Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Факультет комп’ютерної інженерії та управління
(повна назва)

Кафедра електронних обчислювальних машин
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА

Пояснювальна записка

Рівень вищої освіти другий (магістерський)

Методи обробки гнучких голосових запитів в автоматизованих системах керування завданнями

(тема)

Виконав:

студент II курсу, групи КСМм-23-1

Давидов Я.А.
(прізвище, ініціали)

Спеціальність

123 «Комп’ютерна інженерія»
(код і повна назва спеціальності)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма

Комп’ютерні системи та мережі
(повна назва освітньої програми)

Керівник: доц.Барковська О.Ю.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри ЕОМ _____ Коваленко А.А.
(підпис) (прізвище, ініціали)

2025 р.

Харківський національний університет радіоелектроніки

Факультет комп’ютерної інженерії та управління

Кафедра електронних обчислювальних машин

Рівень вищої освіти другий (магістерський)

Спеціальність 123 «Комп’ютерна інженерія»
(код і повна назва)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма Комп’ютерні системи та мережі
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
“____” _____ 20__ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

студенту Давидову Ярославу Андрійовичу
(прізвище, ім’я, по батькові)

1. Тема роботи Методи обробки гнучких голосових запитів в автоматизованих системах керування завданнями

затверджена наказом по університету від “ 22 ” листопада 2024 р. № 1237 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 20 січня 2025 р.

3. Вхідні дані до роботи

Тренувальний датасет Mozilla Common Voice

Програмне забезпечення розробки ОС Windows 10

Апаратне забезпечення Intel Core i5-6500 CPU 3.20GHz

4. Перелік питань, що потрібно опрацювати у роботі

Огляд методів розпізнавання голосових команд

Оцінка ефективності голосового управління на основі існуючих рішень

Розробка структурної схеми системи розпізнавання мовлення із гнучкими запитами

Проведення експериментів оцінки продуктивності ResNet у залежності від рівня шуму та методів виділення ключових ознак

Аналіз результатів експериментів

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) 17 слайдів

6. Консультанти розділів роботи (заповнюється за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Огляд методів розпізнавання голосових команд	26.11.24-30.11.24	
2	Оцінка ефективності існуючих систем	02.12.24-05.12.24	
	голосового управління		
3	Вибір та обґрунтування методики дослідження	06.12.24-10.12.24	
4	Вибір інструментальних засобів	11.12.24-21.12.24	
5	Розробка структурної схеми системи	23.12.24-03.01.25	
	розпізнавання мовлення із гнучкими запитами		
6	Оформлення матеріалів кваліфікаційної роботи	04.01.25-07.01.25	
7	Подання кваліфікаційної роботи керівникові	08.01.25-11.01.25	
	та її попередній захист		
8	Подання кваліфікаційної роботи на рецензу-	13.01.25-17.01.25	
	вання		

Дата видачі завдання 25 листопада 2024 р.

Студент _____ (підпис)

Керівник роботи _____ (підпис) доц.Барковська О.Ю. (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка кваліфікаційної роботи: 90 с., 28 рис., 16 табл., 22 джерел.

ГНУЧКІ ГОЛОСОВІ ЗАПИТИ, MFCC, RESNET, ASR, АВТОМАТИЗОВАНІ СИСТЕМИ.

Метою кваліфікаційної роботи є розробка ефективних методів обробки гнучких голосових запитів для автоматизованих систем керування завданнями.

У ході виконання кваліфікаційної роботи досліджено методи обробки гнучких голосових запитів в автоматизованих системах керування завданнями. Було розроблено архітектуру системи розпізнавання мовлення, яка включає базу даних синонімів для покращення обробки гнучких голосових запитів. Проведено експерименти в умовах високого та низького рівня зашумленості без направленого мікрофона на джерело звуку та з ним.

Окрім цього, виконано порівняння ефективності запропонованого підходу на основі нейромережевої моделі ResNet із визначеними пайплайнами обробки аудіорядів. Результати дослідження підтвердили, що запропонована архітектура забезпечує покращену точність розпізнавання запитів у складних акустичних умовах.

ABSTRACT

Master’s thesis: 90 pages, 28 figures, 16 tables, 22 sources.

FLEXIBLE VOICE QUERIES, MFCC, RESNET, ASR, AUTOMATE SYSTEMS.

The major goal of this thesis is to develop effective methods for processing flexible voice queries for automated task management systems.

In order to investigate methods of processing flexible voice queries in automated task management systems. The architecture of the speech recognition system was developed, which includes a synonym database to improve the processing of flexible voice queries. Experiments were conducted in high and low noise conditions without and with a microphone pointed at the sound source.

In addition, the effectiveness of the proposed approach based on the ResNet neural network model was compared with the defined audio sequence processing pipelines. The study results confirmed that the proposed architecture provides improved query recognition accuracy in complex acoustic conditions.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

АСК – автоматизовані системи керування

США – Сполучені Штати Америки

ASR – автоматичне розпізнавання мовлення (англ., Automatic Speech Recognition)

BERT – двоспрямовані кодувальні представлення з трансформерів (англ., Bidirectional Encoder Representations from Transformers)

CQT – постійне Q-перетворення (англ., Constant-Q Transform)

CNN – згорткова нейронна мережа (англ., Convolutional Neural Networks)

CPU – центральний процесор (англ., Central Processing Unit)

DCT – дискретне косинусне перетворення (англ., Discrete Cosine Transform)

DNN – глибокі нейронні мережі (англ., Deep Neural Network)

DSP – процесор цифрових сигналів (англ., Digital Signal Processor)

FFT – швидке перетворення Фур'є (англ., Fast Fourier Transform)

GPT – генеративний попередньо тренований трансформер (англ., Generative Pre-trained Transformer)

GPU – графічний процесор (англ., Graphics Processing Unit)

GRU – вентильні рекурентні вузли (англ., Gated Recurrent Units)

LSTM – довга короткочасна пам'ять (англ., Long Short-Term Memory)

MFCC – мелчастотні кепстральні коефіцієнти (англ., Mel-Frequency Cepstral Coefficients)

MVC – Mozilla Common Voice

NLG – генерування природної мови (англ., Natural Language Generation)

NLP – обробка природної мови (англ., Natural Language Processing)

RNN – рекурентні нейронні мережі (англ., Recurrent Neural Networks)

SNR – співвідношення сигнал/шум (англ., Signal-to-Noise Ratio)

STFT – віконне перетворення Фур’є (англ., Short-Time Fourier Transform)

TPU – тензорний блок обробки (англ., Tensor Processing Unit)

ВСТУП

У сучасному світі автоматизовані системи керування завданнями (АСК) відіграють важливу роль у підвищенні продуктивності та оптимізації робочих процесів у різних галузях. Ці системи полегшують розподіл ресурсів, спрощують управління складними проєктами та автоматизують рутинні завдання. Однією з ключових технологій, що значно розширює можливості таких систем, є голосові асистенти, які використовують голосові запити для взаємодії з користувачами.

Використання голосових запитів значно спрощує роботу з АСК роблячи їх більш доступними та інтуїтивно зрозумілими. Однак існуючі методи обробки голосових команд мають певні обмеження, особливо щодо точності розпізнавання складних запитів та інших факторів. Це підкреслює необхідність розробки більш гнучких методів, здатних ефективно справлятися з такими завданнями.

Актуальність дослідження полягає в тому, що вдосконалення методів обробки гнучких голосових запитів сприятиме підвищенню загальної ефективності автоматизованих систем управління завданнями. Вивчення та розробка нових підходів у цій галузі не тільки покращить користувацький досвід, але й відкриє нові можливості для впровадження голосових технологій у різних галузях.

В умовах постійного зростання обсягу завдань, які потребують швидкої обробки та організації, автоматизовані системи керування завданнями стають незамінними інструментами. Інтеграція голосових технологій у ці системи спрощує їх використання, надає швидкий доступ до важливої інформації та забезпечує зручну взаємодію між користувачем і системою. Як наслідок, ці автоматизовані системи набувають більшої гнучкості та адаптивності, що підвищує їхню ефективність у широкому спектрі завдань – від управління проєктами до застосування в охороні здоров'я та освіті.

1 ОГЛЯД ПРОБЛЕМНОЇ ОБЛАСТІ

1.1 Визначення проблемної області

Автоматизовані системи керування завданнями (АСК) – це програмні інструменти, призначені для планування, моніторингу та автоматизації процесів виконання завдань. Ці системи допомагають оптимізувати розподіл ресурсів, підвищити продуктивність і сприяють кращій організації робочих процесів.[1]

Завдяки гнучким можливостям конфігурації та інтеграції з іншими програмами, АСК дозволяють адаптувати робочі процеси до конкретних потреб організації. Вони підтримують такі функції, як планування, визначення пріоритетів, відстеження прогресу та обмін інформацією, що дозволяє командам ефективно співпрацювати та швидко реагувати на зміни в робочому середовищі.

АСК широко використовуються в різних галузях завдяки своїй здатності підвищувати продуктивність, впорядковувати процеси та спрощувати управління проєктами. На рисунку 1.1 показано класифікаційну схему основних сфер застосування автоматизованих систем керування завданнями.

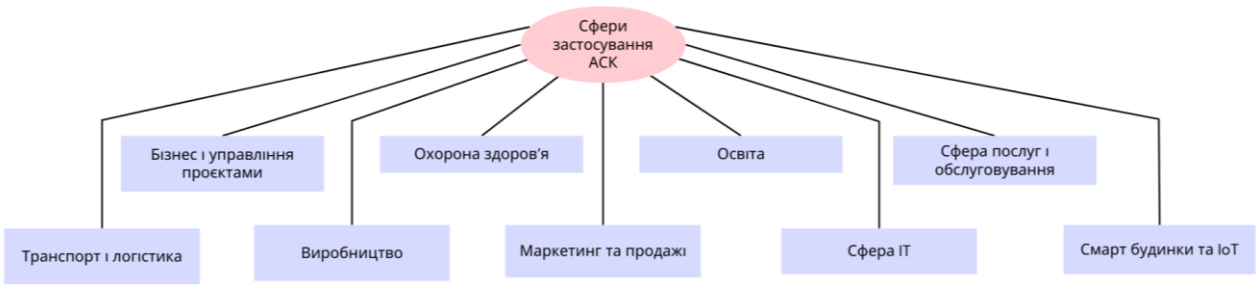

Рисунок 1.1 – Сфери застосування АСК

В таблиці 1.1 показано сфери застосування АСК, а також приклади їх використання в конкретних сферах діяльності.

Таблиця 1.1 – Сфери застосування АСК

Сфера застосування	Приклади задач
Бізнес та управління проєктами	АСК допомагають автоматизувати управління проєктами, розподіляючи завдання, відстежуючи дедлайни та забезпечуючи своєчасну звітність. Вони також оптимізують використання ресурсів, підвищуючи ефективність роботи команди та контролюючи якість виконання проєктів.
Охорона здоров’я	АСК використовуються для систематизації та автоматичного оновлення медичної документації пацієнтів, допомагають планувати візити до лікаря та контролювати дотримання пацієнтами призначень, що полегшує управління процесом лікування.
Освіта	АСК спрощують планування навчального процесу, автоматизуючи розклад занять і контроль виконання завдань. Вони також сприяють дослідницькій роботі, покращуючи управління проєктами та забезпечуючи ефективний моніторинг результатів навчання.
Сфера послуг та обслуговування	АСК забезпечують ефективне управління зверненнями клієнтів шляхом автоматизації розподілу запитів та відстеження їхнього прогресу. Вони також допомагають швидко реагувати на скарги, покращуючи якість обслуговування, та оптимізують робочий графік працівників.
Транспорт та логістика	АСК використовуються для відстеження товарів на всіх етапах ланцюга поставок, від виробництва до доставки. Вони також оптимізують маршрути доставки та забезпечують автоматичне поповнення складських запасів, що підвищує точність і ефективність процесів.

Продовження таблиці 1.1

Сфера застосування	Приклади задач
Маркетинг та продажі	АСК використовуються для планування та управління рекламними кампаніями, забезпечуючи автоматичне відстеження результатів та ефективності. Вони також допомагають контролювати процес укладання угод та аналізують ринок для оптимізації стратегій продажів.
Сфера IT	АСК допомагають керувати розробкою програмного забезпечення, забезпечуючи розподіл завдань, дотримання термінів і тестування. Вони також підтримують процеси DevOps та управління релізами, сприяючи безперервному розвитку й оновленню продуктів.
Виробництво	АСК оптимізують робочі процеси, допомагаючи планувати та контролювати ресурси для забезпечення стабільності та якості продукції. Вони також автоматизують замовлення матеріалів у постачальників, підвищуючи ефективність виробництва.
Смарт-будинки та IoT	В розумних будинках АСК керують різноманітними пристроями, автоматизуючи освітлення, опалення та безпеку. Вони також забезпечують автоматичне реагування на тривоги й оптимізують енергоспоживання, що підвищує комфорт і заощаджує ресурси.

1.2 Актуальність дослідження

З розвитком технологій голосових помічників та АСК, зростає потреба у вдосконаленні методів обробки гнучких голосових запитів. Зручність і

швидкість взаємодії з системами за допомогою голосових команд значно спрощують процеси управління, особливо в тих сферах, де потрібна швидка і безперервна взаємодія з системами. Впровадження таких рішень розширює функціональність АСК, роблячи їх більш доступними для ширшого кола користувачів, в тому числі для людей з обмеженими можливостями.

Голосовий асистент – це програмне забезпечення, яке використовує технологію розпізнавання природної мови для взаємодії з користувачами за допомогою голосових команд. Голосові асистенти дозволяють користувачам виконувати різноманітні завдання, такі як пошук інформації, керування пристроями, планування подій та автоматизація різних дій, без необхідності фізичного втручання. Завдяки здатності обробляти та інтерпретувати людську мову, голосові асистенти забезпечують зручний, інтуїтивно зрозумілий спосіб взаємодії з пристроями та системами, що підвищує ефективність і знижує навантаження на користувача.[2]

Голосові запити – це команди або запитання, які користувачі вимовляють вголос для отримання інформації або виконання певних дій за допомогою голосового асистента або іншого програмного забезпечення для розпізнавання мови. Голосові запити дозволяють взаємодіяти з різними системами без ручного введення, що робить їх особливо корисними в умовах багатозадачності або коли руки користувача зайняті.[3]

Класифікація голосових запитів показана на рисунку 1.2.

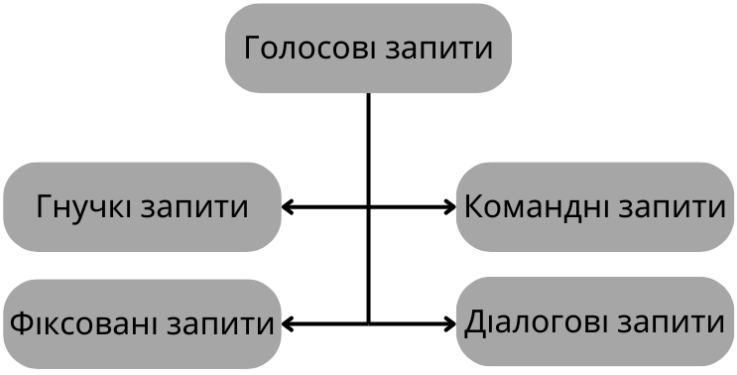

Рисунок 1.2 – Класифікація голосових запитів

У таблиці 1.2 наведено тип голосового запиту, його характеристика та приклад голосового запиту.

Таблиця 1.2 – Класифікація голосових запитів

Тип запиту	Характеристика	Приклад
Фіксовані голосові запити	Запити з чітко визначеними командами та обмеженим набором опцій. Використовуються в системах де важлива точність і простота команд.	"Увімкни світло", "Постав будильник на 7.00".
Гнучкі голосові запити	Запити у довільній формі, де користувач може виразити команду різними способами. Система повинна «розуміти» різні варіанти фраз.	"Постав будильник на 7 ранку", "Знайди мені найближче кафе"
Діалогові голосові запити	Запити, які є частиною постійного діалогу між користувачем і системою. Система зберігає контекст діалогу для коректної інтерпретації	"Яка погода сьогодні?"– "А завтра?"
Командні голосові запити	Запити, які спрямовані на виконання певної дії. Вони можуть бути простими або складними, залежно від завдання.	"Відкрий браузер", "Запусти музику"

Використання голосових запитів за допомогою природньої мови в АСК значно спрощує взаємодію користувача з системою, дозволяючи виконувати складні дії, такі як створення, оновлення та відстеження завдань, за допомогою голосових команд. Така інтеграція робить систему більш інтуїтивно зрозумілою та доступною для ширшого кола користувачів.

За допомогою голосових запитів у системах керування завданнями користувачі можуть створювати та керувати завданнями, оновлювати їхні статуси завдань, додавати нотатки та деталі, встановлювати нагадування та сповіщення, а також їх можна інтегрувати в інші системи. Наприклад, користувачі можуть швидко додавати нові завдання, вказуючи, що потрібно зробити, а також встановлювати терміни і пріоритети. Команда на кшталт «Додати завдання підготувати звіт до п'ятниці» може миттєво створити нове завдання із зазначеним дедлайном.

Голосові команди також спрощують оновлення статусів завдань, дозволяючи користувачам вказувати на завершення або поточний етап виконання завдання. Наприклад, команда «Позначити завдання як виконане» може закрити завдання без необхідності шукати його в системі.

Голосові запити також можуть спростити додавання нотаток і деталей, дозволяючи користувачам диктувати відповідну інформацію, уточнювати вимоги або залишати коментарі для колег, що значно полегшує процес документування.

Нагадування та сповіщення можна налаштувати за допомогою голосових команд, щоб попередити користувачів про наближення дедлайнів або дії, які необхідно виконати. Наприклад, команда «Нагадай мені про зустріч завтра о 10:00» автоматично встановить відповідне нагадування.

Важливо, що голосові команди можуть інтегруватися з іншими інструментами, такими як календарі або електронна пошта, автоматизуючи ще більшу частину робочих процесів. Така команда, як «Надіслати оновлений статус завдання на електронну пошту», може виконати цю дію автоматично.

Загалом, використання голосових запитів для керування завданнями спрощує їх виконання, покращує ефективність, економить час і підвищує доступність для ширшого кола користувачів, зокрема і для людей з обмеженими можливостями.

Обробка гнучких голосових запитів має важливе значення для розвитку автоматизованих систем, оскільки вона забезпечує більш природний та інтуїтивний спосіб взаємодії з системами, значно покращуючи користувацький досвід. Гнучкість у розпізнаванні та обробці голосових , що система може розуміти запити незалежно від того, як вони сформульовані, що робить систему більш адаптованою до різних користувачів і ситуацій.

Обробка гнучких голосових запитів має ряд переваг, які дозволяють уникнути низки проблем, що обмежують продуктивність та доступність традиційних автоматизованих систем. До них відносяться:

- зменшення залежності від формальних команд (здатність розпізнавати команди в різних формах, що зменшує потребу в чітких формулюваннях);
- забезпечення доступності для різних користувачів (система здатна адаптуватися до різних акцентів, діалектів та стилів мовлення);
- покращення продуктивності в багатозадачних середовищах (голосові запити надають можливість взаємодіяти з не відволікаючись від основного завдання, що є критично важливим).

Розвиток обробки гнучких голосових запитів сприяє створенню автоматизованих систем, які стають більш адаптивними та ефективними, покращуючи взаємодію користувачів з технологією та роблячи її доступною для широкого застосування в різних сферах.

Інтеграція гнучких голосових запитів у різні сфери життя спрощує виконання завдань і підвищує продуктивність. Це стає можливим завдяки зручності та швидкості взаємодії з технологіями. Використання голосових запитів не тільки оптимізує робочі процеси, але й зменшує навантаження на користувачів, дозволяючи їм зосередитися на більш важливих завданнях.

У таблиці 1.3 представлена інформація про різні сфери, де можуть використовуватися гнучкі голосові запити, а також детально описаний опис їх застосування.

Таблиця 1.3 – Сфери застосування гнучких голосових запитів

Сфера застосування	Використання гнучких голосових запитів
Бізнес та управління проєктами	Керівники та працівники можуть швидко створювати, оновлювати та переглядати завдання, не перериваючи робочий процес, що підвищує ефективність і продуктивність.
Охорона здоров'я	Медичний персонал використовує голосові запити для введення даних про пацієнта, оновлення записів та доступу до інформації під час лікування.
Домашні смарт-системи	Голосові запити дозволяють керувати пристроями в розумному будинку, що спрощує виконання повсякденних завдань і підвищує зручність для користувачів.
Транспорт і логістика	Водії та працівники складів можуть оновлювати інформацію, відстежувати вантажі та планувати маршрути, що підвищує швидкість і точність операцій.
Освіта	Студенти та викладачі використовують голосові запити для пошуку інформації, планування уроків та доступу до освітніх ресурсів, що підвищує ефективність навчання.

1.3 Сучасний стан технологій обробки голосових запитів

Сучасні технології обробки голосових запитів базуються на досягненнях в області розпізнавання мови (ASR), обробки природної мови (NLP) та машинного навчання.

Розпізнавання мови (ASR – Automatic Speech Recognition) – це процес перетворення аудіосигналів у текстові дані, що дозволяє використовувати текст для подальшого аналізу та виконання завдань. Сучасні технології ASR

дозволяють досягти високої точності розпізнавання навіть у складних умовах, таких як наявність фонового шуму або різноманітних акцентів. Це досягається завдяки глибоким нейронним мережам, які підвищують продуктивність і адаптивність системи.[4]

Обробка природної мови (NLP – Natural Language Processing) дозволяє аналізувати текстові запити, розпізнавання наміри користувача та визначати відповідні дії для виконання завдань. Використання передових моделей, таких як BERT і GPT, значно покращує здатність системи розуміти контекст, дозволяючи їй ефективніше обробляти складні запити. Це також покращує семантичну обробку, забезпечуючи точну інтерпретацію запитів користувача.[5]

Генерування природної мови (NLG – Natural Language Generation) використовується для створення природних, зрозумілих відповідей на запити користувачів. Завдяки цій технології системи можуть формулювати більш людські та релевантні відповіді, що підвищує якість взаємодії та зручність користування. NLG дозволяє адаптувати відповіді під конкретного користувача та контекст, створюючи враження природної спілкування.[6]

Машинне навчання, особливо глибоке навчання (NLP), є основою для розвитку систем розпізнавання мови та обробки природної мови. Використання моделей машинного навчання дозволяє системам адаптуватися до нових даних, таких як різні акценти або нестандартні мовленнєві моделі, що значно покращує точність розпізнавання та взаємодію з користувачем. Глибокі нейронні мережі можуть обробляти великі обсяги даних і навчатися на попередніх прикладах, забезпечуючи все більш точні результати.[7]

У таблиці 1.4 наведено приклади реальних завдань, пов’язаних із сучасними технологіями обробки голосових запитів, а також конкретні продукти, що використовуються для їх реалізації. Ця таблиця демонструє різноманітність застосувань, включаючи розпізнавання мови (ASR), обробку природної мови (NLP) та генерацію природної мови (NLG).

Таблиця 1.4 – Приклади задач та продуктів у сфері обробки голосових запитів

Категорія	Задача	Опис	Застосування/ Приклади
Розпізнавання мови (ASR)	Перетворення голосових запитів на текст	Процес, під час якого голосові сигнали перетворюються на текстові дані, що дозволяє системам розуміти команди або запити користувачів.	Віртуальні асистенти, автоматизовані кол-центри, системи диктування тексту. Приклади: Google Assistant, Apple Siri, Amazon Alexa
	Визначення акценту та мови	Розпізнавання мови та акценту користувача для підвищення точності розпізнавання, адаптація системи до особливостей мови та діалекту.	Міжнародні сервіси з підтримкою різних мов, системи для навчання мов. Приклади: Microsoft Azure Speech Service, IBM Watson Speech to Text
	Розпізнавання команд в умовах шуму	Здатність системи розпізнавати голосові команди за наявності фонового шуму, наприклад, розмови або музики.	Голосові системи в автомобілях, навушники з активним шумозаглушенням. Приклади: Apple AirPods Pro, Bose Noise Cancelling Headphones

Продовження таблиці 1.4

Категорія	Задача	Опис	Застосування/ Приклади
Обробка природної мови (NLP)	Аналіз намірів користувача	Визначення намірів користувача на основі текстового запиту, що дозволяє системі вибрати відповідні дії.	Чат-боти, системи підтримки клієнтів, рекомендаційні системи. Приклади: Dialogflow (Google), Microsoft LUIS
	Витягнення сутностей	Виявлення і класифікація ключових інформаційних одиниць (імен, дат, місць тощо) з тексту, що допомагає системі зрозуміти важливі деталі.	Системи автоматичного аналізу текстів, резюме документів, юридичні системи. Приклади: Amazon Comprehend, SpaCy
	Аналіз настроїв	Визначення емоційного забарвлення тексту (позитивне, негативне, нейтральне), що допомагає зрозуміти відгуки користувачів.	Дослідження ринку, соціальні мережі, аналітика зворотного зв'язку. Приклади: Sentiment Analysis API (Google Cloud), Lexalytics

Продовження таблиці 1.4

Категорія	Задача	Опис	Застосування/ Приклади
Генерація природної мови (NLG)	Автоматичне створення відповідей	Генерація природних та зрозумілих відповідей на запити користувачів на основі аналізу тексту і контексту.	Чат-боти, системи підтримки клієнтів, автоматичні системи новин. Приклади: OpenAI ChatGPT, Google Bard
	Створення описів продуктів	Автоматизоване написання описів товарів або послуг на основі характеристик, що підвищує ефективність контенту.	Інтернет-магазини, платформи продажу. Приклади: Copy.ai, Jasper
	Генерація контенту для маркетингових кампаній	Автоматичне створення рекламних слоганів або постів у соціальних мережах, що дозволяє заощаджувати час на розробку контенту.	Маркетингові агенції, платформи для управління соціальними мережами. Приклади: Writesonic, Phrasee

1.4 Недоліки та обмеження існуючих рішень

Серед основних аналогів, що підтримують функції голосових асистентів, найпопулярнішими є рішення, розроблені провідними технологічними компаніями. Ці асистенти відрізняються за

функціональністю, рівнем інтеграції з екосистемою пристроїв, а також можливостями обробки голосових запитів та інтерпретації намірів користувача. До них належать:

- Apple Siri;
- Google Assistant;
- Amazon Alexa;
- Microsoft Cortana.

Siri – голосовий асистент, розроблений компанією Apple, доступний на всіх пристроях з iOS та macOS. Siri підтримує широкий набір базових команд, таких як пошук інформації, керування налаштуваннями пристрою, а також інтегрується з деякими програмами, наприклад, Apple Music та iMessage. Асистент також може обробляти запити різними мовами і використовувати контекст для розуміння розмовних команд.

На рисунку 1.3 показано інтерфейс Siri: просте, мінімалістичне екранне представлення, що відображає запитання користувача або команди та відповіді, надані асистентом у вигляді тексту та іконок.

Рисунок 1.3 – Інтерфейс голосового помічника Siri

Google Assistant – універсальний голосовий асистент від Google, інтегрований з пристроями на базі Android і Google Home. Він підтримує широкий спектр команд, від пошуку інформації в інтернеті до управління

розумним домом, а також може працювати з Google-сервісами, такими як Календар та Карти. Асистент відомий високою точністю розпізнавання голосу та підтримкою багатьох мов, що робить його популярним серед користувачів різних країн.

На рисунку 1.4 показано інтерфейс Google Assistant, який відображає вікно чату з текстовими відповідями на запитання користувача, а також додатковими пропозиціями команд.

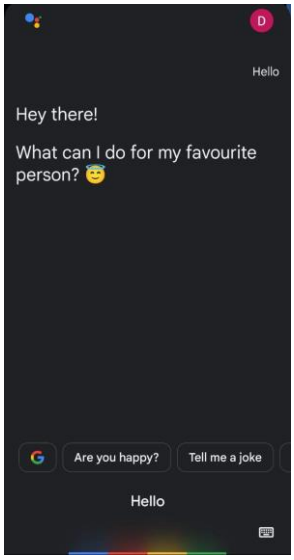

Рисунок 1.4 – Інтерфейс голосового помічника Google Assistant

Amazon Alexa – голосовий асистент від Amazon, який найчастіше використовується на пристроях Echo. Alexa здатна взаємодіяти з іншими розумними пристроями у домі, такими як освітлення або термостати, і підтримує безліч додаткових «навичок» (skills), які дозволяють виконувати різноманітні функції: від замовлення товарів до відтворення музики. Вона добре підходить для керування домашньою автоматизацією.

На рисунку 1.5 показано інтерфейс Amazon Alexa: простий голосовий екран з опціями для активації команд і відображення інформації на розумних дисплеях, включаючи пропозиції команд і варіанти управління іншими пристроями.

Рисунок 1.5 – Інтерфейс голосового помічника Amazon Alexa

Microsoft Cortana – голосовий асистент, розроблений для пристроїв на базі Windows. Основна функція Cortana – допомога в керуванні завданнями, такими як встановлення нагадувань, пошук документів і керування подіями календаря. Cortana добре інтегрується з програмами Microsoft Office та іншими службами Windows. Однак на мобільних пристроях функціональність Cortana обмежена.

На рисунку 1.6 показано інтерфейс Microsoft Cortana: вікно з відповідями на текстові та голосові запити користувача, а також опціями інтеграції з календарем, контактами та іншими додатками Microsoft.

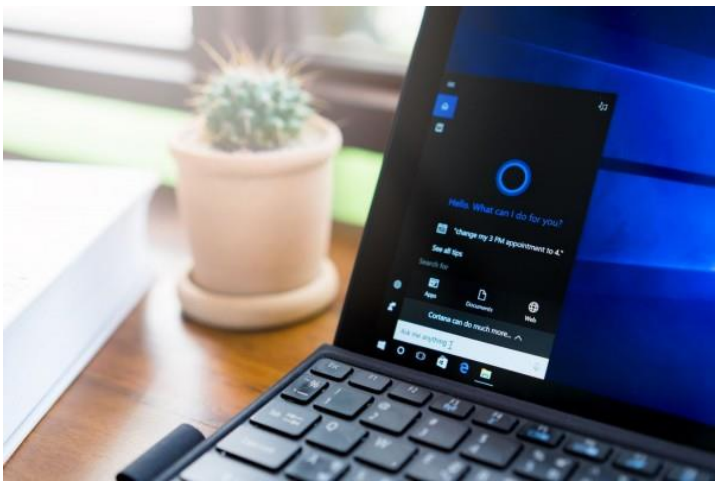

Рисунок 1.6 – Інтерфейс голосового помічника Microsoft Cortana

Незважаючи на значний прогрес у технологіях обробки голосових запитів, існуючі рішення все ще мають певні обмеження, що впливають на їх ефективність та зручність для користувачів.

Одним з таких недоліків є обмежена адаптивність до різних акцентів та мовних особливостей. У голосових асистентах таких, як Siri або Google Assistant, інколи бувають труднощі з розпізнаванням акцентів, наприклад, індійського, ірландського або австралійського. Це може призводити до неправильної інтерпретації запитів або помилок обробки, що знижує якість користувацького досвіду. Відомий випадок стався, коли користувачі з Шотландії повідомили, що Siri часто неправильно інтерпретувала їхні запити через місцевий акцент.

Багато сучасних систем не завжди ефективно обробляють складні запити, які потребують аналізу контексту або містять двозначність. Наприклад, запит “Чи буде зустріч перенесена на наступний тиждень?” може неправильно інтерпретуватися, якщо користувач не уточнить, про яку зустріч йдеться.

Багато голосових систем, таких як Amazon Alexa і Google Assistant, працюють через хмарні сервіси, що вимагає постійного і надійного підключення до інтернету. Це обмежує їхнє використання в умовах відсутності або нестабільного доступу до мережі, особливо у віддалених регіонах, подорожей або блекаутів.

На малопотужних смартфонах, старих моделях або пристроях без спеціалізованого процесора для обробки мови, таких як Digital Signal Processor (DSP), голосові асистенти можуть працювати повільно або не запускатися взагалі. Це особливо актуально для нових моделей голосових систем, таких як Google Assistant, які постійно вдосконалюються і потребують все більше обчислювальних ресурсів для обробки складних запитів.

Ще одним недоліком є питання конфіденційності та безпеки. Відомий інцидент трапився з Amazon Alexa, коли пристрій випадково записав

приватну розмову користувача і відправив її третій стороні. Такі інциденти викликають занепокоєння щодо безпеки, оскільки дані користувачів передаються на віддалені сервери для обробки, що створює потенціал для витоку даних або зловживань.

У медичній сфері, де використовується специфічна термінологія, системи розпізнавання голосу можуть неправильно інтерпретувати медичні терміни або назви ліків. Це може стати критичним фактором для лікарів, які використовують голосові асистенти для створення нотаток чи оновлення інформації про пацієнта.

У багатьох багатомовних країнах, таких як Канада, де використовується англійська та французька, голосовим асистентам часто важко перемикатися між мовами в межах одного запиту. Наприклад, запит “Покажіть мені погоду на сьогодні та de demain” може бути неправильно розпізнаний або система може відповісти лише на одну частину запиту.

Для порівняння голосових асистентів можна розглянути такі критерії, як:

- точність розпізнавання;
- адаптивність;
- стійкість до фонового шуму;
- час реакції;
- вимоги до обчислювальних ресурсів;
- зручність інтеграції;
- конфіденційність та безпека даних;
- можливість локальної обробки.

Точність розпізнавання відображає здатність асистента правильно інтерпретувати запити користувача. Висока точність забезпечує надійне розуміння команд та мінімізує кількість помилкових інтерпретацій. Точність залежить від здатності системи враховувати мовні особливості, акценти та інші мовні варіації. У багатомовних системах важливою є підтримка високої точності для різних мов.

Адаптивність – це здатність системи навчатися та підлаштовуватися під індивідуальні особливості мовлення користувача, такі як унікальний стиль мовлення, часті запити та особисті вподобання. Адаптивність також включає здатність асистента розпізнавати контекст діалогу, зберігати його для подальших запитів і враховувати історію взаємодії для надання більш релевантних відповідей.

Завадостійкість оцінює, наскільки ефективно асистент розпізнає голос користувача шумному середовищі або серед інших звукових перешкод. Висока завадостійкість важлива для використання асистента в реальних умовах таких як вулиця, офіс чи під час багатозадачності. Це досягається завдяки алгоритмам фільтрації шуму та ізоляції голосу від фонових звуків.

Час відгуку – це швидкість, з якою асистент реагує на запити користувача. Висока швидкість обробки та швидка відповідь важливі для асистентів, що працюють у режимі реального часу, мінімізуючи затримки та підвищуючи зручність для користувача. Час відгуку залежить від алгоритмів, обчислювальної потужності пристрою і здатності обробляти локальні запити.

Вимоги до обчислювальних ресурсів оцінюють потужність, необхідну для стабільної роботи асистента. Системи, які виконують складну обробку мови і глибокий аналіз запитів, часто вимагають потужного процесора або графічного процесора. Водночас менш ресурсомісткі асистенти можуть працювати на більш доступних пристроях з обмеженою обчислювальною потужністю.

Легкість інтеграції визначає здатність асистента з’єднуватись з іншими програмами та пристроями. Асистент з високим рівнем інтеграції дозволяє легко взаємодіяти з різними сервісами (наприклад, календарі, месенджерами) та іншими додатками. Асистенти, інтегровані в екосистему певного бренду (наприклад, Apple або Google), зазвичай краще працюють з пристроями того ж бренду.

Конфіденційність і безпека даних оцінюють, наскільки добре захищена інформація користувача під час використання асистента. Багато голосових

асистентів використовують хмарні сервери для обробки даних, що може збільшити ризик витоку або несанкціонованого доступу. Високі стандарти конфіденційності передбачають локальну обробку даних і належне шифрування переданої інформації.

Можливість локальної обробки даних має вирішальне значення в ситуаціях, коли стабільне інтернет-з'єднання недоступне або коли дані користувачів не повинні надсилатися на хмарні сервери. Асистенти з можливістю локальної обробки можуть виконувати основні функції навіть без доступу до Інтернету, зберігаючи конфіденційність і скорочуючи час відгуку.

У таблиці 1.5 представлено порівняння основних голосових асистентів за критеріями, важливими для автоматизованих систем обробки гнучких голосових запитів.

Таблиця 1.5 – Порівняння голосових асистентів

Критерій	Apple Siri	Google Assistant	Amazon Alexa	Microsoft Cortana
Точність розпізнавання	Висока точність для англійської, середня точність для інших мов	Відмінна точність для багатьох мов	Точне розпізнавання в умовах низького рівня шуму	Достатня точність для основних команд
Адаптивність	Обмежена адаптивність до певних типових команд	Висока гнучкість контекстних запитів	Гнучке налаштування, але лише в межах екосистеми Amazon	Добре інтегрується з програмами Microsoft

Продовження таблиці 1.5

Критерій	Apple Siri	Google Assistant	Amazon Alexa	Microsoft Cortana
Час відгуку	Швидка реакція на пристроях Apple	Миттєва реакція, особливо на Android	Час відгуку залежить від пристрою	Плавна відповідь на Windows
Завадостійкість	Середня стійкість до шуму	Ефективно розпізнає голос навіть при фоні	Добре розпізнає команди в умовах контрольова-ного шуму	Низька
Вимоги до обчислюваль-них ресурсів	Оптимізова-но для пристроїв iOS	Потребує середніх ресурсів для забезпечення стабільності	Потребує більших ресурсів для повноцінної роботи	Підтримує мінімальні вимоги для Windows
Зручність інтеграції	Глибока інтеграція з продуктами Apple	Широка сумісність з екосистемою Google	Повна інтеграція з продуктами Amazon	Відмінно взаємодіє з програмами Microsoft

Продовження таблиці 1.5

Критерій	Apple Siri	Google Assistant	Amazon Alexa	Microsoft Cortana
Конфіденційність та безпека даних	Високий рівень захисту даних користувачів	Забезпечує керування налаштуваннями безпеки.	Використовує хмару, що підвищує ризик для безпеки даних.	Підвищена безпека для корпоративного середовища
Можливість локальної обробки	Підтримує локальну обробку на iOS	Обробка без Інтернету на Google Pixel	Здебільшого використовує хмару	Основні функції працюють локально на Windows

1.5 Мета та задачі дослідження

Мета роботи полягає у розробці ефективних методів обробки гнучких голосових запитів для автоматизованих систем керування завданнями.

Для досягнення поставленої мети дослідження необхідно вирішити наступні задачі:

- огляд методів розпізнавання голосових команд;
- оцінка ефективності голосового управління на основі існуючих рішень;
- розробка архітектури системи розпізнавання мовлення із гнучкими запитами;
- проведення експериментів оцінки продуктивності ResNet у залежності від рівня шуму та методів виділення ключових ознак;
- дослідження впливу використання направлених в сторону джерела звуку мікрофона на точність детектування голосових команд;

- аналіз результатів експериментів.

Майбутні шляхи розвитку можуть включати впровадження нових алгоритмів обробки природної мови, які зможуть краще адаптуватися до змін у мові та обробляти більш складні запити. Важливим кроком також стане інтеграція технологій для локальної обробки голосових команд, що зменшить залежність від хмарних сервісів і прискорить роботу системи.

Крім того, варто досліджувати можливість покращення точності систем у мультикультурних та багатомовних середовищах, розробляючи моделі, що можуть працювати з різними мовними варіантами та діалектами. У майбутньому також можна зосередити зусилля на підвищенні рівня захисту даних користувачів і конфіденційності під час обробки запитів за рахунок використання сучасних криптографічних методів і технологій децентралізованого зберігання даних.

2 МЕТОДОЛОГІЧНЕ ПІДҐРУНТЯ ДОСЛІДЖЕННЯ

2.1 Огляд технологій, методів і алгоритмів

Для реалізації системи обробки гнучких голосових запитів необхідно використати кілька ключових технологій та алгоритмів, кожен із яких відповідає за певні аспекти процесу розпізнавання, аналізу та обробки мовлення і тексту.

Методи машинного навчання є фундаментом сучасних систем розпізнавання мовлення та аналізу голосових запитів. Вони дозволяють автоматично виявляти закономірності в даних, аналізувати їх і прогнозувати. У контексті голосових запитів це означає, що система навчається розуміти аудіодані, асоціювати певні мовленнєві патерни з відповідними текстовими інтерпретаціями та належним чином реагувати на запити користувача.

Ці методи покладаються на алгоритми класифікації та регресії для ідентифікації мовних характеристик (наприклад, тембр, акцент, швидкість мовлення) та зв'язування їх з текстом. Під час навчання моделі використовують великі бази даних аудіозаписів і відповідних текстових транскрипцій, що дозволяє машині поступово покращувати свої результати і підвищувати точність розпізнавання. Сучасні системи використовують такі алгоритми, як градієнтний бустинг, метод опорних векторів (SVM) та ансамблі дерев рішень для класифікації мовних сигналів.

На рисунку 2.1 зображено основні методи обробки голосових запитів.

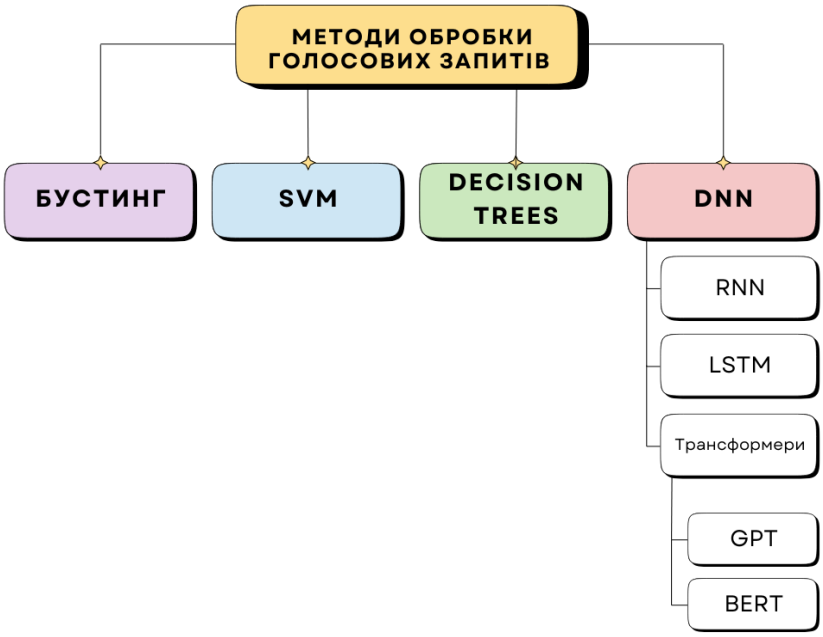

Рисунок 2.1 – Основні методи обробки голосових запитів

Бустинг – це ансамблевий метод машинного навчання, який об'єднує кілька слабких моделей (класифікаторів або регресорів) для створення одного сильного класифікатора. Ідея бустингу полягає в послідовному навчанні моделей, де кожна модель фокусується на виправленні помилок своїх попередників.

Основні етапи бустингу включають:

- призначення вагової функції для кожної моделі в ансамблі, щоб збалансувати її внесок;

- Навчання кожної наступної моделі фокусуватися на точках даних, неправильно класифікованих попередньою моделлю.

Метод опорних векторів використовує гіперплощину для відокремлення класів даних у просторі ознак. Він будує гіперплощину, яка максимізує відстань між найближчими точками (опорними векторами) різних класів:

$$\omega \cdot x - b = 0,$$

де ω – вектор ваг, x – вхідні дані, а b – зміщення.

Оптимізація цього методу полягає в максимізації відстані до гіперплощини з обмеженням, що точки різних класів розташовані по різні сторони від неї.

Дерева рішень спираються на ієрархічну структуру вузлів, де кожен вузол представляє тест певної умови або атрибуту, а кожна гілка – результат. Математичний принцип дерев рішень полягає в тому, щоб зменшити ентропію або приріст інформації при кожному розбитті:

$$IG(T, X) = H(T) - H(T|X),$$

де IG – приріст інформації, H(T) – ентропія до розбиття, H(T|X) – ентропія після розбиття по ознаці X.

Глибокі нейронні мережі (DNN) є однією з найефективніших технологій для розпізнавання мовлення та контексту голосових запитів. DNN складаються з багатьох шарів штучних нейронів, які здатні обробляти складні залежності в даних. У контексті розпізнавання мовлення, DNN навчаються перетворювати звукові сигнали (акустичні особливості) у текстові послідовності.

Зокрема, системи розпізнавання мови використовують рекурентні нейронні мережі (RNN) або їхню вдосконалену версію – довготривалу короткочасну пам'ять (LSTM), здатну враховувати послідовність звуків та попередній контекст. Ця здатність дозволяє моделям точно розпізнавати навіть складні фрази або слова, які залежать від попередніх звукових сигналів. Архітектури трансформерів, що забезпечують паралельну обробку послідовностей, дозволяють ще більше підвищити ефективність систем, особливо в реальному часі.

Основною перевагою DNN є їхня адаптивність до різних умов мовлення (шуму, акцентів) завдяки структурі глибокого навчання, що дозволяє ефективно обробляти складні звукові дані. Це дозволяє створювати

точні системи ASR, які можуть працювати з різними користувачами та умовами.

Обробка природної мови (NLP) відіграє вирішальну роль у розумінні та інтерпретації тексту, який є результатом перетворення голосових запитів. Моделі NLP допомагають аналізувати семантику та синтаксис запитів і розуміти контекст і наміри користувача. Сучасні моделі NLP, такі як BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pretrained Transformer) та інші, базуються на архітектурі трансформерів, що дозволяє враховувати контекст не тільки окремих слів, але й усього тексту.

BERT – це двонаправлена модель, яка одночасно аналізує слова в контексті попередніх і наступних слів. Це дозволяє їй розуміти багатозначність слів і складні граматичні структури. Наприклад, слово може мати різне значення залежно від контексту, і BERT здатний враховувати це при обробці запитів. GPT, у свою чергу, є генеративною моделлю, яка може не тільки інтерпретувати запити, але й генерувати відповіді на основі контексту. Це робить GPT корисною для створення діалогових систем або чат-ботів.

Використання моделей NLP забезпечує ефективне розпізнавання та інтерпретацію голосових запитів, що дозволяє системі точно реагувати на складні або неоднозначні запити користувачів.

Обробка може відбуватися як на локальних обчислювальних станціях, так і з використанням ресурсів хмарних обчислень, забезпечуючи гнучкість та ефективність залежно від потреб системи.

Хмарні обчислення дозволяють обробляти великі обсяги даних у режимі реального часу. Сучасні системи розпізнавання мовлення та NLP вимагають значних обчислювальних ресурсів для роботи з великими масивами даних і виконання складних обчислювальних завдань, таких як розпізнавання мовленнєвих патернів, навчання нейронних мереж або обробка природної мови. Хмарні платформи, такі як Google Cloud, AWS або Microsoft

Azure, дозволяють розподіляти обчислювальні навантаження та забезпечувати масштабованість системи.

Хмарні обчислення дозволяють інтегрувати гнучкі системи розпізнавання мовлення з реальними додатками, обробляючи дані в режимі реального часу навіть при високих навантажень. Це особливо важливо для систем, які обробляють великі потоки даних або потребують високої продуктивності для точного і швидкого аналізу запитів користувачів. Хмарні ресурси надають можливість навчати великі моделі машинного навчання та підтримувати високу обчислювальну потужність для роботи з глибокими нейронними мережами та NLP.

В таблиці 2.1 наведено порівняння переваг та недоліків використання локальних обчислювальних станцій на основі CPU і GPU, а також віддалених хмарних серверів для обробки даних.

Таблиця 2.1 – Переваги і недоліки локальних обчислювачів на CPU, GPU і хмарних сервісів

Тип обчислення	Переваги	Недоліки
Домашні обчислювачі (CPU)	Низька вартість використання (після придбання обладнання); легка доступність; відсутність затримок у мережі.	Низька потужність для паралельних обчислень; обмежена масштабованість.
Домашні обчислювачі (GPU)	Висока продуктивність для паралельних задач (глибоке навчання, обробка зображень); підходить для моделювання в реальному часі; легкий доступ до даних.	Висока початкова вартість на придбання GPU; обмежена продуктивність для дуже великих моделей.

Продовження таблиці 2.1

Тип обчислення	Переваги	Недоліки
Хмарні сервери	Практично необмежена масштабованість ресурсів; доступ до найсучасніших GPU та TPU; можливість обробляти великі дані в режимі реального часу; оплата лише за використані ресурси.	Постійна залежність від інтернету; затримки при передачі даних; висока вартість для великих обчислень.

2.2 Аналіз фреймворків, бібліотек та апаратного забезпечення

Для створення системи обробки гнучких голосових запитів необхідно використовувати комплекс технологій, що забезпечують ефективне розпізнавання мови, аналіз тексту та інтеграцію моделей машинного навчання.

Такі фреймворки, як TensorFlow та PyTorch, широко використовуються для побудови та навчання моделей глибокого навчання. Вони дозволяють створювати моделі автоматичного розпізнавання мовлення (ASR) та обробки природної мови (NLP). Вибір між ними залежить від зручності використання та конкретних вимог проєкту, оскільки обидва підтримують велику кількість попередньо навчених моделей і бібліотек для різних завдань. Kaldi – це фреймворк відкритим вихідним кодом, який надає інструменти для побудови високопродуктивних систем ASR, дозволяючи гнучко налаштовувати моделі та обробляти аудіофайли. DeepSpeech, заснований на глибокому навчанні, спрощує процес перетворення мовлення в текст завдяки наскрізному підходу (end-to-end).[8][9]

Бібліотеки для обробки мови і тексту можуть інтегрувати розпізнавання та аналіз голосових даних. SpeechRecognition дозволяє

перехоплювати голосовий потік через мікрофон і перетворювати його в текст, підтримуючи різні системи розпізнавання мовлення, такі як Google Speech API або Microsoft Bing Voice Recognition. Бібліотеки NLP, такі як NLTK або spaCy, надають можливості обробки природної мови. NLTK орієнтована на академічні дослідження та базову аналітику NLP, тоді як spaCy зосереджена на високопродуктивних реальних додатках, підтримуючи складні моделі для токенізації, синтаксичному та семантичному аналізу, розпізнавання іменованих сутностей тощо. Hugging Face Transformers надає доступ до широкого спектру попередньо навчених моделей, таких як BERT і GPT, що значно спрощує процес інтеграції потужних функцій NLP для обробки текстових запитів і генерації відповідей. [10]

Високочутливі професійні мікрофони, такі як Shure SM7B або Blue Yeti, можна використовувати для якісного збору аудіоданих, забезпечуючи чистий звук з мінімальним рівнем шуму, що підвищує точність розпізнавання мови. Аудіоінтерфейси, такі як Focusrite Scarlett, забезпечують якісну передачу аудіосигналів на комп'ютер для подальшої обробки. Для навчання великих моделей машинного навчання можна використовувати сервери з високою обчислювальною потужністю, наприклад, з графічними процесорами NVIDIA Tesla, що дозволяє значно прискорити навчання на великих наборах даних.

Загальний процес створення системи включає кілька етапів. Спочатку проводиться збір аудіоданих за допомогою мікрофонів та аудіоінтерфейсів. Потім аудіосигнал перетворюється в текст за допомогою фреймворків для автоматичного розпізнавання мовлення, таких як Kaldi або DeepSpeech. Після цього текст обробляється за допомогою NLP-бібліотек, таких як spaCy або NLTK. Нарешті, для аналізу та інтерпретації запитів використовуються попередньо навчені моделі, надані Hugging Face Transformers, або користувальницькі моделі, створені за допомогою TensorFlow чи PyTorch. [11]

2.3 Дискретизація голосового сигналу

Дискретизація мовного сигналу є критично важливим етапом у процесі обробки мовлення, що передбачає перетворення безперервного звукового сигналу в цифровий формат, який може бути проаналізований комп'ютерними системами. Цей процес дозволяє ідентифікувати ключові звукові характеристики, необхідні для точного розпізнавання мови, і складається з декількох етапів, на кожному з яких виконуються певні операції.

Одним з найбільш поширених методів представлення мовного сигналу в цифровому вигляді є обчислення Мел-частотних кепстральних коефіцієнтів (MFCC). Цей метод передбачає низку послідовних етапів обробки сигналу, кожен з яких додає нову важливу інформацію, що підвищує ефективність розпізнавання мови.[12]

На рисунку 2.2 зображено процес обробки голосового сигналу для отримання мел-частотних кепстальних коефіцієнтів.

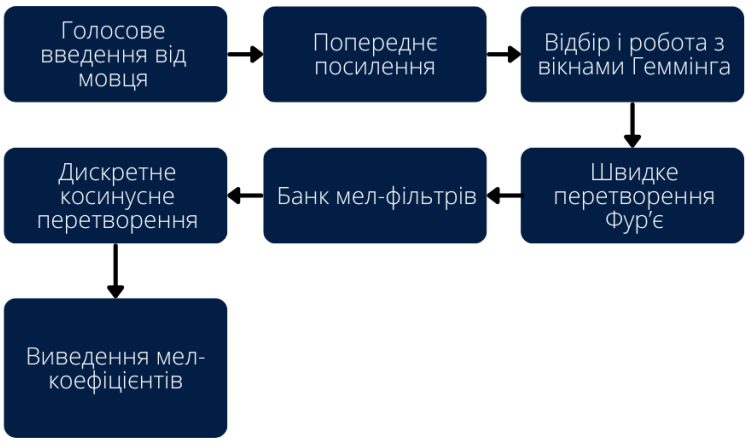

Рисунок 2.2 – Процес обробки сигналу за допомогою MFCC

Перш ніж перетворити сигнал, його потрібно попередньо посилити. Цей етап важливий для підвищення чутливості системи до аудіосигналів, особливо якщо вихідний сигнал має низьку амплітуду. Для цього

використовуються фільтри, які виділяють високі частоти, сприяючи кращій якості подальшої обробки.

Далі сигнал розбивається на кадри, де безперервний сигнал сегментується на короткі інтервали по 20-30 мс. Така сегментація дозволяє аналізувати мову окремими короткими сегментами, причому кожен кадр містить достатньо інформації для розпізнавання частотних характеристик мови. Така сегментація дозволяє системі ігнорувати незначні варіації, які можуть виникати між окремими звуками мови.

Для кожного кадру застосовується віконна функція Геммінга, яка згладжує сигнал на початку і в кінці кожного кадру, мінімізуючи розриви між ними. Це дозволяє уникнути спотворень, які можуть виникнути через різкі розриви сигналу на межі кадру, і зберігає природну якість звуку в його цифровому представленні.

Після цього кожен кадр перетворюється з часової області в частотну за допомогою швидкого перетворення Фур'є (FFT). Цей крок дозволяє ідентифікувати частотні компоненти, присутні в сигналі, та їхні амплітудні значення. Таке перетворення дозволяє системі розпізнавати окремі частоти в мовному сигналі, що має вирішальне значення для подальшого аналізу мови.

Далі до частотного спектру застосовується шкала Мела – перетворення, яке імітує чутливість людського вуха до різних частот. Більш високі частоти сприймаються у вужчому діапазоні, що допомагає системі краще відтворити слухове сприйняття людини. Для цього спектральна енергія, розрахована на попередньому етапі, проходить через фільтр з набором трикутних фільтрів, розташованих на частотній шкалі, побудованій за шкалою Мела. Результатом цього процесу є «спектр Мела», який представляє спектр сигналу в частотній області, адаптований до характеристик людського слуху.

Останнім кроком є дискретне косинусне перетворення (DCT), яке перетворює спектр Мела в набір коефіцієнтів. Ці коефіцієнти є найбільш важливими характеристиками сигналу і називаються Mel-частотними кепстральними коефіцієнтами (MFCC). Вони представляють ключові

акустичні характеристики мовного сигналу, необхідні для розпізнавання мови, і є одними з найбільш важливих параметрів при цифровому представленні мови.

Таким чином, MFCC забезпечує детальне кодування акустичних характеристик мовного сигналу. Таке представлення дозволяє автоматизованим системам ефективно розпізнавати мову, підвищуючи точність і надійність алгоритмів, що використовуються для обробки голосових запитів. До основних переваг MFCC можна віднести його здатність адаптуватися до різних умов мовлення та зменшувати вплив фонового шуму, що робить цей метод одним з найпоширеніших у сфері розпізнавання голосу.

Ще одним важливим інструментом для аналізу голосових сигналів є спектрограма. У той час як MFCC надає компактне уявлення про частотні характеристики мовлення, спектограма забезпечує детальну візуалізацію змін частотного спектра з плином часу. Вона дозволяє спостерігати, як різні частоти змінюються протягом мовлення, що особливо корисно для розпізнавання мови, аналізу інтонації та визначення емоційного стану мовця.

На рисунку 2.3 показано приклад спектрограми чоловічого голосу.

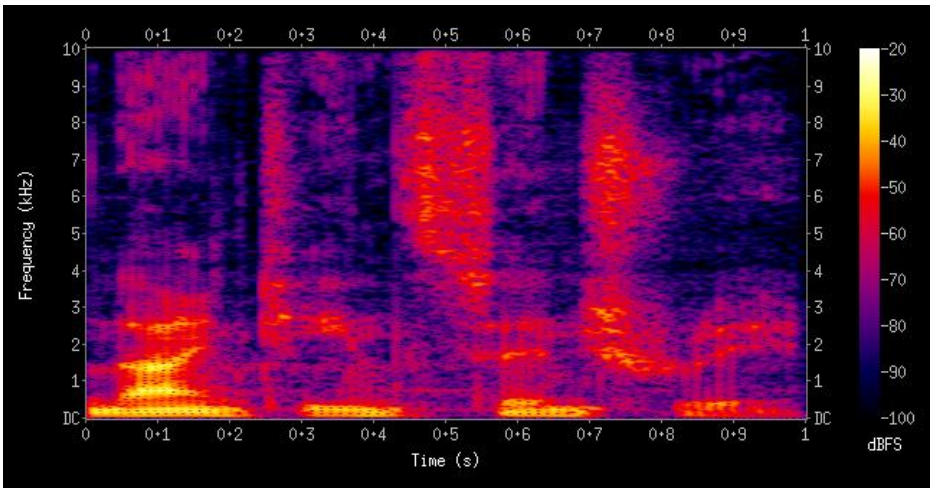

Рисунок 2.3– Спектрограма чоловічого голосу

На спектрограмі горизонтальна вісь представляє час, показуючи, як частотні компоненти змінюються з часом. Вертикальна вісь представляє частоти, що дозволяє нам побачити, які частотні компоненти присутні в сигналі і на якому рівні. Колір або інтенсивність спектрограми вказує на рівень енергії або амплітуду кожної частоти в певний момент часу. Як правило, більш насичені кольори або темні відтінки означають вищу енергію сигналу на певній частоті, тоді як світліші відтінки вказують на нижчі рівні енергії.

Спектрограма зберігає важливі акустичні характеристики сигналу, такі як зміни висоти тону, гучності та тембру, які є критично важливими для розпізнавання мови. Наприклад, у розмовній мові підвищення або зниження висоти тону може сигналізувати про питання або твердження, тоді як зміна інтенсивності може підкреслити емоційні нюанси. Спектрограма також допомагає ідентифікувати такі особливості, як форманти – основні частотні компоненти вимовлених звуків, які мають вирішальне значення для розпізнавання фонем і слів.

В автоматизованих системах розпізнавання мови спектрограма використовується як один з методів представлення акустичних особливостей сигналу. Це допомагає комп'ютерним алгоритмам розрізняти звуки, аналізуючи зміни частотних і енергетичних складових сигналу, що призводить до більш точного розпізнавання слів і розуміння контексту в голосових командах.

3 ПРОЄКТУВАННЯ ТА РЕАЛІЗАЦІЯ СИСТЕМИ ОБРОБКИ ГНУЧКИХ ГОЛОСОВИХ ЗАПИТІВ

3.1 Speech Recognition Software

Програмне забезпечення розпізнавання мовлення (Speech Recognition Software) – це технологія, яка дозволяє сприймати усні висловлювання природною мовою та перетворювати їх у текстовий формат із високою точністю. Цей процес реалізується завдяки використанню сучасних методів штучного інтелекту, алгоритмів машинного навчання та інструментів обробки природної мови.

Програмне забезпечення для розпізнавання мовлення виконує когнітивну функцію, що імітує людські дії. Подібно до того, як люди розуміють мовлення, запам’ятовують почуте та реагують відповідним чином, ця технологія надає машинам аналогічні можливості.

На рисунку 3.1 зображено процес розпізнавання мовлення.

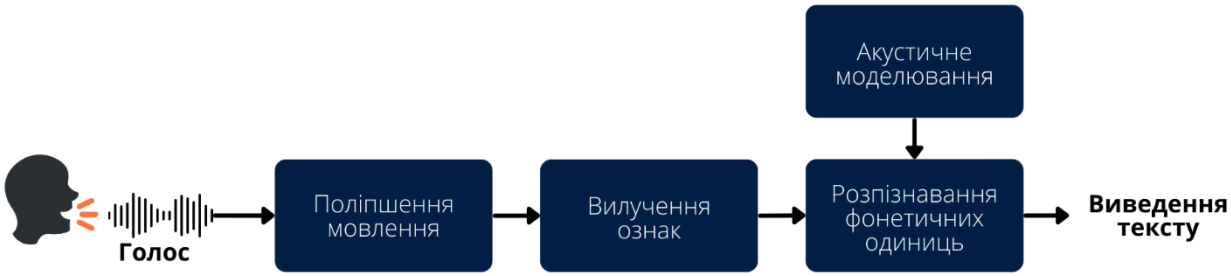

Рисунок 3.1 – Процес розпізнавання мовлення

До кращих сервісів розпізнавання мовлення відносяться Alibaba Cloud Intelligent Speech Interaction [13], Amazon Transcribe [14], Nuance Dragon [15], Deepgram [16] та Google Speech-to-Text API [17]. У таблиці 3.1 наведено порівняння цих сервісів, де розглядаються їх пропозиції, точність розпізнавання мовлення, ціноутворення та підсумкові висновки.

Таблиця 3.1 – Порівняння найкращих програм для розпізнавання мовлення

	Пропозиція	Точність	Ціноутворення	Підсумок
Alibaba Cloud Intelligent Speech Interaction	Використовує інноваційну технологію декодування з низькою частотою кадрів, що суттєво зменшує час відгуку, зберігаючи високу точність.	Компанія не розкриває точний рівень точності, проте сервіс самонавчається за своєю суттю	Вартість становить від 1,00$ за годину для обробки записаних файлів і від 1,40$ за годину для розпізнавання мовлення в режимі реального часу.	Платформа пропонує широкий набір функцій і добре підходить для розпізнавання коротких фраз. В той же час перше знайомство із середовищем може бути доволі складним.
Amazon Transcribe	Приділяє особливу увагу безпеці, конфіденційності та відповідності стандартам. Тобто, для таких галузей як охорона здоров'я впроваджуються спеціальні заходи.	Близько 80%	Безкоштовні 60 хвилин на місяць і після коштує 0.00780$ за хвилину.	Забезпечує високий рівень налаштувань, проте інтеграція у вашу систему може вимагати значних зусиль.

Продовження таблиці 3.1

	Пропозиція	Точність	Ціноутворення	Підсумок
Nuance Dragon	Легкий у використанні та впровадженні, а також непогано підходить для бізнес-користувачів.	Забезпечує 99% точності	Від 500$ за окреме видання	Є лідером у сегменті ПЗ штучного інтелекту та розпізнавання мовлення. Однак користувачі відзначають можливі проблеми з пунктуацією.
Deepgram	Забезпечує найвищу швидкість транскрипції в галузі, дозволяючи обробити годинний запис всього за три секунди.	Понад 90% точності при навчанні моделі.	Починаються від 0,0125$ за хвилину.	Відрізняється високою масштабова-ністю та можливістю локального розгортання, проте його використання є обмеженим у сценаріях безконтактних центрів.

Продовження таблиці 3.1

	Пропозиція	Точність	Ціноутворення	Підсумок
Google Speech-to-Text API	Забезпечує функції шумозаглушення, багатоканального розпізнавання та фільтрації нецензурної лексики, що спрощує навчання моделі та роботу розробників.	Близько 80-85%	Безкоштовні перші 60 хвилин, надалі 0,004 за кожні 15 секунд.	Здатна розпізнавати мовлення у складних середовищах, але для початку потрібні технічні навички, зокрема контейнерне розгортання.

Більшість із цих сервісів забезпечують високу точність розпізнавання мовлення та широкий функціонал, що робить їх ефективними інструментами для різних задач. Однак використання цих сервісів накладають певні обмеження з кількох причин. По-перше, вони є запатентованими, що може ускладнити їх застосування у певних випадках. По-друге, всі ці сервіси мають платну основу, що робить їх менш доступними для проектів з обмеженим бюджетом або для тих, кому потрібні безкоштовні альтернативи.

Тому виникає потреба у створенні власних безкоштовних і незапатентованих систем. Особливістю таких систем має стати здатність адаптуватися до гнучких голосових запитів, забезпечуючи доступність та ефективність для ширшого кола користувачів.

Крім того, розробка власних систем розпізнавання мовлення відкриває можливість глибшої інтеграції з конкретними галузевими рішеннями та внутрішніми бізнес-процесами. Це дозволяє враховувати специфіку певної

сфери діяльності, налаштовувати систему під вузькопрофільну термінологію та забезпечувати вищий рівень конфіденційності даних, оскільки обробка інформації може здійснюватися локально без залучення сторонніх сервісів. Такий підхід сприяє підвищенню продуктивності, зменшенню витрат і кращому контролю за якістю обробки голосових запитів.

3.2 ASR сфери та визначення їх проблем

ASR використовується в багатьох сферах: від створення в реальному часі субтитрів у соцмережах до більш важливих галузей, таких як медицина, банкінг, енергетика та багато інших.

Попри значні відмінності у системах охорони здоров’я, США та Великобританія стикаються з проблемою тривалого очікування на медичну допомогу. Технології розпізнавання мовлення можуть допомогти скоротити цей час, надаючи лікарям змогу швидше створювати нотатки, переводячи мовлення у текст замість ручного введення. Це дозволяє збільшити кількість пацієнтів, яких лікар може прийняти за день.

Крім того, передові системи розпізнавання мовлення, як-от автоматичне розпізнавання мовлення (ASR), дають змогу усунути потребу в посередниках. Багато медичних установ уже використовують ці технології для визначення симптомів і з'ясування, чи потрібна пацієнтові консультація лікаря.

Однак виникають питання щодо безпеки даних, які обробляються такими системами. Для збереження конфіденційності та надійності інформація повинна бути підтверджена відповідними медичними установами.

Незважаючи на ці виклики, розпізнавання мовлення в охороні здоров’я є очевидним кроком вперед. Кожна секунда, зекономлена завдяки таким технологіям, може врятувати життя.

В сфері енергетики завдяки інтеграції голосових асистентів, таких як Alexa чи Google Assistant, енергетичні компанії надають користувачам доступ до інформації про споживання енергії, пропонують поради щодо

економії та підвищують ефективність використання. Наприклад, компанія Octopus дозволяє своїм клієнтам дізнатися, коли електроенергія найдешевша, і планувати енергозатратні завдання на цей час, щоб зменшити рахунки за електроенергію.

Голосові технології також стають важливим інструментом у фінансовій індустрії, покращуючи обслуговування клієнтів і забезпечуючи персоналізацію. Великі банки США, як Bank of America [18], використовують голосових агентів для перевірки балансу, налаштування сповіщень і інших завдань, тоді як інноваційні банки, як Atom [19], застосовують голосову біометрику для підвищення безпеки. Зростаюча довіра до технології сприяє популярності голосових платежів, особливо для оплати невеликих покупок і підписок.

На рисунку 3.2 зображено рівень поширення голосових платежів у Сполучених Штатах Америки. Ця статистика відображає динаміку впровадження голосових платежів у США у період з 2017 по 2022 роки [20].

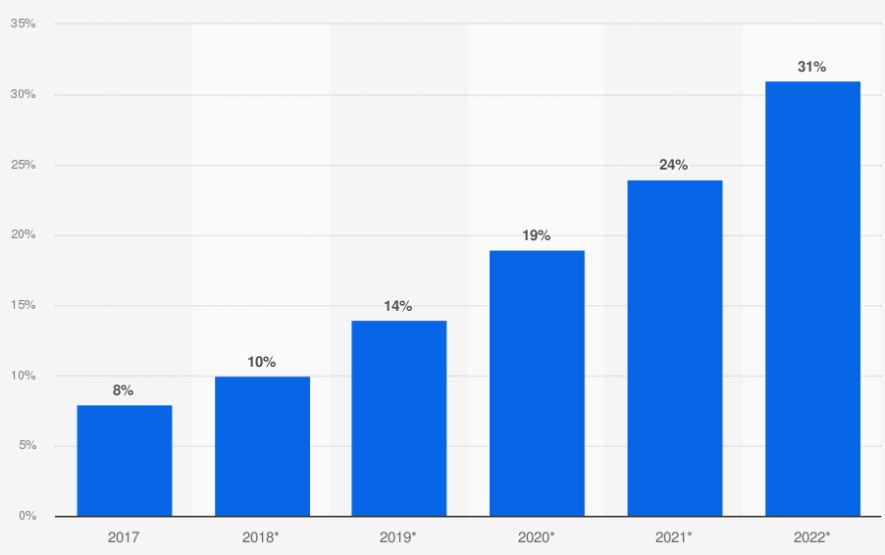

Рисунок 3.2 – Рівень поширення голосових платежів у США

В Україні голосову біометрію запровадив ПриватБанк, що дозволяє клієнтам підтверджувати свою особу за допомогою голосу. Ця технологія підвищує безпеку та зручність, оскільки голос кожної людини є унікальним,

що унеможливлює несанкціонований доступ. Для активації голосової біометрії клієнтам необхідно записати свій голосовий зразок, який надалі використовуватиметься для ідентифікації під час звернень до банку. Це спрощує процес автентифікації та зменшує потребу у використанні традиційних методів підтвердження особи [21].

Попри всі досягнення технології автоматичного розпізнавання мовлення (ASR), вона досі стикається з низкою важливих проблем.

Основні проблеми ASR:

- лексична неоднозначність;
- фоновий шум;
- варіативність вимови;
- затримки та паузи;
- редукція звуків.

Лексична неоднозначність обумовлена тим що одні й ті ж самі слова можуть мати різні значення залежно від контексту. Тобто, слово коса може означати сільськогосподарське знаряддя або ж сплетене волосся. ASR-система може неправильно інтерпретувати слово, якщо не враховує контекст.

Розпізнавання мовлення погіршується в умовах шумного середовища. Тобто, голосові команди можуть бути спотворені звуками вітру, музики чи розмов інших людей. Це призводить до того що ASR може неправильно розпізнати команду або взагалі не зрозуміти.

Варіативність вимови спричинена індивідуальними особливостей вимови людини, які залежать від акценту, діалекту, швидкості мовлення чи емоційного стану. Наприклад, англійське слово "water" з британським акцентом може звучати як "watah".

Затримки та паузи в живій мові зустрічають доволі часто, це також можуть бути повторення чи вставні слова, по типу "емм", "ну", "типу". Наприклад, фраза "Ну, емм, постав будильник" може бути сприйнята як складна для розпізнавання. ASR має вміти ігнорувати такі заповнювачі мовлення.

Також у розмовній мові люди спрощують вимову слів. Наприклад, замість "Що робиш?" часто звучить "Шо робиш". Такі спрощення ускладнюють розпізнавання мовлення, особливо якщо ASR налаштована на стандартну вимову.

3.3 Гнучкі голосові запити і їх особливості

Гнучкі голосові запити – це голосові команди, які формулюються користувачем у довільній або нестандартній формі, з використанням різних слів, фраз і мовних конструкцій для вираження одного й того ж наміру. Вони відрізняються від чітких та шаблонних команд тим, що не мають фіксованої структури й можуть містити варіативні елементи мовлення.

Основні особливості гнучких голосових запитів:

- можливість користувача формулювати запити у довільній формі без обмеження певними шаблонами чи фразами, наприклад, "Чи тепло буде завтра?" або "Які плани на вечір?";
- здатність системи розпізнавати контекст і розуміти не лише окремі слова, а й їхню логічну взаємодію в реченні, тобто після запиту "Хто був президентом України у 2004 році?" вона зрозуміє продовження "А до нього?";
- підтримка складених запитів, які можуть містити кілька команд або завдань для активації різних функцій системи, наприклад: "Постав будильник на 7 ранку і ввімкни музику для медитації";
- здатність системи коректно обробляти неточні або помилкові запити, враховуючи можливі помилки у формулюванні, тобто на запит "Пісня того, що співав про дорогу" система може запропонувати "Highway to Hell";
- персоналізація взаємодії відповідно до індивідуальних налаштувань і особливостей користувача, наприклад, на запит "Заплануй спортзал після роботи" система автоматично обере час на основі графіка користувача.

3.4 Огляд обраного мовного корпусу для розпізнавання мовлення

Мовний корпус – це структурована база даних, яка включає аудіозаписи мовлення та відповідні текстові транскрипції цих записів. Аудіофайли можуть містити як окремі слова, так і фрази, речення чи діалоги, записані в різних умовах, таких як тиша, шумове середовище або з використанням різних типів мікрофонів. Текстові транскрипції забезпечують точне відображення того, що сказано в аудіофайлах, і можуть містити додаткову інформацію, наприклад, розділові знаки, інтонаційні позначки чи акценти.

Такі корпуси використовуються для створення акустичних моделей – математичних моделей, які описують звукові особливості мовлення. Акустичні моделі є ключовим компонентом систем розпізнавання мови, оскільки вони дозволяють програмам інтерпретувати звукові хвилі і зіставляти їх із відповідними текстовими даними.

Завдяки цьому мовні корпуси відіграють важливу роль у розвитку технологій розпізнавання мови, таких як голосові помічники, автоматичні системи перекладу чи диктування тексту.

Для розпізнавання мовлення було обрано корпус Mozilla Common Voice (MCV). MVC – це відкритий мовний корпус, створений для розвитку технологій розпізнавання мовлення. Його головна перевага полягає у вільному доступі до великої кількості якісних голосових даних, що дозволяє дослідникам і розробникам ефективно використовувати цей ресурс без ліцензійних обмежень. Завдяки різноманітності мовних даних, корпус включає записи 133 мов, що забезпечує широку мовну підтримку й адаптивність систем до користувачів із різних мовних середовищ [22].

Станом на 6 грудня 2024 року загальний обсяг доступних мовних даних у MCV становить 33 150 годин, із яких 22 108 годин були підтверджені спільнотою шляхом краудсорсингової перевірки якості. Такий підхід забезпечує надійність і точність даних. Крім того, постійне оновлення корпусу сприяє його актуальності та зростанню, що робить його ідеальним

ресурсом для дослідження методів обробки гнучких голосових запитів в автоматизованих системах керування завданнями.

3.5 Концепція та архітектура запропонованої системи

Еталонна система розпізнавання мовлення складається з кількох основних етапів, які забезпечують ефективне перетворення мовного сигналу в текстову форму.

Першим етапом є вхідне мовлення, яке надходить до системи у вигляді звукового сигналу. Цей сигнал спочатку проходить через блок попередньої обробки, де він очищається від шумів, нормалізується рівень гучності та видаляються непотрібні фонові звуки. Це покращує якість сигналу для подальшої обробки.

Далі сигнал передається до модуля вилучення ознак, який аналізує акустичні властивості мовлення та виділяє ключові характеристики, такі як частотні спектри чи мел-частотні кепстральні коефіцієнти (MFCC). Ці ознаки є основою для розпізнавання мовних одиниць.

Отримані ознаки надходять до блоку класифікації шаблонів (або декодування). Тут використовуються дві основні моделі: акустична модель та мовна модель. Акустична модель створюється на основі навчальних даних і використовується для перетворення акустичних ознак у ймовірні фонеми або інші мовні одиниці. Мовна модель, яка також навчається на текстових даних, допомагає визначити найімовірніші послідовності слів, покращуючи точність розпізнавання. Додатково застосовується словник, який містить правила вимови слів і пов’язує звукові форми зі словами.

Фінальним етапом є формування розпізнаних слів, які є результатом обробки вхідного мовного сигналу. Така система забезпечує базову, але ефективну обробку мовлення, дозволяючи перетворювати голосові команди або запити у текстовий вигляд.

На рисунку 3.3 зображено архітектуру системи розпізнавання мовлення.

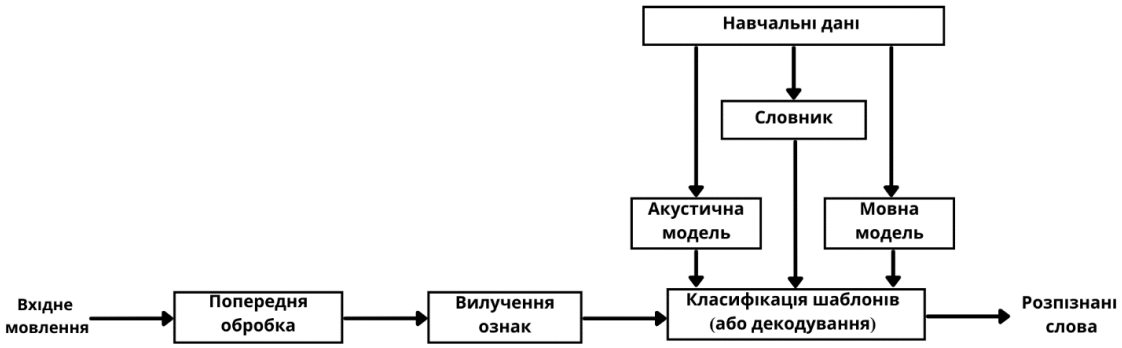

Рисунок 3.3 –Компоненти існуючих систем розпізнавання мовлення

Запропонована система обробки голосових запитів складається з кількох взаємопов’язаних компонентів, які забезпечують ефективне перетворення мовленнєвого сигналу в текстовий формат із подальшою обробкою. На першому етапі система отримує мовленнєвий сигнал від користувача. Цей сигнал надходить до модуля вилучення ознак, де здійснюється попередній аналіз акустичних властивостей мовлення. Отримані акустичні ознаки передаються до акустичної моделі мовних одиниць, яка була попередньо навчена на великому обсязі мовних даних. Акустична модель тісно взаємодіє з декодером мовлення, що аналізує акустичні ознаки та перетворює їх у текстовий формат.

На рисунку 3.4 представлено архітектуру запропонованої системи.

Для підвищення точності розпізнавання мови використовується статистична мовна модель, яка навчається на текстових даних і допомагає передбачити ймовірні послідовності слів. Лексична модель, що містить словник вимови, забезпечує правильне розпізнавання слів із різними варіантами вимови. Одним із ключових компонентів є база даних синонімів, яка значно покращує гнучкість і адаптивність системи. Наприклад, у випадку супермаркету кожен магазин може мати власні найменування товарів: один супермаркет може називати продукт "помідор", інший – "томат". Завдяки базі синонімів система зможе коректно обробити запит користувача незалежно від того, як саме він сформулює запит, і підібрати відповідний товар із конкретної бази даних супермаркету.

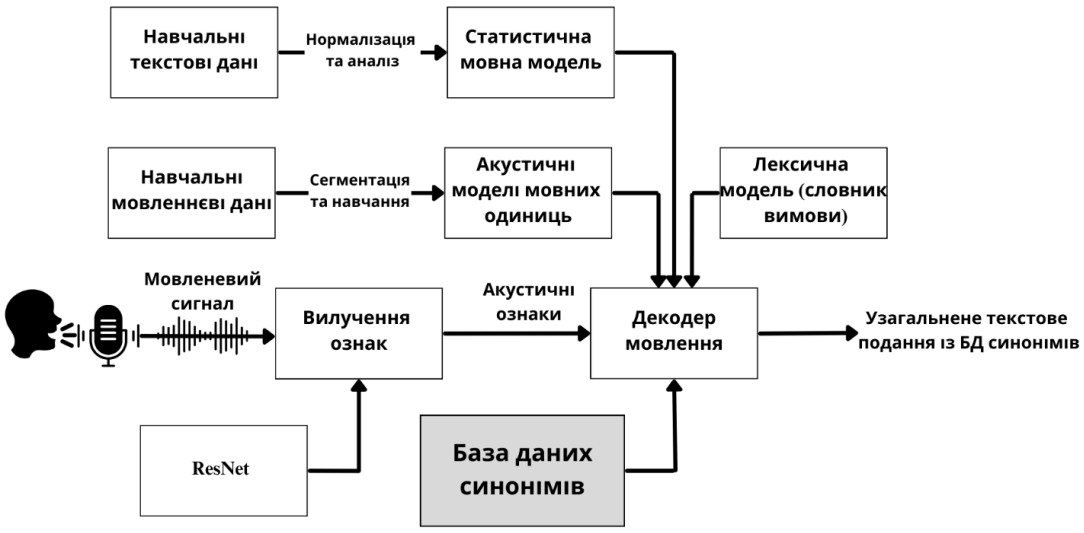

Рисунок 3.4 – Архітектура запропонованої системи розпізнавання гнучких голосових запитів

Використання бази синонімів має низку переваг. По-перше, це підвищує точність розпізнавання, оскільки система враховує варіативність мовлення. По-друге, це забезпечує більшу гнучкість у роботі з різними джерелами даних, адаптуючись до специфіки кожного магазину чи платформи. По-третє, це покращує користувацький досвід, адже користувачам не потрібно підлаштовувати свою мову під систему – вона сама адаптується до їхнього стилю мовлення. У підсумку, така система забезпечує ефективну, точну й зручну обробку голосових запитів у різних сферах застосування.

3.6 Огляд бази даних на прикладі супермаркету

Супермаркет є оптимальним прикладом для демонстрації бази даних синонімів у системах обробки гнучких голосових запитів, оскільки поєднує в собі складну товарну номенклатуру, різноманітність клієнтських запитів і високий попит на автоматизацію. Великий асортимент товарів у супермаркеті передбачає наявність різних назв для однієї й тієї самої продукції. Наприклад, покупці можуть використовувати слова "хліб",

"батон", "булка" або "випічка", маючи на увазі одну товарну категорію. Для ефективного обслуговування система повинна коректно розпізнавати ці синоніми та відповідати на запити незалежно від їх формулювання.

Крім того, клієнти супермаркету можуть по-різному формулювати свої запити. Хтось скаже "покажи знижки", інший – "які акції зараз діють", що вимагає від системи здатності до гнучкої інтерпретації синонімів і контексту. Саме тому впровадження бази синонімів дозволяє забезпечити точне розпізнавання різних варіантів запитів і підвищити ефективність обслуговування.

До того ж супермаркети активно впроваджують автоматизовані системи, такі як голосові асистенти в мобільних додатках, інтерактивні термінали й системи самообслуговування. У цих умовах обробка гнучких голосових запитів стає важливим елементом підвищення зручності для клієнтів і оптимізації бізнес-процесів. Це робить супермаркет ідеальним середовищем для впровадження методів обробки гнучких голосових запитів в автоматизованих системах керування завданнями.

Синоніми певних товарів і послуг в супермаркеті представлено в таблиці 3.3 і лістингу 3.1.

Лістинг 3.1 –База даних синонімів на прикладі розташуваввння асортименту супермаркету

```
CREATE TABLE Synonyms (
    ID INT PRIMARY KEY,
    MainWord VARCHAR(50),
    Synonyms TEXT,
    Category VARCHAR(50)
);

INSERT INTO Synonyms (ID, MainWord, Synonyms, Category) VALUES
(1, 'Хліб', 'Батон, Булка, Випічка', 'Продукти'),
(2, 'Ковбаса', 'Салямі, Сосиски', 'Продукти'),
(3, 'Макарони', 'Вермішель, Паста', 'Продукти'),
(4, 'Олія', 'Масло' 'Соняшникова', 'Продукти'),
(5, 'Десерт', 'Торт, Тістечко, Печиво', 'Продукти'),
(6, 'Фрукти', 'Ягоди, Цитрусові, Плоди', 'Продукти'),
```

```
(7, 'Вода', 'Мінералка, Газованка', 'Напої'),
(8, 'Кава', 'Еспресо, Капучино, Американо', 'Напої'),
(9, 'Корм для тварин', 'Корм, Їжа для котів, Вологий корм',
'Напої'),
(10, 'Пакет', 'Сумка, Кульок', 'Товари'),
(11, 'Доставка', 'Привезти, Доставити', 'Послуга'),
(12, 'Знижка', 'Акція, Розпродаж, Вигода', 'Обслуговування'),
(13, 'Кошик', 'Корзина, Кошелка, Кошик для покупок',
'Обслуговування');
```

В таблиці 3.3 представлено синоніми певних послуг і товарів в супермаркеті.

Таблиця 3.3 – Приклад бази даних синонімів на прикладі супермаркету

ID	Основне слово	Синоніми	Категорія
1	Хліб	Батон, булка, випічка…	Продукти
2	Ковбаса	Салямі, сосиски…	Продукти
3	Макарони	Вермішель, паста…	Продукти
4	Олія	Масло, Соняшникова…	Продукти
5	Десерт	Торт, тістечко, печиво…	Продукти
6	Фрукти	Ягоди, цитрусові, плоди…	Продукти
7	Вода	Мінералка, газованка…	Напої
8	Кава	Еспресо, капучино…	Напої
9	Корм для тварин	Корм, їжа для котів, вологий корм…	Напої
10	Пакет	Сумка, кульок…	Товари
11	Доставка	Привезти, доставити…	Послуга
12	Знижка	Акція, розпродаж, вигода…	Обслуговування
13	Кошик	Корзина, кошелка, кошик для покупок	Обслуговування

Використання гнучких голосових запитів є надзвичайно важливим для ефективного функціонування АСК, особливо в умовах супермаркету. Люди не є роботами і природно спілкуються живою, емоційною та індивідуальною

мовою, яка суттєво відрізняється від чітких і формалізованих команд. Кожна людина має власний стиль мовлення, використовує різні слова, синоніми, діалекти та інтонації, що створює додаткові виклики для автоматизованих систем.

Наприклад, один користувач може сказати "покажи акції на батон", інший – "чи є знижки на хліб?", а третій – "що вигідного з випічки?". Усі ці запити мають однаковий зміст, але сформульовані по-різному. Якщо система працює лише з чіткими командами, вона не зможе коректно обробити такі запити. Саме тому важливо, щоб система могла розпізнавати різні слова, розуміти синоніми, контекст і адаптуватися до індивідуальних особливостей мовлення.

Гнучка обробка голосових запитів дозволяє зробити взаємодію з користувачем більш природною та інтуїтивною. Це підвищує зручність користування, зменшує кількість помилок у спілкуванні та сприяє ефективнішому виконанню завдань. Завдяки використанню бази синонімів система стає здатною розуміти не лише прямі запити, а й варіації фраз, що робить її доступною для ширшого кола користувачів незалежно від їх мовних звичок.

3.7 Архітектура нейронної мережі

Обробка голосу – це складна задача, яка потребує ефективного аналізу часових і частотних ознак звукових сигналів. Для цієї мети використовуються різні типи нейронних мереж, кожен з яких має свої сильні сторони.

У задачах обробки голосу використовуються такі типи нейронних мереж як CNN (Convolutional Neural Networks), RNN (Recurrent Neural Networks), LSTM (Long Short-Term Memory), GRU (Gated Recurrent Units), WaveNet, ResNet та інші. В таблиці 3.4 представлено порівняння цих нейронних мереж для обробки голосу.

Таблиця 3.4 – Порівняння нейронних мереж для обробки голосу

Модель	Сильні сторони	Слабкі сторони
ResNet	Вирішення проблеми зникання градієнтів; ефективне навчання глибоких моделей; висока узагальнювальність.	Вимагає оптимізації для великих часових послідовностей
CNN	Добре працюють із двовимірними спектрограмами; швидке навчання.	Обмежена здатність враховувати глобальні часові залежності.
RNN	Ідеально підходять для аналізу послідовностей; простота в реалізації.	Проблеми зі зникання градієнтів у тривалих послідовностях; складно навчати дуже довгі залежності.
LSTM і GRU	Ефективно зберігають довготривалі часові залежності; знижують проблему зникання градієнтів.	Повільне навчання; більш складна архітектура порівняно з CNN.
WaveNet	Відмінна якість синтезу мовлення; генеративний підхід для аудіо.	Складність масштабування; високі вимоги до обчислювальних потужностей.

Серед них особливу увагу привертає ResNet (Residual Networks), яка поєднує високу ефективність і гнучкість у навчанні глибоких моделей.

Резидуальні нейронні мережі (ResNet) стали важливим інструментом у сфері обробки голосових запитів, де ефективність і точність моделі мають вирішальне значення. Використання резидуальних блоків дозволяє мережам

успішно подолати проблему градієнтного затухання, яка особливо актуальна для глибоких архітектур, необхідних для моделювання складних акустичних і мовних сигналів.

У задачах обробки голосових запитів, таких як розпізнавання мовлення, синтез голосу та аналіз емоцій, резидуальні мережі використовуються для виявлення тонких структурних і тимчасових залежностей в аудіосигналах. Ключова перевага ResNet полягає у здатності моделі зосереджуватися на залишкових змінах, що дозволяє зберігати основну інформацію про сигнал навіть після багаторазової трансформації даних через шари мережі. Резидуальні з'єднання, які передають інформацію між шарами, мінімізують втрату критично важливих особливостей, таких як тональні і ритмічні характеристики голосу.

ResNet також показує високу ефективність у поєднанні з техніками часово-частотного представлення аудіосигналів, такими як перетворення сигналів у спектрограми. Завдяки цьому архітектура адаптується до виявлення ключових патернів, які характеризують мовлення, акценти або навіть шум у фоні. У задачах розпізнавання голосових команд ResNet дозволяє поліпшити точність класифікації, оскільки модель ефективно ідентифікує навіть незначні відхилення у звучанні різних слів.

Якщо позначити вхідні дані як x, а бажану вихідну функцію як H(x), то мережа вчиться резидуальній функції F(x) = H(x) - x. Тоді вихід резидуального блоку визначається як H(x) = F(x) + x.

На рисунку 3.6 зображено архітектуру резидуального блоку у глибокій резидуальній мережі.

Архітектура ResNet складається з послідовності таких резидуальних блоків, що дозволяє створювати дуже глибокі мережі з сотнями шарів без втрати ефективності тренування. Це досягається завдяки тому, що кожен блок вчиться моделювати лише різницю між вхідними та вихідними даними, а не повну трансформацію, що спрощує процес оптимізації.

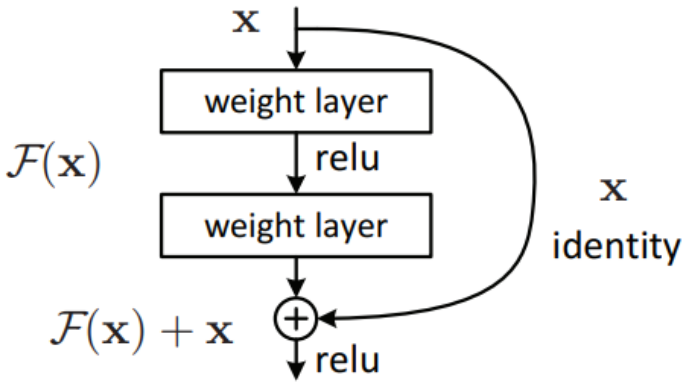

Рисунок 3.6 – Резидуальний блок у глибокій резидуальній мережі

На рисунку 3.7 представлено архітектуру 18-шарової ResNet, яка використовується для обробки гнучких голосових запитів.

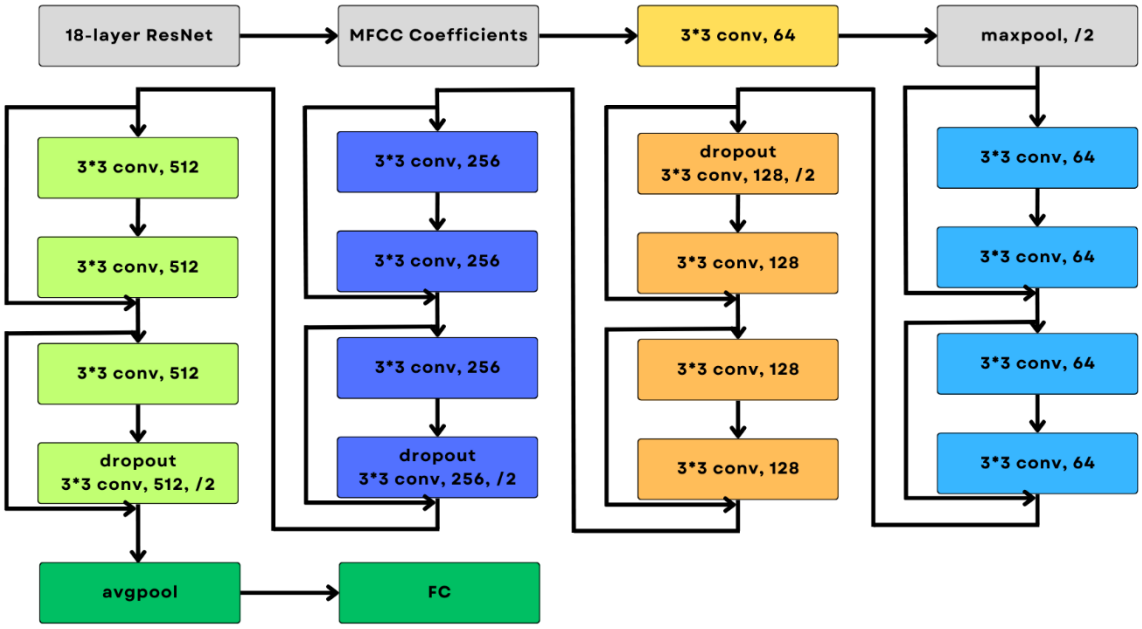

Рисунок 3.7 – Архітектура ResNet

Архітектура 18-шарового ResNet базується на залишкових з'єднаннях, що запобігають затуханню градієнтів. Вхідні дані (наприклад, MFCC коефіцієнти) проходять через 3x3 згортку з 64 фільтрами та пулінг для зменшення розмірності.

Далі використовуються блоки згорток із фільтрами (64, 128, 256, 512), підвибіркою (stride = 2) і dropout для регуляризації. Завершується архітектура глобальним усередненим пулінгом і повнозв’язним шаром для класифікації.

Завдяки резидуальним з’єднанням і модульній структурі ResNet-18 ефективно витягує ознаки навіть із складних даних.

3.8 Обробка голосового запиту

Процес обробки голосового запиту складається з кількох послідовних етапів, кожен з яких виконує певну роль у перетворенні звукового сигналу на текст.

На першому етапі відбувається завантаження аудіосигналу у форматі .wav. Цей файл містить записані звукові коливання у цифровому вигляді. Сигнал може бути монофонічним або стерео, і цей етап є підготовчим для подальшої обробки. Відображення першого етапу представлено на рисунку 3.8.

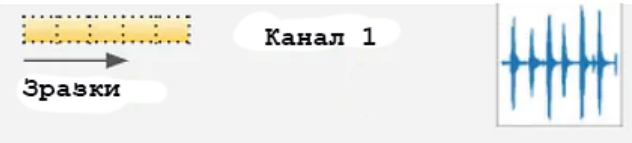

Рисунок 3.8 – Завантаження аудіо з файлу .wav

Після завантаження аудіо виконується передискретизація, тобто перетворення частоти дискретизації на стандартне значення, яке використовуватиметься в системі. Якщо сигнал був монофонічним, його перетворюють у стерео, додаючи другий канал. На рисунку 3.9 показано цей етап.

Рисунок 3.9 – Передискретизація аудіо сигналу

Далі змінюється розмір сигналу до фіксованої довжини. Це дозволяє привести всі вхідні дані до однакового формату для зручності обробки, навіть якщо оригінальний запис був занадто коротким або довгим. Відображення цього показано на рисунку 3.10.

Рисунок 3.10 – Зміна розміру до фіксованої довжини

Четвертий етап – зсув у часі, що є частиною аудіоаргументації. Це означає, що сигнал може бути зміщений у часі для створення варіацій даних, що підвищує стійкість моделі до змін у записах. На рисунку 3.11 представлено цей етап.

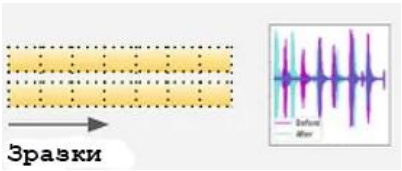

Рисунок 3.11 – Зсув у часі аудіоаргументації

На наступному етапі аудіосигнал конвертується в мел-спектрограму. Це перетворення дозволяє візуалізувати частотний спектр сигналу у вигляді двовимірного зображення, що показує розподіл енергії в часі та частоті. Це відображення можна побачити на рисунку 3.12.

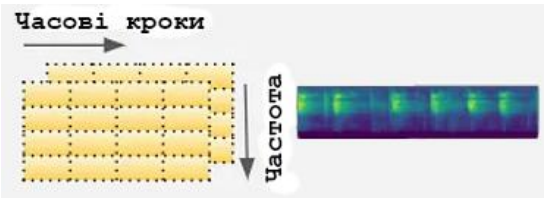

Рисунок 3.12 – Конвертація в спектограму Мел

Після створення мел-спектрограми до неї застосовується аргументація. Це можуть бути зміни кольору, шуму або інші варіації, які допомагають моделі бути стійкою до різних умов запису. Відображення цього показано на рисунку 3.13.

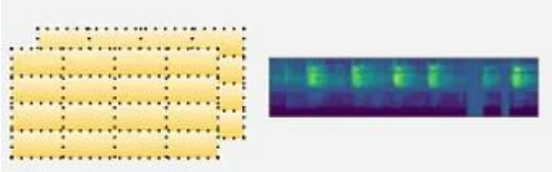

Рисунок 3.13 – Аргументація спектограми

Заключний етап – це екстракція мел-кепстральних коефіцієнтів (MFCC), які є ключовими ознаками, що використовуються для розпізнавання мовлення. Отримані коефіцієнти стають вхідними даними для подальшого аналізу. Цей етап зображено на рисунку 3.14.

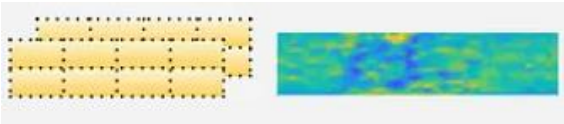

Рисунок 3.14 – Екстракція мел-кепстральних коефіцієнтів

Після цього підготовлені дані зіставляються з мітками (y), які представляють текстову транскрипцію сказаного. Цей текст може бути закодований у вигляді ідентифікаторів символів для полегшення обробки. Результат зображено на рисунку 3.15.

Рисунок 3.15 – Текстова транскрипція

Повний процес обробки перетворення голосових хвиль на зображення спектрограми зображено на рисунку 3.16.

Таким чином, цей процес дозволяє перетворити аналоговий звуковий сигнал у цифрове представлення, яке можна аналізувати та обробляти за допомогою алгоритмів машинного навчання. Завдяки послідовній обробці, включно з передискретизацією, нормалізацією, зсувом у часі та конвертацією в мел-спектрограму, створюється наочне та інформативне зображення звукового сигналу. Це забезпечує точність і стійкість у подальшому аналізі мовлення, відкриваючи можливості для розробки ефективних систем розпізнавання та обробки голосу.

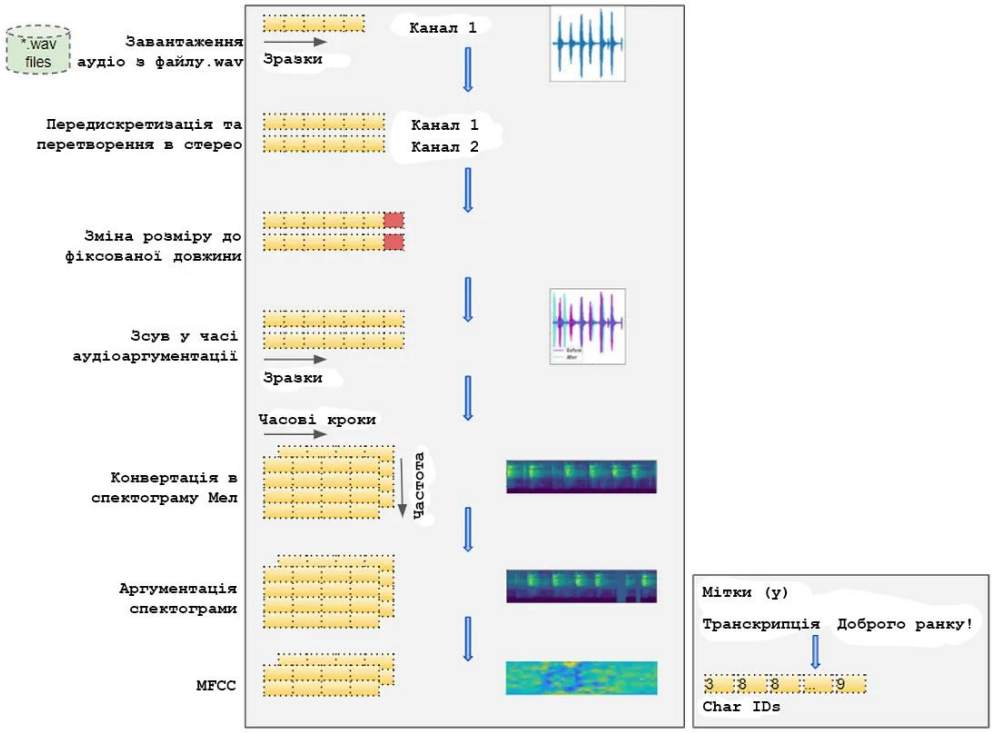

Рисунок 3.16 – Перетворення голосових хвиль на зображення спектограми

3.9 Результати і продуктивність системи

Для оцінки якості розпізнавання мовлення було обрано 4 показники такі, як точність (accuracy), влучність (precision), повнота (recall), F-міра (F-measure).

Accuracy (A) – це частка правильно класифікованих випадків (як позитивних, так і негативних) серед загальної кількості випадків. Вона показує, наскільки добре система розпізнає як мовлення, так і відсутність мовлення. Обчислюється за формулою:

$$A = \frac{TP + TN}{TP + TN + FP + FN},$$

де TP – кількість справжніх позитивних;

TN – кількість справжніх негативних;

FP – кількість хибно позитивних;

FN – кількість хибно негативних.

Presicion (P) – це частка правильно класифікованих позитивних випадків серед усіх випадків, які система визначила як позитивні. Обчислюється за формулою:

$$P = \frac{TP}{TP + FP} * 100.$$

Повнота (R) – це частка правильно розпізнаних позитивних випадків (наприклад, правильних розпізнавань мовлення) серед загальної кількості реальних позитивних випадків. Вона відображає здатність системи виявляти всі релевантні випадки. Обчислюється за формулою:

$$P = \frac{TP}{TP + FN} * 100.$$

F-міра (F_1) – це гармонійне середнє між точністю та повнотою, яке використовується для балансування цих двох показників. F-міра обчислюється за формулою:

$$F_1 = \frac{P * R}{P + R.}$$

Таблиця 3.5 – Опис методики проведення експериментів без використання направлених мікрофонів

№ експе-рименту	Рівень зашумле-ності	Метод виділення ознак	Метод фільтрації шуму	Мета експерименту
Експе-рименти №1-12	Низький (SNR ≥ 30 дБ)	Mel-Spectrogram	Без фільтрів	Оцінка базової точності розпізнавання (точність, влучність, повнота, F-міра) голосової команди за умов низької зашумленості та використання MFCC
			Low-pass фільтр	
			High-pass фільтр	
			Комбінований (low-pass + high-pass)	
		STFT	Без фільтрів	Оцінка базової точності розпі-знавання голо-сової команди за умов низької зашумленості та використання STFT
			Low-pass фільтр	
			High-pass фільтр	
			Комбінований (low-pass + high-pass)	

Продовження таблиці 3.5

№ експе-рименту	Рівень зашумле-ності	Метод виділення ознак	Метод фільтрації шуму	Мета експерименту
Експе-рименти №1-12	Низький (SNR ≥ 30 дБ)	CQT	Без фільтрів	Оцінка базової точності розпі-знавання голо-сової команди за умов низької зашумленості та використання CQT
			Low-pass фільтр	
			High-pass фільтр	
			Комбінований (low-pass + high-pass)	
Експе-рименти №13-24	Високий (SNR ≤ 15 дБ)	Mel-Spectrogram	Без фільтрів	Оцінка базової точності розпі-знавання голо-сової команди за умов високої зашумленості та використання MFCC
			Low-pass фільтр	
			High-pass фільтр	
			Комбінований (low-pass + high-pass)	
		STFT	Без фільтрів	Оцінка базової точності розпі-знавання голо-сової команди за умов високої зашумленості та використання STFT
			Low-pass фільтр	
			High-pass фільтр	
			Комбінований (low-pass + high-pass)	

Продовження таблиці 3.5

№ експе-рименту	Рівень зашумле-ності	Метод виділення ознак	Метод фільтрації шуму	Мета експерименту
Експе-рименти №13-24	Високий (SNR ≤ 15 дБ)	CQT	Без фільтрів	Оцінка базової точності розпізнавання голосової команди за умов високої зашумленості та використання CQT
			Low-pass фільтр	
			High-pass фільтр	
			Комбінований (low-pass + high-pass)	

Методики проведення експериментів, які оцінюють вплив обраних методів виділення ключових ознак аудіоряду, методів пригнічення шуму, використання направлених мікрофонів і рівня зашумленості аудіодоріжки на точність розпізнавання голосової команди показано в таблиці нижче. Опис методики проведення експериментів без використання направлених мікрофонів наведено в таблиці 3.5.

Аналіз отриманих результатів допоможе зрозуміти який підхід бажано використовувати для різного ступеня зашумленості аудіоряду. Визначення пайплайну для низькозашумлених та високозашумлених аудіодоріжок будуть бейслайнами для оцінки впливу використання направлених мікрофонів на точність розпізнавання голосової команди.

На рисунку 3.17 представлено результати проведення експерименту з низьким рівнем зашумленності без використання направлених мікрофонів у вигляді діаграми.

Таблиця 3.6 – Результати проведення експерименту з низьким рівнем зашумленності без використання направлених мікрофонів

Номер експери-менту	Метод вилучення ознак	Метод фільтрації шуму	Accuracy (A), %	Precision (P), %	Повнота (R), %	F-міра (F1), %
1	Mel-Spectrogram	Без-фільтрів	80,21	81,10	79,90	80,49
2	Mel-Spectrogram	Low-pass	82,10	82,90	81,70	82,29
3	Mel-Spectrogram	High-pass	81,80	82,60	81,40	81,99
4	Mel-Spectrogram	Комбіно-ваний	85.10	85,90	84,80	85,34
5	STFT	Без-фільтрів	78,10	78,90	77,50	78,19
6	STFT	Low-pass	79,50	80,30	79,00	79,60
7	STFT	High-pass	79,20	80,00	78,70	79,34
8	STFT	Комбіно-ваний	81,10	81,90	80,60	81,19
9	CQT	Без-фільтрів	76,50	77,20	75,80	76,49
10	CQT	Low-pass	77,80	78,40	77,00	77,73
11	CQT	High-pass	78,10	78,10	76,80	77,44
12	CQT	Комбіно-ваний	79,10	79,80	78,50	79,14

Результати для низького рівня зашумленості показують, що Mel-Spectrogram демонструє найвищу ефективність серед усіх методів виділення ознак, особливо у поєднанні з комбінованими фільтрами шуму, досягаючи точності 85.10%. Інші методи, такі як STFT і CQT, мають нижчі показники,

але також виграють від використання комбінованих фільтрів, що свідчить про їхню ефективність для таких умов. Комбіновані фільтри забезпечують помітне підвищення продуктивності у порівнянні з окремими Low-pass та High-pass фільтрами.

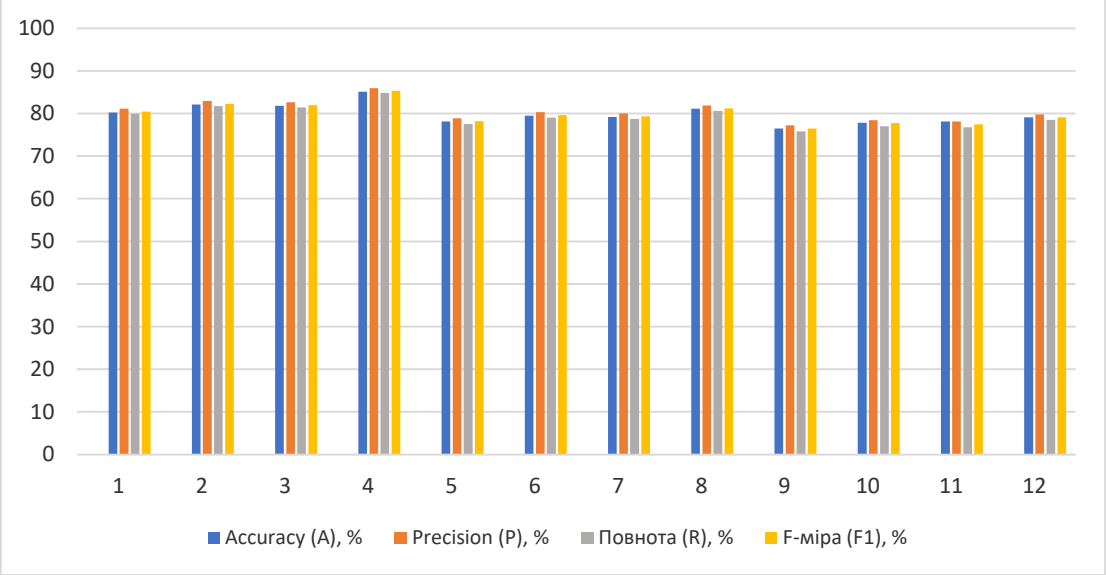

Рисунок 3.17 – Результати проведення експерименту з низьким рівнем зашумленності без використання направлених мікрофонів

В таблиці 3.7 представлено результати експериментів з високим рівнем зашумленості.

Таблиця 3.7 – Результати проведення експерименту з високим рівнем зашумленності без використання направлених мікрофонів

№ експе-рименту	Метод вилучення ознак	Метод фільтрації шуму	Accuracy (A), %	Precision (P), %	Повнота (R), %	F-міра (F1), %
13	Mel-Spectrogram	Без-фільтрів	72,21	73,10	71,90	72,49
14	Mel-Spectrogram	Low-pass	73,50	74,40	73,20	73,70

Продовження таблиці 3.7

Номер експери-менту	Метод вилучення ознак	Метод фільтрації шуму	Accuracy (A), %	Precision (P), %	Повнота (R), %	F-міра (F1), %
15	Mel-Spectrogram	High-pass	73,20	74,10	72,80	73,45
16	Mel-Spectrogram	Комбіно-ваний	75,30	76,00	75,00	75,49
17	STFT	Без-фільтрів	70,10	71,00	69,80	70,39
18	STFT	Low-pass	71,30	72,20	71,00	71,50
19	STFT	High-pass	71,00	71,80	70,60	71,20
20	STFT	Комбіно-ваний	72,50	73,30	72,10	72,63
21	CQT	Без-фільтрів	68,50	69,40	68,10	68,66
22	CQT	Low-pass	69,70	70,60	69,30	69,86
23	CQT	High-pass	69,40	70,20	69,00	69,60
24	CQT	Комбіно-ваний	70,90	71,80	70,50	71,06

На рисунку 3.18 представлено результати проведення експерименту з високим рівнем зашумленності без використання направлених мікрофонів у вигляді діаграми.

За умов високого рівня зашумленості Mel-Spectrogram зберігає своє лідерство, досягаючи 75.30% точності з комбінованими фільтрами. Усі методи демонструють зниження продуктивності в умовах значного шуму, проте комбіновані фільтри допомагають зменшити цей вплив, покращуючи результати на 2-3% порівняно з використанням лише Low-pass або High-pass фільтрів. Методи STFT і CQT залишаються менш ефективними, але отримують приріст від застосування комбінованих фільтрів.

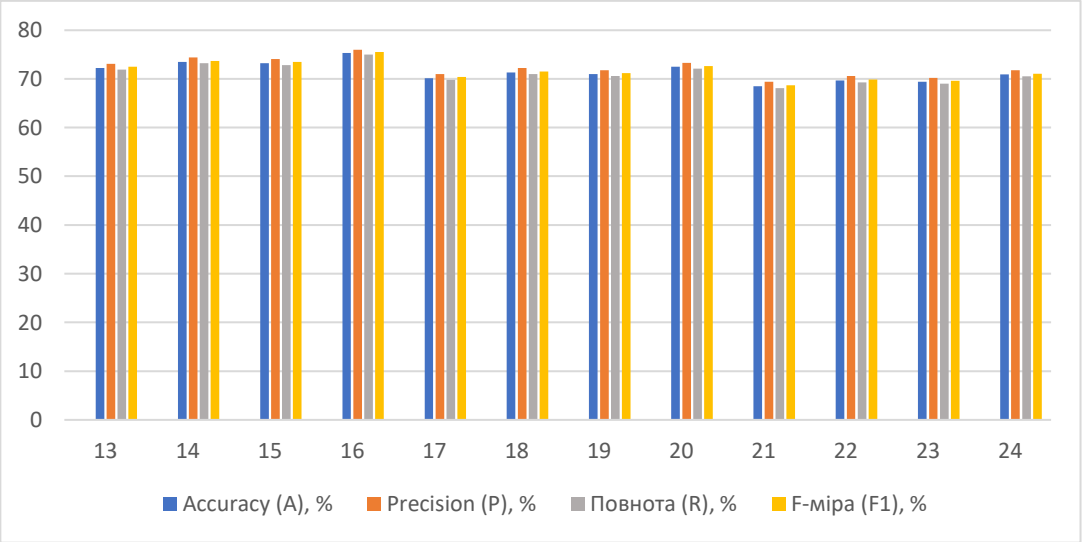

Рисунок 3.18 – Результати проведення експерименту з високим рівнем зашумленності без використання направлених мікрофонів

Наступні експерименти будуть побудовані наступним чином – буде проведена оцінка точності, влучності, повноти та F-міри розпізнавання голосової команди за умови використання направлених в сторону джерела звуку мікрофонів. Опис методики проведення експериментів представлено в таблиці 3.8.

Таблиця 3.8 – Опис методики проведення експериментів для оцінки точності, влучності, повноти, F-міри розпізнавання голосової команди за умови використання направлених в сторону джерела звуку мікрофонів

№ експерименту	Рівень зашумленості	Обраний пайплайн обробки аудіоряду
25	Низький (SNR ≥ 30 дБ)	Заповнюється по результатам експериментів 1-12
26	Високий (SNR ≤ 15 дБ)	Заповнюється по результатам експериментів 13-24

За результатами таблиці 3.7 варто обрати для експериментів 25 і 26 метод вилучення ознак Mel-Spectrogram і метод фільтрації шуму – комбінований фільтр, так як саме ці методи разом показали найвищі значення точності, влучності, повноти та F-міри.

В таблиці 3.9 представлено результати з низьким і високим рівнем зашумленості з використанням направлених до джерела звуку мікрофонів.

Таблиця 3.9 – Результати експериментів низького і високого рівня зашумленості з використанням направлених мікрофонів

Номер експери-менту	Метод вилучення ознак	Обраний пайплайн обробки аудіоряду	Accuracy (A), %	Precision (P), %	Повнота (R), %	F-міра (F1), %
25	Mel-Spectrogram	Mel-Spectrogram + Комбінований	89,10	89,90	88,70	89,25
26	Mel-Spectrogram	Mel-Spectrogram + Комбінований	76,50	77,20	76,00	76,56

Порівняння результатів використання направлених і ненаправлених мікрофонів свідчить про значний вплив цього параметра на ефективність системи розпізнавання мовлення.

У випадку ненаправлених мікрофонів, точність (Accuracy) для методу з комбінованими фільтрами досягала 85,10%, що є хорошим результатом. Однак направлені мікрофони дозволили підвищити цей показник до 89.10%, що свідчить про їхню здатність зменшувати вплив фонових шумів і покращувати якість вхідного сигналу.

Інші метрики також демонструють позитивну динаміку: влучність (Presicion) зросла з 85,90% до 89,90%, повнота (Recall) – з 84,80% до 88,50%,

а F-міра – з 85,34% до 89,10%. Такі покращення підкреслюють ефективність направлених мікрофонів у фокусуванні на джерелі мовлення, що є особливо важливим у середовищах із високим рівнем зашумленості.

Результати аналізу вказують, що направлені мікрофони дозволяють не лише підвищити точність моделі, але й забезпечити стабільніші показники продуктивності за всіма ключовими метриками. Це підтверджує їхню доцільність у задачах розпізнавання мовлення, особливо за умов, коли якість вхідного аудіосигналу є критичним фактором.

Останнім експериментом буде порівняння запропонованого підходу розпізнавання гнучких голосових запитів на основі нейромережевої моделі ResNet із визначеними пайплайнами обробки аудіорядів, включаючи можливість використання направлених мікрофонів, за умов низького та високого рівня зашумленості із існуючими підходами.

На точність розпізнавання мовлення впливає низка параметрів, пов’язаних із характеристиками аудіосигналу та методами його обробки. Зокрема, ключовими факторами є:

- коефіцієнти MFCC (Mel Frequency Cepstral Coefficients) – базовий набір характеристик, що відображають частотні особливості мовлення. MFCC є основою для побудови систем розпізнавання;
- середнє значення (Mean) – відображає середній рівень інтенсивності сигналу та допомагає визначити загальний рівень амплітуди мовлення;
- стандартне відхилення (Standard Deviation) – характеризує ступінь варіації в інтенсивності сигналу, що є важливим для розпізнавання емоційних чи інтонаційних відтінків;
- амплітуда (Amplitude) – додаткова характеристика, що впливає на ідентифікацію окремих звуків або тонів;
- перетин нуля (Zero-Crossing) – показник, що описує частоту зміни знаку сигналу, важливий для аналізу тембру та голосових характеристик.

Щоб визначити, як кожен із перелічених факторів впливає на точність розпізнавання мовлення, необхідно провести серію експериментів:

- використати лише коефіцієнти MFCC як базову характеристику;
- додати середнє значення (Mean) до MFCC;
- додати стандартне відхилення (Standard Deviation) до MFCC;
- використати амплітуду (Amplitude) разом із MFCC;
- додати показник перетину нуля (Zero-Crossing) до MFCC.

Таблиця 3.10 – Результати оцінювання ефективності розпізнавання мовлення

Вилучення ознак	Accuracy (A), %	Presicion (P), %	Повнота (R), %	F-міра (F1), %
MFCC	80,21	81,10	79,90	80,49
MFCC + Mean	82,45	83,30	81,60	82,44
MFCC + Standard Deviation	84,75	85,50	83,90	84,69
MFCC + Amplitude	86,20	87,00	85,40	86,19
MFCC + Zero-Crossing	87,10	87,90	86,50	87,19
MFCC + Комбіновані фільтри + Направлені мікрофони	89,10	89,90	88,50	89,20

На рисунку 3.19 зображено графік створений по результатам таблиці для наочного бачення результатів.

Результати аналізу свідчать, що використання базових коефіцієнтів MFCC забезпечує прийнятну точність, яка становить 80.21%. Це слугує основою для оцінки впливу додаткових ознак. Додавання середнього значення (Mean) до базових коефіцієнтів покращує всі метрики, зокрема влучність (P), що свідчить про позитивний вплив цієї ознаки на продуктивність системи. Включення стандартного відхилення (Standard Deviation) забезпечує стабільне зростання повноти та F-міри, підвищуючи здатність моделі виявляти всі релевантні результати.

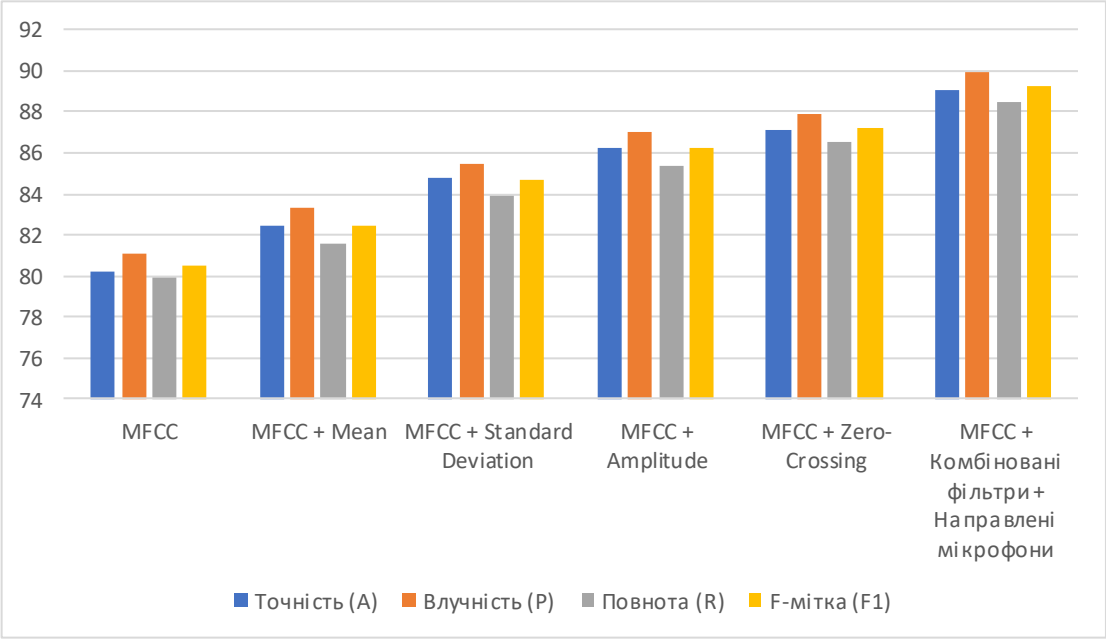

Рисунок 3.19 – Міри точності розпізнавання

Амплітуда (Amplitude) надає додатковий контекст, що дозволяє ще більше підвищити загальну точність і забезпечити кращу адаптацію моделі до змін у мовному сигналі. Частота перетину нуля (Zero-Crossing) додає корисну інформацію про структуру сигналу, завдяки чому підвищуються всі ключові метрики.

Застосування комбінованих фільтрів у поєднанні з направленим мікрофоном (MFCC + Комбіновані фільтри + Направлений мікрофон) демонструє найвищу продуктивність системи. Точність досягає 89.10%, а решта метрик, такі як влучність, повнота та F-міра, також знаходяться на максимальному рівні. Це підтверджує, що поєднання комбінованих фільтрів і направленого мікрофона дозволяє ефективно врахувати різні аспекти мовного сигналу, знижуючи вплив шуму та підвищуючи результати розпізнавання.

ВИСНОВКИ

У процесі виконання кваліфікаційної роботи було досліджено методи обробки гнучких голосових запитів в автоматизованих системах керування завданнями. Отримані результати підтверджують актуальність розробки таких систем, враховуючи стрімке зростання попиту на голосові технології в різних сферах діяльності.

Наукова новизна роботи полягає у створенні архітектури системи, здатної адаптуватися до гнучких голосових запитів, що значно підвищує її універсальність і точність. У роботі вперше запропоновано використовувати базу даних синонімів для врахування варіативності мовлення, що дозволяє підвищити ефективність роботи системи в умовах реального використання.

Практична значущість результатів полягає в тому, що розроблена система може бути застосована для автоматизації бізнес-процесів, інтеграції з платформами електронної комерції та впровадження у повсякденне використання, як-от розпізнавання голосових запитів у супермаркетах. Особливо важливим є те, що система є незапатентованою і може бути використана як основа для подальших відкритих розробок.

У подальших дослідженнях доцільно зосередитися на розширенні бази даних синонімів, що дозволить системі враховувати ширший спектр галузевої термінології. Також перспективним напрямком є інтеграція сучасних технологій машинного навчання для забезпечення адаптації системи до нових мовних даних, що дозволить покращити її продуктивність. Окрему увагу варто приділити розробці мобільної версії системи, яка розширить можливості її використання та забезпечить додаткову зручність для користувачів.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Olesia Barkovska, Ihor Velykodnyi, Oleksii Liashenko, Ihor Ivanisenko, Yaroslav Davydov "Study of the architectural features of the ResNet neural network model for solving the task of speaker recognition" 2024 IEEE 5th KhPI Week on Advanced Technology. IEEE Conference, October 7 – 11, 2024, Kharkiv, Ukraine.

2. Давидов Я.А., Барковська О.Ю Методи обробки гнучких голосових запитів в автоматизованих системах керування завданнями // Проблеми інформатизації : XII міжнародна науково-технічна конференція. – 21-22 листопада 2024. –с.73. doi: https://doi.org/10.32620/PI.24.t2.

3. Rai, S., Li, T., & Lyu, B. Keyword spotting - "Detecting commands in speech using deep learning" / arXiv Preprint. – 2024. [Електронний ресурс] – Режим доступу: https://doi.org/10.48550/arXiv.2312.05640 – 23.10.2024.

4. Hamza Kheddar, Rached Hamila, and Hiba Al-Shamma'a. "Automatic Speech Recognition using Advanced Deep Learning Approaches: A Surveyp". – 2024. [Електронний ресурс] – Режим доступу: https://doi.org/ 10.48550/arXiv.2403.01255 – 23.10.2024.

5. Li, Y., Zhang, X., & Shao, Y. "A Survey of the Usages of Deep Learning for Natural Language Processing". arXiv. – 2020. [Електронний ресурс] – Режим доступу: https://doi.org/10.48550/arXiv.1807.10854 – 23.10.2024.

6. Chenglei Si et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". arXiv. – 2023. [Електронний ресурс] – Режим доступу: https://doi.org/10.48550/arXiv.2005.11401 – 23.10.2024.

7. Fu-Lian Yin, Xing-Yi Pan, Xiao-Wei Liu and Hui-Xin Liu, "Deep neural network language model research and application overview," 2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu – 2015. – pp. 55-60.

[Електронний ресурс] – Режим доступу: https://ieeexplore.ieee.org/document/7493906 – 23.10.2024.

8. F. Ertam and G. Aydın, "Data classification with deep learning using Tensorflow," 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey. – 2017. – pp. 755-758. [Електронний ресурс] – Режим доступу: https://ieeexplore.ieee.org/document/8093521 – 23.10.2024.

9. M. Lambeta et al., "PyTouch: A Machine Learning Library for Touch Processing," 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China – 2021. – pp. 13208-13214. [Електронний ресурс] – Режим доступу: https://ieeexplore.ieee.org/document/9561084 – 23.10.2024.

10. S. Li, J. You and X. Zhang, "Overview and Analysis of Speech Recognition," 2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), Dalian, China, 2022. – pp. 391-395. [Електронний ресурс] – Режим доступу: https://ieeexplore.ieee.org/document/9919050 – 23.10.2024.

11. P. N. Singh and S. Behera, "The Transformers' Ability to Implement for Solving Intricacies of Language Processing," 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), Ravet, India. – 2022. – pp. 1-7. [Електронний ресурс] – Режим доступу: https://ieeexplore.ieee.org/document/9909423 – 23.10.2024.

12. Shalbbya Ali, Salfar Tanweer, Syed Sibtain Khalid and Naseem Rao, "Mel Frequency Cepstral Coefficient: A Review," 2021 Conference: Proceedings of the 2nd International Conference on ICT for Digital, Smart, and Sustainable Development, New Delhi, India. – 2020. – pp. 92-101. [Електронний ресурс] – Режим доступу: https://eudl.eu/doi/10.4108/eai.27-2-2020.2303173.

13. Intelligent Speech Interaction [Електронний ресурс] – Режим доступу: https://www.alibabacloud.com/en/product/intelligent-speech-interaction?_p_lc=1. – 16.01.2024р. – Заголовок з екрану.

14. Amazon Transcribe [Електронний ресурс] – Режим доступу: https://aws.amazon.com/transcribe/?nc1=h_ls. – 16.01.202

15. Dragon Speech Recognition Solutions [Електронний ресурс] – Режим доступу:https://www.nuance.com/dragon.html?srsltid=AfmBOopY8I1KVZ4dUIka0_t9ipFRVuDzinAYNA7OxXhsSHb3uEGiyPyv. – 16.01.2024р. – Заголовок з екрану.

16. Deepgram Voice AI [Електронний ресурс] – Режим доступу: https://deepgram.com/. – 16.01.2024р. – Заголовок з екрану.

17. Cloud Speech-to-Text. [Електронний ресурс] – Режим доступу: https://cloud.google.com/speech-to-text. – 16.01.2024р. – Заголовок з екрану.

18. Erica – Voice Financial Assistant. [Електронний ресурс] – Режим доступу:https://promotions.bankofamerica.com/digitalbanking/mobilebanking/erica. – 16.01.2024р. – Заголовок з екрану.

19. Atom's use of biometrics. [Електронний ресурс] – Режим доступу: https://www.atombank.co.uk/blog/2015/11/why-is-atom-using-biometrics/ – 16.01.2024р. – Заголовок з екрану.

20. Voice payment adoption rate in the United States in 2017 with forecasts from 2018 to 2022. [Електронний ресурс] – Режим доступу: https://www.statista.com/statistics/917933/voice-payments-adoption-rate-usa/. – 16.01.2024р. – Заголовок з екрану.

21. Голосова біометрія (аутентифікація). [Електронний ресурс] – Режим доступу: https://privatbank.ua/voice-biometrics. – 16.01.2024р. – Заголовок з екрану.

22. Common Voice 20 is Now Available. [Електронний ресурс] – Режим доступу: https://foundation.mozilla.org/en/blog/common-voice-20-is-now-available/ – 16.01.2024р. – Заголовок з екрану.