УДК 004.8:004.932

РОЗРОБЛЕННЯ СИСТЕМИ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ МОВЛЕННЯ З ВИКОРИСТАННЯМ ШТУЧНОГО ІНТЕЛЕКТУ

В. І. Сабат, О. Б. Баса
Національний університет «Львівська політехніка»,
вул. С. Бандери, 12, Львів, 79013, Україна
volodymyr.i.sabat@lpnu.ua, oleh.basa.mm.2022@lpnu.ua

Вирішення проблем автоматичного розпізнавання мовлення сьогодні набирає все більшої актуальності особливо в умовах стрімкого розвитку інформаційних технологій та впровадження штучного інтелекту у всі сфери людської діяльності. Насамперед це пов'язано з тим, що засоби масової інформації стають все більш ефективними у плані швидкості передачі та опрацювання інформації, яка спрямовується на потреби і запити користувача. Сучасні тенденції у медіапросторі такі як мультимедійність та медіаконвергенція ставлять певні вимоги до цього, щоб інформаційне середовище охоплювало усі можливі канали передачі інформації та методи її сприйняття. Так, наприклад, поряд із звуковим мовленням про новини з'являється текстова біжуча стрічка новин, на дикторський супровід накладається відео з останніх подій, текстові синхронні переклади з будь-якої мови, можна інтерактивно вибирати з меню перегляду на екрані монітора і т.д. Розглянута в статті система автоматичного розпізнавання мовлення з використанням Штучного інтелекту дозволяє в автономному режимі конвертувати аудіо-інформацію голосового мовлення у текстовий матеріал, готовий для подальшого опрацювання та друку.

Для розроблення системи автоматичного розпізнавання мовлення важливо визначити основні проблеми та параметри, які б в подальшому визначали якість та достовірність такого процесу. В мережевих онлайн-додатках та програмах з підтримкою ШІ існують методи машинного навчання «людина-машина (система розпізнавання мовлення)», при цьому важливі підготовчі етапи навчання та визначення типових характеристик розпізнавання мовлення людини, які в подальшому можуть корегувати правильність цього чи іншого текстового матеріалу. Також важливим фактором є метод вибору мови, хоча сучасні системи це можуть визначати автоматично, але краще все-таки встановлювати пріоритети вибору мов.. Сьогодні такі методи навчання проводяться для і на основі систем штучного інтелекту, де роль перекладача може бути замінена автоматизованими

програмними комплексами, але як свідчить практика, найкращий результат, наприклад, в системі синхронного перекладу забезпечується комплексним поєднанням досвіду людини-оператора та можливостей штучного інтелекту. При цьому кінцева достовірність збереженого і опрацьованого текстового матеріалу відводиться людині, що вимагає від неї постійного вдосконалення та високого фахового рівня.

У статті розглянуті методи створення програмного застосунку з використанням новітніх технологій ШІ, які дозволяють розпізнавати людське мовлення і конвертувати його у текстовий матеріал, що в подальшому може оброблятись у доступних усім офісних програмах — текстових редакторах.

***Ключові слова**: штучний інтелект, ASR, автоматичне розпізнавання мовлення, Whisper, Python, нейронні мережі, транскрипція аудіо.*

Постановка проблеми. Автоматичне розпізнавання мовлення (ASR, Automated Speech Recognition), або іншими словами технології автоматичного перетворення мовлення людини у текстовий формат мають свою давню історію. Однак, незважаючи на те, що вони були усім давно відомими, значного практичного розвитку та впровадження їх у життя ASR досягли завдяки таким технологіям, як AI (штучний інтелект). Технології автоматичного розпізнавання мовлення сьогодні активно використовуються у мобільних пристроях, голосових помічниках, системах відеоконференцій, медіааналітиці, журналістиці, освіті, телекомунікаціях, цифровому архівуванні, інклюзивних технологіях та системах документообігу. Особливо актуальними такі системи стали після стрімкого розвитку нейронних мереж та методів глибокого навчання. Процес конвертації звуку в текст забезпечує швидкі й точні результати, що дозволяє застосовувати такі методи у реальних мережевих сервісах та на різних платформах. Наприклад, деякі популярні програми, такі як Tik Tok, Spotify, Zoom, YouTube, вбудовують цей процес у свої мобільні програми. Завдяки цьому сьогодні ASR стала однією з найпопулярніших технологій в мережних застосунках [1].

Сучасні інформаційні технології все активніше інтегруються у процеси обробки мультимедійних даних. Одним із найбільш перспективних напрямів розвитку систем штучного інтелекту є автоматичне розпізнавання мовлення (Automatic Speech Recognition, ASR). Такі системи дозволяють перетворювати аудіозаписи у текстовий формат, що значно спрощує обробку голосової інформації у різних сферах діяльності. ASR — це технологія на базі штучного інтелекту, що забезпечує трансляцію акустичного сигналу людського мовлення у текстовий формат за допомогою алгоритмів машинного навчання (ML) та глибоких нейронних мереж (DNN). В сучасних ASR-системах використовуються архітектури Transformers (наприклад, Whisper, Conformer)

або RNN (LSTM/GRU). Вони одночасно моделюють акустичні (acoustic models) та мовні (language models) залежності, що гарантує високу точність (низький рівень WER — Word Error Rate) навіть у складних умовах: за наявності фонового шуму, реверберації, різних акцентів чи високого темпу мовлення. [2]

Традиційні системи розпізнавання мовлення базувались на статистичних моделях та прихованих марковських процесах. Проте сучасні системи ASR дедалі частіше використовують архітектури на основі глибоких нейронних мереж та трансформерів, що забезпечують значно вищу точність розпізнавання мовлення навіть у складних акустичних умовах.

Однією з найперспективніших моделей автоматичного розпізнавання мовлення є Whisper від компанії OpenAI. Дана модель забезпечує підтримку багатьох мов світу, стійкість до шумів, підтримку різних форматів аудіо та високу точність транскрипції. Важливою перевагою Whisper є можливість локального запуску без використання хмарних сервісів, що дозволяє зменшити залежність від зовнішніх API та забезпечити конфіденційність даних.

Актуальність роботи полягає у необхідності створення доступних локальних систем автоматичного розпізнавання мовлення, які можуть використовуватись без постійного підключення до мережі Інтернет та не потребують дорогих комерційних сервісів. Крім того, важливим є питання підтримки української мови у системах ASR, оскільки більшість популярних рішень орієнтовані переважно на англомовний сегмент.

Проте на ринку програмних продуктів не так і багато існує автономних програмних додатків, які б дозволили у напівавтоматичному режимі опрацьовувати великі масиви аудіо інформації у текстову форму. Такі програмні продукти дозволяють перетворювати архівні записи радіопередач, виступи відомих особистостей, або просто записи людської мови на диктофон на якісний текстовий матеріал, який зручний для подальшої роботи та редагування. Це особливо важливо для виготовлення поліграфічних та електронних мультимедійних видань, коли запис звукового супроводу можна прочитувати у титрах на екрані і навпаки текстовий матеріал озвучувати, що є також сьогодні досить популярним і доступним завдяки впровадженню нових технологій ШІ.

У зв’язку з цим виникає необхідність у розробці автономного програмного забезпечення, яке б не вимагало великих ресурсів і дозволяло б здійснювати якісну конвертацію акустичного сигналу людського мовлення у текст.

Аналіз останніх досліджень та публікацій. Проблематика автоматичного розпізнавання мовлення є одним із найбільш динамічних напрямів розвитку сучасних систем штучного інтелекту. За останні роки кількість досліджень у сфері ASR суттєво зросла, що пов’язано зі стрімким

розвитком глибокого навчання, нейронних мереж та високопродуктивних обчислювальних систем.

Однією з фундаментальних праць у сфері сучасних нейромережевих моделей є дослідження «Attention Is All You Need» авторів A. Vaswani та ін., у якому було запропоновано архітектуру Transformer. У статті [3] представлена архітектура Transformer, яка стала основою для сучасних великих мовних моделей, таких як GPT (ChatGPT), BERT та інших.

Значний внесок у розвиток ASR-систем зробила компанія OpenAI, яка представила модель Whisper. Особливістю даної моделі є навчання на великому обсязі мультимовних аудіоданих, що дозволило забезпечити високу якість розпізнавання мовлення навіть за наявності шумів та різних акцентів. [4]

Серед сучасних систем автоматичного розпізнавання мовлення також варто відзначити Google Speech-to-Text, Mozilla DeepSpeech та Microsoft Azure Speech Services. Дані системи демонструють високі показники точності, проте більшість із них орієнтовані на використання хмарної інфраструктури [5,6].

У наукових публікаціях [7-9] особлива увага приділяється питанням:

— підвищення точності розпізнавання мовлення;
— зменшення впливу шумів;
— підтримки багатомовності;
— оптимізації швидкодії систем;
— локального виконання моделей без використання зовнішніх серверів.

Питання підтримки української мови у системах ASR залишається актуальним, оскільки значна частина існуючих рішень орієнтована переважно на англомовний контент. Саме тому використання моделей Whisper для створення локальних систем автоматичного розпізнавання українського мовлення є перспективним напрямом досліджень.

Аналіз сучасних досліджень показав, що поєднання нейромережевих моделей, локальної обробки аудіо та графічного інтерфейсу користувача дозволяє створити ефективні програмні системи транскрипції мовлення.

Метою роботи є розроблення програмної системи автоматичного перетворення аудіозаписів у текст із використанням сучасних методів штучного інтелекту.

Для досягнення поставленої мети було сформульовано такі завдання:

— проаналізувати сучасні технології автоматичного розпізнавання мовлення;
— дослідити принципи роботи моделей Whisper;
— реалізувати програмну систему транскрипції аудіо;
— створити графічний інтерфейс користувача;
— реалізувати різні режими якості розпізнавання;
— провести експериментальне дослідження роботи системи.

Виклад основного матеріалу дослідження. Автоматичне розпізнавання мовлення є однією з ключових задач сучасного штучного інтелекту. Основною метою ASR-систем є автоматичне перетворення голосового сигналу у текстову інформацію.

Сучасні системи розпізнавання мовлення можна умовно поділити на декілька поколінь:

1. Системи на основі шаблонного порівняння.
2. Статистичні системи на базі прихованих марковських моделей.
3. Нейромережеві системи.
4. Трансформерні моделі.

Перші системи ASR мали обмежений словниковий запас та вимагали попереднього навчання під конкретного користувача. Подальший розвиток обчислювальної техніки дозволив перейти до статистичних методів, зокрема прихованих марковських моделей (Hidden Markov Models, HMM), які тривалий час залишались основою більшості систем розпізнавання мовлення.

Однак із розвитком технологій глибокого навчання основний акцент змістився у бік нейронних мереж. Глибокі нейронні мережі дозволили значно підвищити якість розпізнавання мовлення, особливо у складних умовах із шумами та різними акцентами.

Сучасні системи ASR використовують архітектури Transformer, які забезпечують ефективну обробку послідовностей даних та враховують контекст мовлення. Саме на основі трансформерів побудована модель Whisper. [10]

Розробка програмної системи автоматичного розпізнавання мовлення

Для реалізації програмної системи було використано мову програмування Python [11]. Вибір Python обумовлений широкою підтримкою бібліотек штучного інтелекту, простотою інтеграції нейромережевих моделей та наявністю засобів швидкої розробки графічних інтерфейсів.

Архітектура системи

Розроблена система складається з таких основних модулів:

— модуль завантаження аудіофайлів;
— модуль обробки аудіо;
— модуль розпізнавання мовлення;
— графічний інтерфейс користувача;
— модуль збереження результатів.

Система підтримує роботу з аудіоформатами:

- WAV;
- MP3;
- M4A;
- OGG.

Після вибору аудіофайлу користувач може обрати модель Whisper, мову розпізнавання та режим роботи системи.

Графічний інтерфейс користувача

Для створення графічного інтерфейсу використано бібліотеку Tkinter. Інтерфейс забезпечує взаємодію користувача із системою та реалізує основні функції [12,13]:

— вибір аудіофайлу;
— вибір моделі розпізнавання;
— вибір мови;
— запуск процесу транскрипції;
— перегляд результатів;
— збереження тексту.

У верхній частині інтерфейсу розташовано кнопки керування та випадаючі списки вибору параметрів системи. Основна область програми використовується для відображення розпізнаного тексту (рис. 1).

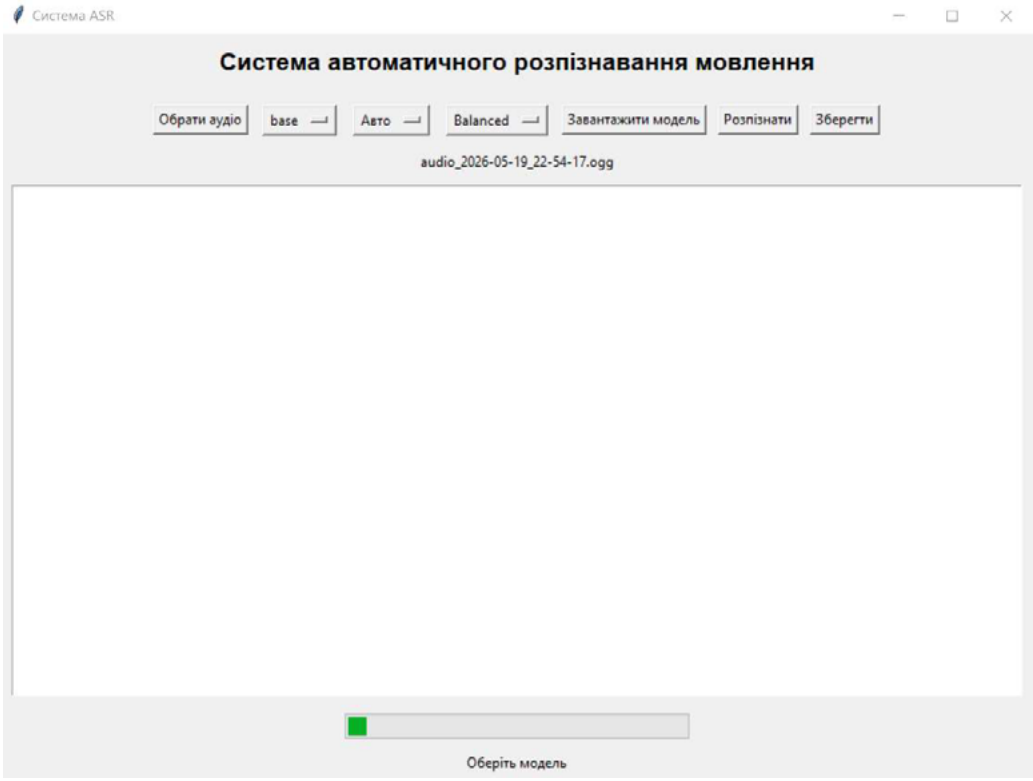

Рис. 1 Вигляд інтерфейсу користувача з можливістю керування параметрами якості (режимами) та моделями розпізнавання мовлення

Режими роботи системи

У розробленій системі реалізовано такі три режими роботи:

1. Fast

Режим Fast забезпечує максимально швидке розпізнавання мовлення. У даному режимі використовуються мінімальні параметри beam search, що дозволяє зменшити час обробки аудіо.

2. Balanced

Balanced є оптимальним режимом для більшості задач. Він забезпечує баланс між швидкістю роботи та якістю розпізнавання.

3. High Quality

Режим High Quality використовує розширені параметри декодування та забезпечує максимальну якість розпізнавання мовлення. Недоліком даного режиму є збільшення часу обробки аудіофайлів.

Реалізація багатопотокової обробки

Для забезпечення стабільної роботи графічного інтерфейсу у процесі розпізнавання мовлення в роботі використано багатопотокову обробку (threading). Це дозволяє уникнути блокування інтерфейсу під час виконання ресурсозатратних операцій. [14,15]

Експорт результатів

Система підтримує експорт результатів транскрипції у формати TXT та DOCX. Для створення документів Microsoft Word використовується бібліотека python-docx.

Експериментальне дослідження системи

Для оцінки ефективності розробленої системи в роботі проведено серію експериментів із використанням різних аудіофайлів та режимів роботи.

Методика експерименту

Тестування проводилось на персональному комп'ютері із процесором Intel Core i5 та операційною системою Windows 10.

Було використано декілька тестових аудіозаписів:

— чисте мовлення українською мовою;
— аудіо із фоновими шумами;
— англомовні записи.

Для кожного тесту оцінювались:

— швидкість обробки;
— якість розпізнавання;
— стабільність роботи системи.

Результати тестування

У процесі експериментального дослідження було встановлено, що режим Fast забезпечує найменший час обробки, однак може допускати помилки при складному аудіо, у якому наявні побічні шуми.

Режим Balanced продемонстрував оптимальне співвідношення між швидкістю та якістю розпізнавання.

Найвищу якість транскрипції забезпечив режим High Quality, проте час обробки аудіофайлів у цьому режимі є найбільшим.
Порівняння режимів роботи системи наведені в табл. 1.

Таблиця 1.

Режим	Швидкість	Якість розпізнавання
Fast	Висока	Середня
Balanced	Середня	Висока
High Quality	Низька	Дуже висока

Отримані результати підтверджують ефективність використання моделі Whisper для задач автоматичного розпізнавання мовлення.

Висновки. У роботі було досліджено сучасні технології автоматичного розпізнавання мовлення та реалізовано програмну систему транскрипції аудіо із використанням технологій штучного інтелекту.
У процесі роботи було проаналізовано принципи функціонування ASR-систем, досліджено можливості моделі Whisper та реалізовано програмний продукт із графічним інтерфейсом користувача.
Розроблена система підтримує роботу з різними форматами аудіо, забезпечує вибір мовного режиму та режимів якості розпізнавання, а також підтримує експорт результатів у текстові формати.

Проведене експериментальне дослідження підтвердило ефективність використання нейромережевих моделей для автоматичного розпізнавання мовлення. Найкращі результати були отримані у режимі High Quality, тоді як режим Balanced забезпечив оптимальне співвідношення між швидкістю та точністю роботи системи.
Перспективами подальшого розвитку системи є інтеграція шумо-подавлення, автоматичне визначення мови, підтримка субтитрів та використання GPU-прискорення.

Список використаних джерел

1. Що таке технологія перетворення мовлення в текст і як вона працює в автоматичному розпізнаванні мовлення. [Електронний ресурс] (Цитовано: 18 вересня 2025) URS: https://uk.shaip.com/blog/automatic-speech-recognitiona-asr/.
2. Автоматичне розпізнавання мовлення (ASR). [Електронний текст] (Цитовано: 05.08.2025). URL: https://denovo.ua/glossary/what-is-asr
3. Ashish Vaswani, Noam Shazeer, Niki Parmar and others. Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017. Pp. 1-11. URL:

https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
4. Пропонуємо до вашої уваги Whisper. [Електронний текст] (Цитовано: 21.08.2022). URL: https://openai.com/uk-UA/index/whisper/.
5. Максимович О. Розпізнавання мовлення від Google Cloud: навіщо використовувати цей сервіс. [Електронний текст] (27.10.2023) URL: https://cloudfresh.com/ua/cloud-blog/google-speech-to-text-navishcho-vykorystovuvaty/
6. Speech Recognition with DeepSpeech using Mozilla's DeepSpeech. [Електронний текст] (Цитовано: 23.06.2025) URL: https://www.geeksforgeeks.org/deep-learning/speech-recognition-with-deepspeech-using-mozilla-s-deepspeech/
7. Douglas O'Shaughnessy. Trends and developments in automatic speech recognition research. Computer Speech & Language, 2024. Vol. 83. URL: https://doi.org/10.1016/j.csl.2023.101538.
8. Sadeen A., Muna S., Alanoud A., Turkiah A. Automatic Speech Recognition: Systematic Literature Review. IEEE Access PP(99):1-1, 2021. Pp. 1-20. URL: https://doi.org/10.1109/ACCESS.2021.3112535.
9. Yao Liu. A systematic literature review of research on automatic speech recognition in EFL pronunciation. Cogent Education. 2025. Vol. 12. I. 1. Pp. 1-15. URL: https://doi.org/10.1080/2331186X.2025.2466288
10. OpenAI Whisper Documentation. URL: https://github.com/openai/whisper
11. Python Software Foundation. Python Documentation. URL: https://www.python.org/
12. Tkinter Documentation. URL: https://docs.python.org/3/library/tkinter.html
13. PyInstaller Documentation. URL: https://pyinstaller.org/
14. Torch Documentation. URL: https://pytorch.org/
15. NumPy Documentation. URL: https://numpy.org/

References

1. What is speech-to-text technology and how does it work in automatic speech recognition. [Electronic resource] (Cited: September 18, 2025) URS: https://uk.shaip.com/blog/automatic-speech-recognitiona-asr/.
2. Automatic speech recognition (ASR). [Electronic text] (Cited: August 5, 2025). URL: https://denovo.ua/glossary/what-is-asr
3. Ashish Vaswani, Noam Shazeer, Niki Parmar and others. Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017. Pp. 1-11. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
4. Introducing Whisper. [Electronic text] (Cited: 08/21/2022). URL: https://openai.com/uk-UA/index/whisper/.
5. Maksimovich O. Speech Recognition from Google Cloud: Why Use This Service. [Electronic text] (10/27/2023) URL: https://cloudfresh.com/ua/cloud-blog/google-speech-to-text-navishcho-vykorystovuvaty/

6. Speech Recognition with DeepSpeech using Mozilla's DeepSpeech. [Electronic text] (Cited: 2025-06-23) URL: https://www.geeksforgeeks.org/deep-learning/speech-recognition-with-deepspeech-using-mozilla-s-deepspeech/
7. Douglas O'Shaughnessy. Trends and developments in automatic speech recognition research. Computer Speech & Language, 2024. Vol. 83. URL: https://doi.org/10.1016/j.csl.2023.101538.
8. Sadeen A., Muna S., Alanoud A., Turkiah A. Automatic Speech Recognition: Systematic Literature Review. IEEE Access PP(99):1-1, 2021. Pp. 1-20. URL: https://doi.org/10.1109/ACCESS.2021.3112535.
9. Yao Liu. A systematic literature review of research on automatic speech recognition in EFL pronunciation. Cogent Education. 2025. Vol. 12. I. 1. Pp. 1-15. URL: https://doi.org/10.1080/2331186X.2025.2466288
10. OpenAI Whisper Documentation. URL: https://github.com/openai/whisper
11. Python Software Foundation. Python Documentation. URL: https://www.python.org/
12. Tkinter Documentation. URL: https://docs.python.org/3/library/tkinter.html
13. PyInstaller Documentation. URL: https://pyinstaller.org/
14. Torch Documentation. URL: https://pytorch.org/
15. NumPy Documentation. URL: https://numpy.org/

DOI 10.32403/2411-9210-2026-1-55-113-123

DEVELOPMENT OF AN AUTOMATIC SPEECH RECOGNITION SYSTEM USING ARTIFICIAL INTELLIGENCE

V. I. Sabat, O. B. Basa

Lviv Polytechnic National University,
12 S. Bandera Street, Lviv, 79013, Ukraine
volodymyr.i.sabat@lpnu.ua, oleh.basa.mm.2022@lpnu.ua

Solving the problems of automatic speech recognition is becoming increasingly relevant today, especially in the context of the rapid development of information technologies and the integration of artificial intelligence into all spheres of human activity. This is primarily due to the fact that mass media are becoming more efficient in terms of the speed of information transmission and processing aimed at meeting user needs and requests. Modern trends in the media environment, such as multimedia and media convergence, impose certain requirements for information systems to cover all possible channels of information transmission and methods of perception. For example, along with audio news broadcasting, text news tickers are displayed, video footage of recent events is superimposed on announcer narration,

synchronous text translation from various languages is provided, and interactive menu selection on display devices is available.

The automatic speech recognition system based on artificial intelligence considered in this paper allows autonomous conversion of audio speech information into text material ready for further processing and printing.

For the development of an automatic speech recognition system, it is important to determine the main problems and parameters that further define the quality and reliability of such a process. In network-based online applications and AI-supported software systems, machine learning methods based on "human-machine (speech recognition system)" interaction are widely used. In this case, important preparatory stages include training procedures and the identification of typical characteristics of human speech recognition, which may later improve the correctness of generated text material.

An important factor is also the method of language selection. Although modern systems can determine the language automatically, it is still preferable to set language priorities manually. Today, such training methods are implemented using artificial intelligence systems where the role of the translator can partially be replaced by automated software solutions. However, practical experience shows that the best results, for example in simultaneous translation systems, are achieved through a comprehensive combination of human operator expertise and artificial intelligence capabilities.

At the same time, the final reliability of the processed and stored textual material still depends on the human operator, which requires continuous professional improvement and a high level of expertise.

The article discusses methods of developing a software application using modern AI technologies that enable recognition of human speech and its conversion into text material, which can later be processed in widely available office software and text editors.

Keywords: *artificial intelligence, ASR, automatic speech recognition, Whisper, Python, neural networks, audio transcription.*

Стаття надійшла до редакції 12.03.2026
Received 12.03.2026