as well as pronunciation variability (including accented speech). This made it possible to assess the models' adaptability under conditions of changing linguistic characteristics within the audio stream.

A classical metric, which determined the correspondence between the output and the reference text, was applied to evaluate transcription accuracy. The accuracy indicator for each audio fragment was calculated using the formula

$$accuracy = \frac{N_{matched}}{N_{total}} \cdot 100\,\%, \qquad (1)$$

where $N_{matched}$ and N_{total} refer to the number of words reproduced identically by the model and in the original text, respectively.

Analysis of the obtained results demonstrated a direct correlation between audio processing time and the volume of linguistic content (Fig. 3). Short sentences required fewer computational resources and were processed significantly faster, whereas increases in text length resulted in proportionally longer processing times. The best performance was demonstrated by the Whisper-1-en model, which provided the lowest audio data processing times among all tested architectures (within a few seconds). The Whisper-1 and Whisper Tiny models displayed slightly inferior time characteristics compared to Whisper-1-en. For short language fragments, processing times remained stable, with only a moderate increase as the word count grew. This suggests Whisper Tiny is suitable for real-time use on low-power IoT platforms with limited hardware resources. In contrast, the Whisper Large 2 model exhibited a significant increase in processing time, directly correlated with its resource-intensive architecture. For fragments exceeding 60 words, processing times surpassed 70 seconds, making this model impractical for real-time applications in resource-constrained computational environments.

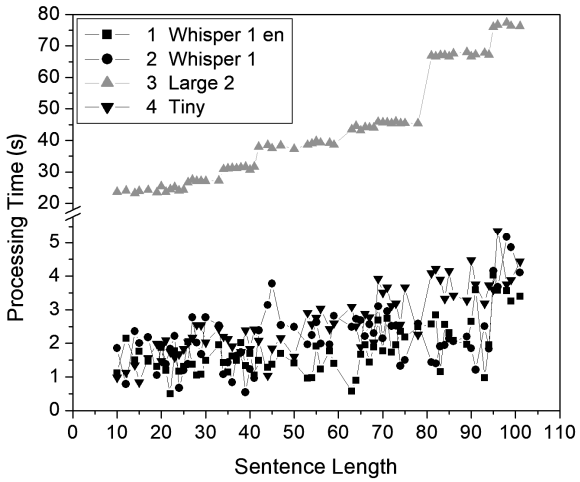


Fig. 3. Dependence of audio data processing time on sentence length for various Whisper-based models

Local models demonstrated a pronounced dependency of performance on the hardware platform characteristics. Processing time remained stable with short texts due to relatively low memory and processor load. However, as the length of the language fragment increased, processing time grew exponentially, particularly for models with large parameter counts, substantially increasing the system's overall load.

At the same time, cloud-based models Whisper-1 and Whisper-1-en provided more stable processing times even for lengthy audio fragments, owing to the use of server-grade hardware with high computational capacity. Under the examined conditions, the average processing time remained within 1.5 to 5 seconds, enabling the use of such models in most scenarios where stability, accuracy, and rapid result acquisition are critical. It is important to note that network latency can significantly affect total processing time in real-world conditions with unstable connectivity.

Preliminary analysis indicated that models with fewer parameters consume less RAM than large-scale models containing billions of parameters. Nonetheless,

this metric cannot be considered an independent efficiency indicator for such integrated systems. This is due to the specifics of the software architecture, which involves the simultaneous operation of several interacting components – Silero and WebRTC for voice activity detection and SpeechBrain for speaker identification. Each of the modules processes the same audio data in parallel with the main Whisper language model, resulting in a proportional increase in total resource consumption.

Since system load is formed by the combined operation of these processes, memory consumption grows linearly with the length of the processed language fragment, regardless of the selected language model (Fig. 4). Even minor advantages of compact architectures in terms of memory usage are neutralized under conditions of integrated system operation. Therefore, absolute memory consumption values of individual models lose priority significance as a criterion for determining suitability for use on low-power devices.

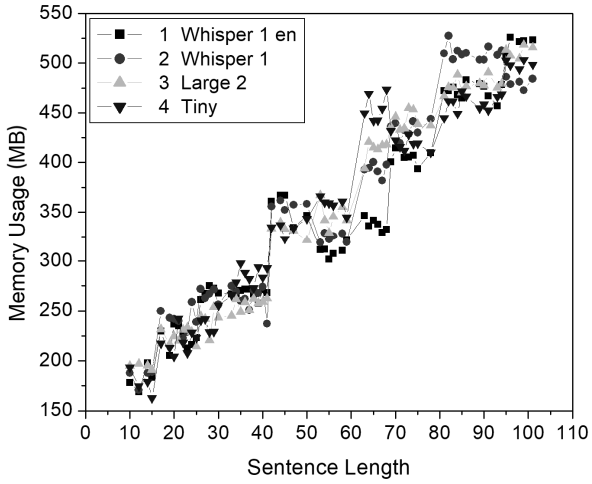


Fig. 4. Dependence of memory consumption on sentence length for various Whisper-based models

The decisive factor is not the RAM volume consumed by a specific algorithm but the system's capacity to maintain stable operation under predictably linear load growth conditions. This parameter eventually determines the appropriateness of a given configuration, both for embedded IoT solutions and hybrid systems with partial reliance on cloud resources for data processing.

The results of the experimental study established that all models exhibited a stable increase in the number of correctly recognized words as text length increased, ensuring high accuracy rates for short and medium-length language fragments (Fig. 5). However, a gradual decline in recognition accuracy was observed for longer sentences when using the Whisper Tiny model, caused by its limited capacity to maintain extended context and process complex syntactic constructions.

In contrast, the Whisper Large 2, Whisper-1, and Whisper-1-en models demonstrated higher and comparable accuracy levels when recognizing long language fragments. This is attributed to a larger number of parameters, extended contextual processing capabilities, and the ability to effectively retain long word sequences, enabling correct handling of syntactically complex sentences regardless of their length. These advantages became particularly evident in fragments with a high word count, where accuracy rates of less resource-intensive models declined.

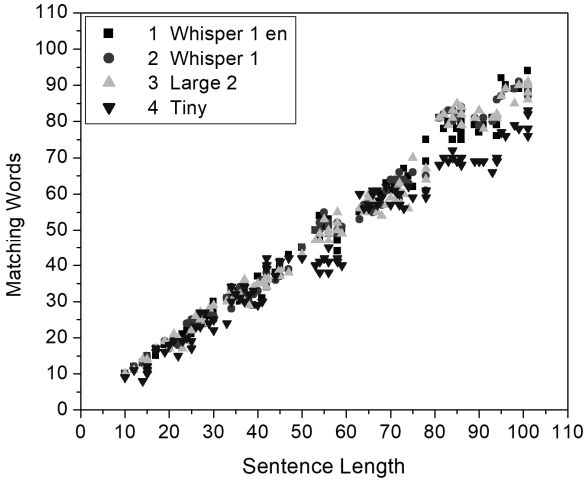


Fig. 5. Relationship between the number of correctly recognized words and sentence length for various Whisper-based models

As illustrated in Fig. 6, compact language models achieve high recognition accuracy on short text fragments with average values exceeding 80–90 %. However, as message length increases beyond 80 words, accuracy rates significantly drop to approximately 67 %, particularly in the presence of specialized terminology and non-standard syntactic structures. This limits their practical application to handling simple, short queries in voice control systems.

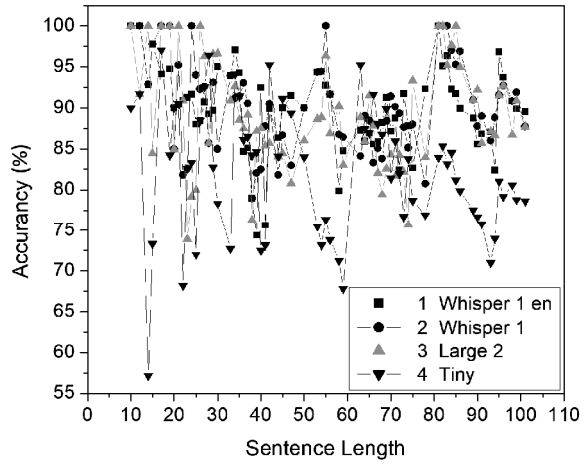


Fig. 6. Dependence of audio recognition accuracy on sentence length for various Whisper-based models

Large-scale architectures maintain consistently high accuracy rates, reaching up to 98 % even on long texts. Nevertheless, their sensitivity to accented or regional pronunciations was identified, which may slightly affect the overall level of performance in specific cases.

Cloud-based models demonstrated the highest versatility, maintaining stable accuracy above 95 % regar-

dless of text length or lexical and grammatical characteristics. Specifically, specialized English-language cloud architectures achieved up to 100 % accuracy when processing texts without pronounced accents or phonetic deviations.

In general, when employing speech transcription models based on Whisper, there is a noticeable variation in accuracy and efficiency depending on the selected model and the characteristics of the involved hardware resources (Table 2).

Table 2

Average values of accuracy, processing time, and memory usage for Whisper-based models

Models	Accuracy, %	Processing time, s	Memory usage, MB
Whisper Tiny	82.31	2.62	355.14
Whisper Large 2	88.99	43.36	355.51
Whisper-1	90.54	2.25	358.05
Whisper-1-en	89.73	1.83	345.55

The Whisper-1 model demonstrates the highest transcription accuracy (90.54 %), combined with moderate memory consumption (358 MB) and an optimal processing time (2.25 seconds). This makes it the most appropriate option for cloud-based data processing systems, where it is essential to maintain high accuracy without compromising processing speed. However, in specific scenarios such as processing English-language audio segments, the Whisper-1-en model has shown a faster processing time (1.83 seconds) with a slightly lower accuracy rate (89.73 %). This characteristic may be advantageous for systems with stricter requirements for processing speed while still ensuring high transcription quality.

On the other hand, the Whisper Large 2 model, despite demonstrating high accuracy (88.99 %), requires significantly more processing time (43.36 seconds), rendering it unsuitable for real-time applications where speed is critical. Nevertheless, this model can be effectively utilized in offline processing scenarios, such as for processing large-scale data batches or pre-processing data intended for subsequent analysis.

The Whisper Tiny model demonstrates the lowest transcription accuracy (82.31 %), which limits its applicability in tasks with stringent accuracy requirements. However, its use can be justified in resource-constrained environments or on devices with limited computational capabilities, where rapid processing of audio fragments is prioritized, even at the expense of some degradation in transcription quality.

The optimal usage of these models can be achieved through adaptive model selection based on the task specifications and hardware resources available. For example, in real-time processing or sensor systems where rapid response is critical, Whisper-1 or Whisper-1-en would be advisable. Conversely, for large-scale batch transcriptions where processing speed is not a primary concern, deploying Whisper Large 2 could be considered, accepting a minor reduction in accuracy in favor of improved handling of data volumes.

VI. CONCLUSION

This study presented a comparative analysis of the efficiency of different Whisper-based models for speech transcription. The Whisper-1, Whisper-1-en, Whisper Large 2, and Whisper Tiny models were assessed based on three key parameters: transcription accuracy, processing time, and memory consumption. The Whisper-1 model demonstrated the highest transcription accuracy (90.55 %) with an optimal processing time (2.25 seconds) and moderate memory usage (358 MB), making it the most effective solution for tasks where both speed and accuracy are critical. The Whisper-1-en model showed slightly lower accuracy (89.74 %) but achieved a faster processing time (1.83 seconds), positioning it as a preferable option for systems where processing time is a decisive factor. Although the Whisper Large 2 model delivered high accuracy (89.00 %), its substantial processing time (43.36 seconds) makes it less suitable for real-time applications requiring prompt responses. The Whisper Tiny model recorded the lowest transcription accuracy (82.32 %), which limits its use in cases of high precision required. However, it remains a viable option for devices with limited hardware resources, where processing speed takes precedence. The analysis of the model metrics allows concluding that the optimal model choice depends on specific requirements for processing time, transcription accuracy, and hardware capabilities. In particular, the Whisper Tiny model is the most practical choice for real-time IoT devices where platform resources are the primary constraint.

References

[1] Sarbast, H. (2024). Voice Recognition Based on Machine Learning Classification Algorithms: A Review. *Indonesian Journal of Computer Science*, 13, 4414–4431. DOI: https://doi.org/10.33022/ijcs.v13i3.4110.

[2] Fatima, I., Fahim, M., Lee, Y. K., & Lee, S. (2013). Analysis and Effects of Smart Home Dataset Characteristics for Daily Life Activity Recognition. *The Journal of Supercomputing*, 66, 760–780. DOI: https://doi.org/10.1007/S11227-013-0978-8.

[3] Luo, X., Zhou, L., Adelgais, K. M., & Zhang, Z. (2024). Assessing the Effectiveness of Automatic Speech Recognition Technology in Emergency Medicine Settings: A Comparative Study of Four AI-powered Engines. DOI: https://doi.org/10.21203/rs.3.rs-4727659/v1.

[4] Wang, X. (2024). Research on Oral English Learning System Integrating AI Speech Data Recognition and Speech Quality Evaluation Algorithm. *Journal of Electrical Systems,* 20, 2466–2477. DOI: https://doi.org/10.52783/jes.2688.

[5] Thandil, R. K., & Basheer, K. P. M. (2020). Accent Based Speech Recognition: A Critical Overview. *Malaya Journal of Matematik,* 8, 1743–1750. DOI: https://doi.org/10.26637/MJM0804/0070.

[6] Subhi, H., Qashi, R., Abdulrahman, L. M., Ayoub, M. & Adil, A. (2023). Performance Analysis of Enterprise Cloud Computing: A Review. *Journal of Applied Science and Technology Trends*, 4, 1–12. DOI: https://doi.org/10.38094/jastt401139.

[7] Sikarwar, S. S. (2025). Computation Intelligence Techniques for Security in IoT Devices. International *Journal on Computational Modelling Applications*, 2(1), 15–27. DOI: https://doi.org/10.63503/j.ijcma.2025.48.

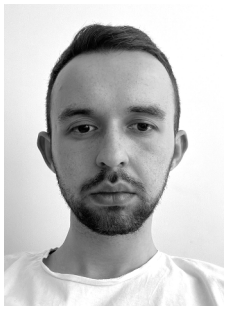
[8] Abnas, M., Imkan, K. M., Ajmal, J. S., Vasudevan A. P., Thampi, S., & Philip, R. K. (2024). Colloquial Language Speech Converter API: A Comprehensive Survey. DOI: https://doi.org/10.20944/preprints202412.2503.v1.

[9] Balan, R. V. S., Vignesh, K., Jose, T., Kalpana, P., & Jothikumar, R. (2024). An Investigation and Analysis on Automatic Speech Recognition Systems. *Journal of Autonomous Intelligence,* 7(3), 1–13. DOI: https://doi.org/10.32629/jai.v7i3.1060.

[10] Cheng, S., Xu, Z., Li, X., Wu, X., Fan, Q., Wang, X., & Leung, V.C.M. (2020). Task Offloading for Automatic Speech Recognition in Edge-Cloud Computing Based Mobile Networks. 2020 IEEE Symposium on Computers and Communications (ISCC), 1–6. DOI: https://doi.org/10.1109/ISCC50000.2020.9219579.

[11] Toshniwal, S., Sainath, T. N., Weiss, R. J., Li, B., Moreno, P., Weinstein, E., & Rao, K. (2018). Multilingual Speech Recognition with a Single End-to-End Model. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4904–4908. DOI: https://doi.org/10.1109/ICASSP.2018.8461972.

[12] Orellana, C., Cereceda-Balic, F., Solar, M., & Astudillo, H. (2024). Enabling Design of Secure IoT Systems with Trade-Off-Aware Architectural Tactics. *Sensors,* 24(22), 7314. DOI: https://doi.org/10.3390/s24227314.



Oleh Osadchuk received the Master's degree in Computer Science at Ivan Franko Lviv National University, Ukraine in 2023. He is presently a PhD student at the Department of Radioelectronic and Computer Systems at Ivan Franko Lviv National University. The main research topics are machine learning applications in embedded systems and data analysis.

Igor Olenych received the degree of PhD Candidate in Physical and Mathematical Sciences from the Institute for Physical Optics, Ukraine in 2010 and the degree of Doctor of Sciences in Physics and Mathematics from Ivan Franko Lviv National University, Ukraine in 2020. He is presently the Head of the Department of Radioelectronic and Computer Systems at Ivan Franko Lviv National University. His current research interests include fuzzy modeling as well as smart solutions and systems.