Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерної інженерії та інформаційних систем

# КВАЛІФІКАЦІЙНА РОБОТА

Кіберфізична система розпізнавання голосу людини на базі алгоритмів машинного навчання
Назва теми

Рівень вищої освіти *другий (магістерський)*

Галузь знань *12 «Інформаційні технології»*
Шифр, назва

Спеціальність *123 «Комп'ютерна інженерія»*
Шифр, назва

Освітня програма *«Комп'ютерна інженерія та програмування»*
Назва

Шифр КвРКІ 240240.11.02.37 ПЗ

Виконав здобувач II курсу, група КІ2м-24-2 _____ Дмитро КРУТИЙ
Підпис Ініціали, прізвище

Керівник канд.-техн. наук, доцент _____ Володимир ГРИГА
Науковий ступінь, учене звання Підпис Ініціали, прізвище

Нормоконтролер д. техн. наук, професор _____ Сергій ЛИСЕНКО
Науковий ступінь, учене звання Підпис Ініціали, прізвище

До захисту допускаю:
завідувач кафедри КІІС _____ Ольга ПАВЛОВА
«01» травня 2026 р. Підпис Ініціали, прізвище
дата

Хмельницький 2026

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет Інформаційних технологій

Кафедра Комп'ютерної інженерії та інформаційних систем

Рівень вищої освіти другий (магістерський)

Галузь знань 12 Інформаційні технології

Спеціальність 123 Комп'ютерна інженерія

Освітня програма «Комп'ютерна інженерія та програмування»

ЗАТВЕРДЖУЮ
Завідувачка кафедри КІІС
_____ Ольга ПАВЛОВА

" 12 " 01 2026 р.

**ЗАВДАННЯ НА КВАЛІФІКАЦІЙНУ РОБОТУ**

Крутому Дмитру Вікторовичу
Прізвище, ім'я, по батькові студента

1. Тема проекту (роботи) Кіберфізична система розпізнавання голосу людини на базі алгоритмів машинного навчання
Керівник проекту (роботи) Грига Володимир Михайлович, к.т.н., доц.
Прізвище, ім'я, по батькові, науковий ступінь, вчене звання

Затверджена наказом ректора університету від 12.01.2026 р. № 6
2. Термін подання здобувачем роботи на кафедру 01.05.2026 р.
3. Вихідні дані до роботи Завдання на кваліфікаційну роботу
4. Зміст пояснювальної записки (перелік питань, які потрібно розробити)
Аналіз предметної галузі розпізнавання голосу
Моделювання кіберфізичної системи розпізнавання голосу людини
Методи та алгоритми машинного навчання для розпізнавання голосу
Проектування та дослідження кіберфізичної системи розпізнавання голосу
5. Перелік графічного матеріалу (із зазначенням обов'язкових креслень)

6. Консультанти розділів кваліфікаційної роботи

| Розділ | Прізвище, ініціали та посада консультанта | Підпис, дата | |
|---|---|---|---|
| | | завдання видав | завдання прийняв |
| | | | |
| | | | |

7. Дата видачі завдання «_12_»_____01_____2026 р.

## КАЛЕНДАРНИЙ ПЛАН

| №з/п | Назва етапів (розділів) кваліфікаційної роботи магістра | Термін виконання етапів проекту (роботи) | Примітка |
|---|---|---|---|
| 1 | Вибір напряму дослідження та узгодження тематики КвРМ з керівником | 12.01.2026 | виконано |
| 2 | Ознайомлення з предметною областю; формулювання мети та задач дослідження; визначення об'єкта та предмета дослідження | 12.01.2026 | виконано |
| 3 | Робота над розділом 1 – аналіз відомих моделей, методів за темою; постановка задачі | 20.01.2026 | виконано |
| 4 | Робота над розділом 2 – розробка моделей для вирішення поставленої задачі | 01.02.2026 | виконано |
| 5 | Робота над науковою статтею | 01.03.2026 | виконано |
| 6 | Робота над розділом 3 – розробка методів для вирішення поставленої задачі | 15.03.2026 | виконано |
| 7 | Робота над розділом 4 – проектування та розробка ПЗ для вирішення поставленої задачі, експериментальна частина | 01.04.206 | виконано |
| 8 | Оформлення пояснювальної записки згідно вимог | 18.04.2026 | виконано |
| 9 | Попередній захист ДРМ | 29.04.2026 | виконано |
| 10 | Захист ДРМ на засіданні ЕК | До 20.05.2026 | |

Здобувач_____ Дмитро КРУТИЙ
Підпис Ім'я, ПРІЗВИЩЕ

Керівник кваліфікаційної роботи_____ Володимир ГРИГА
Підпис Ім'я, ПРІЗВИЩЕ

# РЕФЕРАТ

Тема кваліфікаційної роботи магістра: Кіберфізична система розпізнавання голосу людини на базі алгоритмів машинного навчання

Автор роботи: Дмитро КРУТИЙ

Керівник роботи: Володимир ГРИГА

Пояснювальна записка: 91 с., 8 рис., 4 табл., 2 дод., 81 джерело.

КІБЕРФІЗИЧНІ СИСТЕМИ, РОЗПІЗНАВАННЯ ГОЛОСУ, ПАРАМЕТРИ ГОЛОСУ, МАШИННЕ НАВЧАННЯ, НЕЙРОННІ МЕРЕЖІ, ЦИФРОВА ОБРОБКА СИГНАЛІВ, ФУНКЦІЯ ВТРАТ.

Об’єктом дослідження є процес автоматизованої обробки та інтелектуального аналізу голосових сигналів у кіберфізичних системах.

Предметом дослідження є методи та алгоритми розпізнавання параметрів голосу людини, включаючи його акустичні характеристики на базі нейронних мереж.

Метою кваліфікаційної роботи магістра є підвищення точності систем голосової взаємодії шляхом розробки комплексної архітектури кіберфізичної системи та вдосконалення методів аналізу голосового сигналу на основі гібридних нейронних мереж.

Для розв’язання поставлених задач використовувалися методи цифрової обробки сигналів, методи математичного моделювання, методи глибокого машинного навчання, методи математичної статистики.

Наукова новизна отриманих результатів:

- набула подальшого розвитку модель процесу розпізнавання голосу в кіберфізичній системі, яка, на відміну від існуючих, базується на композиції семи функціональних відображень, що дозволяє враховувати специфіку крайових обчислень для попередньої фільтрації шумів безпосередньо на фізичному рівні.
- дістав подальшого розвитку метод розпізнавання голосових команд на базі гібридної архітектури CNN – LSTM та функції втрат CTC, що дозволило реалізувати наскрізне навчання моделі без необхідності попередньої сегментації

сигналу, підвищивши адаптивність системи до індивідуальних особливостей диктора.

На основі проведених досліджень розроблена архітектура і компоненти програмного забезпечення кіберфізичної системи розпізнавання голосу людини на базі алгоритмів машинного навчання.

Практична значимість отриманих результатів полягає у розробленій архітектурі програмного забезпечення та запропонованих рішень, що можуть бути використані при створенні інтелектуальних голосових інтерфейсів для систем промислового моніторингу, автоматизованих робочих місць та систем управління розумними об’єктами. Реалізований підхід дозволяє зменшити обчислювальне навантаження на мережу та забезпечити високу швидкість реакції системи на голосові вказівки користувача.

В першому розділі проведено огляд існуючих комерційних та відкритих систем, що дозволило виявити їхню залежність від хмарної інфраструктури та вразливість до шумів. На основі аналізу наукових публікацій обґрунтовано актуальність розробки локального гібридного рішення для кіберфізичних систем та сформульовано задачі дослідження.

В другому розділі розроблено багаторівневу структурну модель КФС, яка забезпечує взаємодію між фізичним збором акустичних даних та інтелектуальними алгоритмами обробки. Процес розпізнавання формалізовано як композицію семи функціональних відображень, а також описано математичну модель сигналу та метод виділення ознак за допомогою мел-кепстральних коефіцієнтів.

В третьому розділі виконано порівняльну класифікацію підходів, яка довела перевагу нейромережевих архітектур над класичними статистичними моделями HMM-GMM у складних акустичних умовах. Обґрунтовано та розроблено гібридний метод на базі CNN-LSTM з використанням функції втрат CTC, що дозволяє реалізувати наскрізне навчання без потреби у жорсткому вирівнюванні даних.

Спроектовано програмну архітектуру з розподілом обчислень між edge-

пристроями та сервером, а також реалізовано графічний інтерфейс для моніторингу стану системи у реальному часі. Експериментальні дослідження на наборі даних Google Speech Commands підтвердили високу ефективність рішення: точність склала 93,2%, що на 11,8% вище за класичні методи.

# ЗМІСТ

# СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАКИ

КФС - кіберфізична система

ПЗ - програмне забезпечення

API - Application Programming Interface (інтерфейс прикладного програмування)

ASR - Automatic Speech Recognition (автоматичне розпізнавання мовлення).

CNN - Convolutional Neural Network (згорткова нейронна мережа)

CTC - Connectionist Temporal Classification (зв'язкова часова класифікація)

DNN - Deep Neural Network (глибока нейронна мережа)

FAR - False Acceptance Rate (ймовірність помилкового допуску)

FRR - False Rejection Rate (ймовірність помилкового відхилення)

HMM - Hidden Markov Model (прихована марковська модель)

IoT - Internet of Things (інтернет речей)

LSTM - Long Short-Term Memory (Довга короткочасна пам'ять)

MFCC - Mel-Frequency Cepstral Coefficients (Мел-частотні кепстральні коефіцієнти)

VAD - Voice Activity Detection (Виявлення мовної активності)

# ВСТУП

Голосове керування сьогодні визнано найбільш природним та зручним способом комунікації, що дозволяє користувачеві взаємодіяти з технічними пристроями без використання фізичних маніпуляторів. Це має особливе значення для кіберфізичних систем, що функціонують у промислових середовищах, системах розумного будинку, медицині та робототехніці, де руки оператора можуть бути зайняті або доступ до пристроїв обмежений фізичними перешкодами.

Незважаючи на значні успіхи в галузі автоматичного розпізнавання голосу, впровадження таких технологій безпосередньо в архітектуру кіберфізичних систем стикається з низкою проблем. По-перше, це необхідність роботи в реальному часі при обмежених обчислювальних ресурсах периферійних (edge) пристроїв. По-друге, це високий рівень акустичних шумів та варіативність індивідуальних характеристик голосу людини, що значно знижує точність класичних алгоритмів.

Традиційні статистичні моделі, такі як приховані марковські моделі, поступово витісняються методами глибокого машинного навчання. Використання гібридних архітектур, що поєднують згорткові та рекурентні нейронні мережі, дозволяє значно підвищити точність розпізнавання за рахунок кращого виділення ознак та врахування часового контексту сигналу. Однак питання оптимізації таких моделей для таких систем та реалізація наскрізного (end-to-end) розпізнавання залишаються відкритими.

Сучасні підходи до голосової взаємодії активно переходять від класичних систем командного типу до інтелектуальних асистентів, здатних інтерпретувати природну мову та адаптуватися до контексту середовища. Це означає, що голосове керування перестає бути лише інструментом введення команд і поступово перетворюється на повноцінний інтерфейс людино-машинної взаємодії. У межах кіберфізичних систем така трансформація відкриває нові можливості для підвищення ефективності керування складними технічними

процесами, оскільки оператор отримує змогу взаємодіяти з системою інтуїтивно, без необхідності використання спеціалізованих панелей керування або додаткових пристроїв введення.

Особливої актуальності голосові інтерфейси набувають у розподілених системах, де елементи керування можуть бути фізично віддалені або інтегровані в середовища з високим рівнем автоматизації. У таких умовах голос стає найбільш природним каналом передачі інформації між людиною та системою. Крім того, розвиток Інтернету речей сприяє тому, що кількість підключених пристроїв постійно зростає, і традиційні методи взаємодії стають менш ефективними через перевантаження користувацьких інтерфейсів.

Водночас інтеграція голосових технологій у кіберфізичні системи вимагає врахування специфічних умов їх функціонування. До таких умов належать обмежені обчислювальні ресурси крайових пристроїв, необхідність мінімізації затримок обробки сигналу, а також висока динамічність середовища, в якому відбувається запис голосу. У промислових умовах, наприклад, акустичні сигнали часто спотворюються шумом обладнання, реверберацією приміщень та одночасною роботою кількох джерел звуку. Це створює додаткові виклики для алгоритмів розпізнавання, які повинні забезпечувати стабільну роботу навіть за умов низького співвідношення сигнал/шум.

Ще одним важливим аспектом є персоналізація голосових моделей. Кожен користувач має унікальні характеристики мовлення, такі як тембр, швидкість мовлення, акцент та інтонаційні особливості. Це вимагає застосування адаптивних моделей, які здатні підлаштовуватися під конкретного користувача або групу користувачів без значного зниження продуктивності системи. У цьому контексті особливо перспективними є методи перенесення навчання та донавчання моделей на невеликих наборах індивідуальних даних.

Таким чином, розвиток голосового керування в межах кіберфізичних систем потребує комплексного підходу, що поєднує досягнення в галузі обробки мовленнєвих сигналів, глибокого навчання та розподілених обчислень. Лише інтеграція цих напрямків дозволить створити ефективні, стійкі до шумів та

адаптивні системи голосової взаємодії, придатні для використання в реальних умовах експлуатації.

Актуальність роботи зумовлена в потребі розроблення систем розпізнавання голосу на базі сучасних гібридних нейронних мереж. Це дозволяє створити стійкий до перешкод та швидкий інструмент інтелектуального керування, який здатний адаптуватися до індивідуальних особливостей користувача та специфіки акустичного середовища

Таким чином, розробка кіберфізичної системи розпізнавання голосу людини на базі сучасних алгоритмів машинного навчання є актуальним завданням.

Метою кваліфікаційної роботи магістра є підвищення точності систем голосової взаємодії шляхом розробки комплексної архітектури кіберфізичної системи та вдосконалення методів аналізу голосового сигналу на основі гібридних нейронних мереж.

Поставлена мета досягається розв’язанням таких основних завдань:

- проведення аналізу сучасного стану та тенденції розвитку технологій розпізнавання голосу в контексті кіберфізичних систем;
- моделювання процесу розпізнавання голосу в кіберфізичній системі;
- розроблення методу розпізнавання голосу на основі згорткових і рекурентних нейронних мереж;
- проведення експериментального дослідження розробленого запропонованого рішення та оцінка його точності.

Об’єктом дослідження є процес автоматизованої обробки та інтелектуального аналізу голосових сигналів у кіберфізичних системах.

Предметом дослідження є методи та алгоритми розпізнавання параметрів голосу людини, включаючи його акустичні характеристики на базі нейронних мереж.

Наукова новизна отриманих результатів:

- набула подальшого розвитку модель процесу розпізнавання голосу в кіберфізичній системі, яка, на відміну від існуючих, базується на композиції семи функціональних відображень, що дозволяє враховувати специфіку крайових

обчислень для попередньої фільтрації шумів безпосередньо на фізичному рівні.

~ дістав подальшого розвитку метод розпізнавання голосових команд на базі гібридної архітектури CNN – LSTM та функції втрат CTC, що дозволило реалізувати наскрізне навчання моделі без необхідності попередньої сегментації сигналу, підвищивши адаптивність системи до індивідуальних особливостей диктора.

Практична значимість отриманих результатів полягає у розробленій архітектурі програмного забезпечення та запропонованих рішень, що можуть бути використані при створенні інтелектуальних голосових інтерфейсів для систем промислового моніторингу, автоматизованих робочих місць та систем управління розумними об’єктами. Реалізований підхід дозволяє зменшити обчислювальне навантаження на мережу та забезпечити високу швидкість реакції системи на голосові вказівки користувача.

Для розв’язання поставлених задач використовувалися методи методи цифрової обробки сигналів, методи математичного моделювання, методи глибокого машинного навчання, методи математичної статистики.

За темою кваліфікаційної роботи опубліковано одну публікацію [81] у Збірнику наукових праць за матеріалами XIX Всеукраїнської науково практичної WEB конференції аспірантів, студентів та молодих вчених «Комп’ютерні інтелектуальні системи та мережі» (25-27 березня 2026 р.). – Кривий Ріг: Криворізький національний університет, 2026. – 370 с.

# 1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ РОЗПІЗНАВАННЯ ГОЛОСУ

## 1.1 Аналіз систем розпізнавання голосу

Розпізнавання голосу є одним із ключових напрямів сучасних інформаційних технологій, що знаходиться на перетині обробки сигналів, машинного навчання та біометричних систем.

У загальному випадку під розпізнаванням голосу розуміють процес автоматичного визначення особи за її голосовими характеристиками. Такий підхід відноситься до біометричних методів ідентифікації, оскільки голос, подібно до відбитків пальців або зображення обличчя, містить унікальні ознаки, притаманні конкретній людині [1].

Важливо чітко розмежовувати поняття розпізнавання голосу та розпізнавання мовлення. Якщо розпізнавання мовлення спрямоване на визначення змісту сказаного та перетворення аудіосигналу у текстове представлення, то розпізнавання голосу фокусується саме на ідентифікації мовця незалежно від змісту висловлювання [2]. Таким чином, у першому випадку ключовим є що сказано, тоді як у другому хто говорить.

Системи розпізнавання голосу можна класифікувати за характером задачі, яку вони вирішують. Найбільш поширеними є задачі ідентифікації та верифікації спікера [3, 4].

У задачі ідентифікації система повинна визначити, кому з відомого набору користувачів належить поданий голос. Формально це можна розглядати як задачу багатокласової класифікації, де кожен клас відповідає окремій особі.

У задачі верифікації система перевіряє, чи відповідає голос заявленій особі, тобто вирішує бінарну задачу прийняття або відхилення гіпотези. Верифікація є більш поширеною у практичних застосуваннях, таких як системи доступу або аутентифікації.

Сучасні системи розпізнавання голосу все частіше використовують підходи машинного навчання, які дозволяють автоматично виділяти релевантні ознаки та будувати ефективні моделі [5, 6]. На відміну від класичних методів, де ознаки

визначалися вручну, сучасні нейромережеві підходи дозволяють навчати моделі безпосередньо на сирих або слабо оброблених даних. Це значно підвищує точність розпізнавання та дозволяє адаптувати систему до складних умов експлуатації.

Значною перевагою систем розпізнавання голосу є їх зручність використання та відсутність необхідності у спеціальному обладнанні. Для роботи достатньо стандартного мікрофона, що робить такі системи доступними для широкого кола застосувань. Вони широко використовуються у системах контролю доступу, банківських сервісах, мобільних додатках та розумних пристроях. Крім того, розпізнавання голосу є природним способом взаємодії людини з технічними системами, що підвищує зручність користування.

Разом із тим, системи розпізнавання голосу мають і певні обмеження. Однією з основних проблем є чутливість до шуму та якості запису. Зовнішні завади можуть суттєво впливати на точність розпізнавання, особливо у реальних умовах. Крім того, голос людини може змінюватися під впливом різних факторів, таких як емоційний стан, втома або захворювання, що також ускладнює задачу ідентифікації.

Ще одним викликом є забезпечення безпеки системи, зокрема захист від підробки голосу або використання записів [8].

У контексті кіберфізичних систем розпізнавання голосу набуває особливого значення, оскільки дозволяє забезпечити природну та безконтактну взаємодію користувача з системою. Інтеграція алгоритмів розпізнавання голосу у кіберфізичні системи відкриває широкі можливості для автоматизації процесів та підвищення рівня безпеки. Водночас це висуває додаткові вимоги до швидкодії, надійності та адаптивності таких систем.

Таким чином, системи розпізнавання голосу є важливим компонентом сучасних інформаційних технологій, що поєднують біометричні методи ідентифікації з можливостями машинного навчання. Їх розвиток сприяє створенню більш інтелектуальних і зручних у використанні кіберфізичних систем, здатних ефективно взаємодіяти з людиною у різних умовах.

Розглянемо більш детально існуючі рішення.

Система Voice Match, що використовується у пристроях компанії Google, є однією з найбільш поширених реалізацій технології розпізнавання голосу у споживчому сегменті [9]. Вона інтегрована у такі пристрої, як розумні колонки, смартфони та інші елементи екосистеми Google, і призначена для ідентифікації користувача за його голосовими характеристиками. Основною функцією системи є розпізнавання конкретного користувача серед кількох зареєстрованих осіб та забезпечення персоналізованого доступу до сервісів. Для цього на етапі налаштування формується голосовий профіль користувача, який зберігається у вигляді векторного представлення (embedding) і використовується для подальшого порівняння.

Принцип роботи Voice Match базується на використанні алгоритмів машинного навчання, які аналізують акустичні особливості голосу, такі як тембр, інтонація, спектральні характеристики та динаміка сигналу. Під час взаємодії користувача з пристроєм система виконує обробку вхідного аудіосигналу, виділяє релевантні ознаки та порівнює їх із збереженими профілями. У разі успішного збігу система може, наприклад, надати доступ до персональних даних, таких як календар, повідомлення або налаштування пристрою. Важливою особливістю є те, що система здатна працювати в режимі реального часу, забезпечуючи швидку реакцію на голосові запити.

Серед основних переваг Voice Match слід відзначити високу точність розпізнавання, яка досягається завдяки використанню великих обсягів навчальних даних та сучасних нейромережевих моделей. Інтеграція з екосистемою Google забезпечує широкі можливості використання, включаючи синхронізацію між пристроями та доступ до хмарних сервісів. Крім того, система оптимізована для роботи у фоновому режимі, що дозволяє реалізувати функцію активації голосом без необхідності фізичної взаємодії з пристроєм.

Разом із тим, система має певні обмеження. Одним із них є залежність від інтернет-з’єднання, оскільки значна частина обробки виконується на віддалених серверах. Це може призводити до затримок або зниження якості роботи у разі

нестабільного підключення. Іншою проблемою є чутливість до шумового середовища: у складних акустичних умовах, наприклад у громадських місцях або при наявності фонових звуків, точність розпізнавання може знижуватися. Крім того, система може мати труднощі у випадку схожих голосів або значних змін голосу одного й того ж користувача.

Система розпізнавання голосу, що використовується компанією Apple у межах голосового асистента Siri (часто умовно позначається як Voice ID або Siri Speaker Recognition), призначена для персоналізації взаємодії користувача з пристроями [10]. Вона інтегрована у такі продукти, як iPhone, iPad, HomePod та інші пристрої екосистеми Apple. Основна мета цієї технології полягає у визначенні, хто саме звертається до асистента, що дозволяє надавати індивідуалізовані відповіді, доступ до персональних даних та виконання дій від імені конкретного користувача.

Принцип роботи системи базується на аналізі біометричних характеристик голосу, які перетворюються у векторне представлення та зберігаються у вигляді голосового профілю. Під час первинного налаштування користувач вимовляє кілька фраз, на основі яких формується модель його голосу. Надалі система порівнює вхідний аудіосигнал із збереженим профілем, використовуючи алгоритми машинного навчання для визначення ступеня подібності. Важливою особливістю є те, що значна частина обробки може виконуватися безпосередньо на пристрої, що зменшує залежність від мережевого з’єднання та підвищує швидкість реагування.

Однією з ключових переваг підходу Apple є високий рівень безпеки та конфіденційності. Компанія робить акцент на локальній обробці даних і мінімізації їх передачі на віддалені сервери. Це знижує ризик витоку персональної інформації та відповідає сучасним вимогам захисту даних. Можна оптимізувати алгоритми під конкретні пристрої, що позитивно впливає на точність і стабільність роботи системи. Завдяки цьому розпізнавання голосу відбувається швидко та з мінімальними затримками.

Разом із тим, система має і певні обмеження. Одним із них є закритість

екосистеми Apple, що ускладнює інтеграцію з іншими платформами або сторонніми сервісами. Це обмежує гнучкість використання технології у різних сценаріях, особливо поза межами продуктів компанії. Крім того, можливості налаштування та адаптації системи для специфічних задач є обмеженими, оскільки користувач не має доступу до внутрішніх механізмів її роботи.

Ще одним аспектом є потенційне зниження точності у складних акустичних умовах або при значних змінах голосу користувача. Хоча система добре оптимізована для типових сценаріїв використання, її ефективність може знижуватися у разі сильного шуму або при використанні в нестандартних умовах. Також варто зазначити, що система орієнтована насамперед на персоналізацію взаємодії, а не на повноцінні задачі біометричної ідентифікації у високорівневих системах безпеки.

Система Amazon Alexa Voice Profiles призначена для ідентифікації різних користувачів у межах одного пристрою та забезпечення персоналізованої взаємодії з голосовим асистентом Alexa [11]. Вона широко застосовується у пристроях сімейства Amazon Echo та інших елементах екосистеми «розумного дому». Основна ідея полягає у створенні окремих голосових профілів для кожного користувача, що дозволяє системі розрізняти, хто саме звертається до пристрою, і відповідно адаптувати відповіді та виконувані дії.

Під час налаштування користувач проходить процедуру реєстрації голосу, у межах якої вимовляє задані фрази. На основі цих даних формується унікальне векторне представлення голосу, яке зберігається у системі. У процесі подальшої роботи Alexa аналізує вхідний аудіосигнал, виділяє характерні ознаки та порівнює їх із наявними профілями. У разі успішної ідентифікації система може надати персоналізовану інформацію, наприклад доступ до календаря, списків покупок, музичних вподобань або індивідуальних налаштувань «розумного дому».

Важливою перевагою цієї системи є підтримка кількох користувачів на одному пристрої, що робить її особливо зручною для використання у сімейному середовищі або спільних просторах. Інтеграція зі smart-home дозволяє реалізувати сценарії, у яких система реагує не лише на голосову команду, але й на те, хто її

подав. Наприклад, різні користувачі можуть мати індивідуальні налаштування освітлення, температури або доступу до певних функцій. Це значно підвищує рівень персоналізації та зручність використання кіберфізичних систем.

Разом із тим, система має низку обмежень. Одним із основних є зниження точності розпізнавання у випадках, коли голоси користувачів мають схожі характеристики. У таких ситуаціях система може помилково ідентифікувати користувача, що впливає на якість персоналізації та може створювати ризики безпеки. Крім того, ефективність роботи значною мірою залежить від якості початкового налаштування голосових профілів та умов експлуатації.

Ще одним важливим недоліком є залежність від хмарної інфраструктури. Більшість обчислень, пов’язаних із розпізнаванням голосу, виконується на віддалених серверах Amazon, що вимагає стабільного інтернет-з’єднання. У разі його відсутності або нестабільності функціональність системи може бути обмежена, а швидкість обробки знижена. Також це піднімає питання конфіденційності та захисту персональних даних, оскільки голосові дані передаються на зовнішні сервери.

Сервіс Microsoft Azure Speaker Recognition є складовою платформи хмарних обчислень Microsoft і призначений для реалізації задач ідентифікації та верифікації голосу в прикладних і корпоративних системах [12]. Він надає розробникам готові інструменти у вигляді API, що дозволяють інтегрувати функції голосової біометрії у програмні продукти без необхідності самостійної розробки складних алгоритмів машинного навчання. Сервіс підтримує два основні сценарії: визначення, хто говорить (ідентифікація), та перевірка, чи відповідає голос заявленій особі (верифікація).

Принцип роботи системи базується на формуванні голосових профілів користувачів, які створюються під час реєстрації. Для цього користувач записує кілька зразків голосу, після чого система перетворює їх у компактне векторне представлення. У процесі подальшого використання вхідний аудіосигнал аналізується, перетворюється у відповідний embedding та порівнюється з наявними профілями у базі. Завдяки використанню сучасних нейромережевих

моделей сервіс забезпечує високу точність розпізнавання навіть у складних умовах.

Однією з головних переваг Azure Speaker Recognition є його масштабованість. Оскільки обчислення виконуються у хмарній інфраструктурі, система здатна обробляти великі обсяги даних і підтримувати одночасну роботу великої кількості користувачів. Це робить її придатною для використання у корпоративних середовищах, де важливими є надійність і продуктивність. Наявність готових API значно спрощує інтеграцію сервісу у різноманітні застосунки, скорочуючи час розробки та впровадження.

Ще однією важливою перевагою є підтримка enterprise-рішень. Сервіс інтегрується з іншими компонентами екосистеми Azure, що дозволяє створювати комплексні інформаційні системи з високим рівнем безпеки та керованості. Крім того, платформа надає інструменти для моніторингу, масштабування та управління ресурсами, що є важливим для промислового використання.

Водночас сервіс має певні обмеження. Одним із них є платний доступ, що може бути суттєвим фактором при розробці невеликих або дослідницьких проєктів. Вартість використання залежить від обсягу оброблених даних та кількості запитів, що потребує ретельного планування ресурсів. Іншим недоліком є залежність від стабільного інтернет-з’єднання, оскільки всі основні обчислення виконуються на віддалених серверах. У разі перебоїв у мережі можливе зниження продуктивності або недоступність сервісу.

Також варто враховувати питання конфіденційності даних, оскільки голосові записи передаються та обробляються у хмарі. Хоча платформа забезпечує сучасні механізми захисту інформації, у деяких сценаріях (наприклад, у критичних кіберфізичних системах) це може бути обмеженням.

Kaldi Speaker Recognition Toolkit є однією з найвідоміших відкритих платформ для досліджень і розробки систем розпізнавання голосу [13]. Вона активно використовується як у науковому середовищі, так і в прикладних проєктах, що пов’язані з обробкою мовлення та голосовою біометрією. Платформа надає широкий набір інструментів для побудови повного циклу

системи розпізнавання, від обробки аудіосигналу до навчання моделей і оцінки їх ефективності.

Принцип роботи Kaldi базується на модульній архітектурі, яка дозволяє гнучко комбінувати різні алгоритми та підходи. Система підтримує як класичні статистичні методи, зокрема i-vector, так і сучасні нейромережеві підходи, такі як x-vector. Це дозволяє дослідникам експериментувати з різними моделями та знаходити оптимальні рішення для конкретних задач. Kaldi також забезпечує підтримку роботи з великими датасетами та інтеграцію з іншими інструментами машинного навчання.

Однією з ключових переваг платформи є її гнучкість. Користувач має можливість налаштовувати практично всі етапи обробки даних і навчання моделі, що робить Kaldi потужним інструментом для досліджень. Крім того, платформа забезпечує високу точність розпізнавання, оскільки реалізує сучасні алгоритми, які широко використовуються у наукових роботах. Важливим аспектом є також відкритий вихідний код, що дозволяє адаптувати систему під конкретні потреби та розширювати її функціональність.

Разом із тим, Kaldi має і суттєві недоліки. Основним із них є складність у використанні. Платформа не має зручного графічного інтерфейсу і працює переважно через командний рядок, що ускладнює її освоєння для початківців. Для ефективного використання необхідні глибокі знання у сфері обробки сигналів, машинного навчання та роботи з операційними системами типу Linux. Крім того, налаштування та запуск моделей у Kaldi потребують значного часу та зусиль. Багато процесів, таких як підготовка даних, конфігурація експериментів та оптимізація параметрів, виконуються вручну, що підвищує складність розробки. Це робить платформу менш придатною для швидкого прототипування або комерційного використання без відповідної експертизи.

LIUM Speaker Diarization System є спеціалізованою відкритою системою, призначеною для задач сегментації аудіосигналу та ідентифікації мовців у межах одного запису [14]. Основною функцією цієї системи є визначення того, «хто і коли говорить» у багатокористувацькому аудіо, що особливо актуально для

аналізу конференцій, інтерв'ю, телефонних розмов або записів із систем спостереження. На відміну від класичних систем розпізнавання голосу, які орієнтовані на ідентифікацію конкретної особи, дана система більше зосереджена на структуризації аудіопотоку та розділенні його на сегменти, що відповідають різним мовцям.

Серед основних переваг системи варто відзначити її ефективність при роботі з довготривалими аудіозаписами. Вона здатна обробляти значні обсяги даних і виділяти структуру розмови навіть у складних сценаріях із кількома учасниками. Відкритий вихідний код дозволяє адаптувати систему під конкретні задачі, а також використовувати її як базу для досліджень і експериментів. Це робить LIUM популярним інструментом у науковому середовищі.

Однак система має і ряд обмежень. Одним із основних недоліків є обмежена точність у складних акустичних умовах, зокрема при наявності шуму, перекриття мовлення або великої кількості мовців. Використання класичних статистичних методів, таких як GMM, робить систему менш ефективною порівняно з сучасними нейромережевими підходами, які краще враховують складні залежності в аудіосигналі. Крім того, LIUM не орієнтована на задачі точної біометричної ідентифікації конкретної особи, що обмежує її використання у системах контролю доступу або безпеки.

I-vector системи розпізнавання голосу є класичним підходом у задачах голосової біометрії, який тривалий час залишався стандартом у системах ідентифікації та верифікації особи [15]. Основна ідея цього підходу полягає у представленні мовного сигналу у вигляді компактного вектора фіксованої розмірності (так званого identity vector або i-vector), який відображає індивідуальні характеристики голосу користувача. Такий підхід дозволяє ефективно працювати з аудіосигналами різної тривалості, зводячи їх до уніфікованого представлення, придатного для подальшого аналізу.

Формування i-vector базується на статистичній моделі, яка описує варіації мовного сигналу у просторі ознак. Зазвичай використовується універсальна фонова модель (UBM), побудована на основі гаусових сумішей, яка апроксимує

розподіл ознак для великої кількості мовців. Далі вводиться так звана тотальна варіаційна підпросторова модель, у межах якої кожен мовний сигнал описується як відхилення від середнього значення. У результаті складний багатовимірний сигнал проєктується у низьковимірний простір, де і формується i-vector. Це значно спрощує задачу порівняння голосів, оскільки вона зводиться до обчислення відстані або подібності між векторами.

Однією з ключових переваг i-vector підходу є ефективне зменшення розмірності даних. Замість роботи з великими обсягами сирих або спектральних ознак система оперує компактними векторами, що суттєво знижує обчислювальні витрати. Це робить такі системи придатними для використання у реальному часі та в умовах обмежених ресурсів, зокрема у кіберфізичних системах. Крім того, метод добре формалізований і має зрозумілу математичну основу, що полегшує його аналіз і реалізацію.

Ще однією перевагою є відносно висока швидкість роботи. Після побудови моделі процес ідентифікації або верифікації зводиться до обчислення подібності між векторами, що може виконуватися дуже швидко. Це особливо важливо для систем, де критичною є затримка обробки, наприклад у системах контролю доступу.

Водночас i-vector системи мають низку обмежень. Одним із головних недоліків є чутливість до шуму та змін умов запису. Оскільки модель базується на статистичних припущеннях щодо розподілу ознак, вона може погано адаптуватися до нових або нестандартних умов, таких як фоновий шум, реверберація або зміни мікрофона. Це призводить до зниження точності розпізнавання у реальних сценаріях. Крім того, i-vector підхід поступово витісняється сучасними нейромережевими методами, такими як x-vector, які забезпечують кращу якість розпізнавання за рахунок автоматичного навчання ознак. Нейронні мережі здатні враховувати складні нелінійні залежності у даних, що робить їх більш стійкими до варіацій голосу та умов середовища.

X-vector нейромережеві системи розпізнавання голосу є сучасним підходом у галузі голосової біометрії, який базується на використанні глибоких нейронних

мереж для формування компактних та інформативних представлень голосу, так званих embedding-векторів [16]. На відміну від класичних методів, таких як i-vector, де ознаки формуються на основі статистичних моделей, x-vector підхід дозволяє автоматично навчати релевантні характеристики голосу безпосередньо з даних. Це забезпечує значно вищу якість розпізнавання, особливо у складних умовах.

Архітектурно x-vector система зазвичай реалізується на основі глибоких нейронних мереж типу TDNN (Time Delay Neural Network) або їх сучасних модифікацій. На вході така мережа отримує послідовність ознак (наприклад, MFCC), які описують аудіосигнал у часово-частотній області. Далі через кілька прихованих шарів відбувається поступове узагальнення інформації, а спеціальний шар статистичного агрегування (statistics pooling) дозволяє перетворити змінну за довжиною послідовність у фіксований вектор. Саме цей вектор і є x-vector компактним представленням голосу, що містить інформацію про індивідуальні характеристики мовця.

Однією з головних переваг x-vector систем є висока точність розпізнавання. Завдяки використанню глибоких нейронних мереж такі системи здатні враховувати складні нелінійні залежності в аудіосигналі, що дозволяє ефективно розрізняти навіть схожі голоси. Крім того, вони демонструють високу стійкість до варіацій, таких як зміни інтонації, швидкості мовлення, емоційного стану або умов запису. Це робить x-vector підхід придатним для використання у реальних кіберфізичних системах, де умови експлуатації можуть бути непередбачуваними.

Ще однією важливою перевагою є досягнення результатів рівня state-of-the-art у багатьох задачах голосової біометрії. У порівнянні з класичними методами, x-vector системи забезпечують значно нижчі значення помилки (зокрема, EER), що підтверджується численними науковими дослідженнями. Вони також добре масштабуються та можуть бути інтегровані у складні системи з великою кількістю користувачів.

Разом із тим, x-vector підхід має і певні недоліки. Насамперед це високі обчислювальні витрати, пов’язані з навчанням і використанням глибоких

нейронних мереж. Крім того, ефективність x-vector систем значною мірою залежить від обсягу та якості навчальних даних. Для навчання моделей потрібні великі датасети, що містять записи голосів багатьох мовців у різних умовах. Отримання таких даних може бути складним і ресурсомістким завданням. Також існує ризик зниження якості моделі при недостатній репрезентативності навчального набору.

В таблиці 1.1 наведені порівняльні характеристики систем розпізнавання голосу.

Таблиця 1.1 – Порівняльні характеристики систем розпізнавання голосу

| Система | Тип | Переваги | Недоліки |
|---|---|---|---|
| Google Voice Match | Комерційна | Висока точність, швидкість | Залежність від хмари |
| Apple Voice ID | Комерційна | Безпека, оптимізація | Закрита система |
| Amazon Alexa | Комерційна | Multi-user, smart-home | Помилки при схожих голосах |
| Microsoft Azure | Хмарна | Масштабованість | Платна |
| Kaldi | Open-source | Гнучкість, точність | Складна для користувача |
| LIUM | Open-source | Обробка записів | Нижча точність |
| i-vector | Статистична | Компактність | Чутливість до шуму |
| x-vector | Нейромережева | Висока точність | Ресурсомісткість |

Аналіз існуючих систем показав, що сучасні рішення активно використовують алгоритми машинного навчання, зокрема нейромережеві підходи, які забезпечують високу точність розпізнавання голосу. Водночас більшість комерційних систем залежить від хмарної інфраструктури, що може

бути обмеженням у кіберфізичних системах реального часу. Це обумовлює необхідність розробки ефективних локальних або гібридних рішень, що поєднують точність і низьку затримку.

## 1.2 Огляд сучасних наукових рішень у задачах розпізнавання голосу

Сучасні дослідження у сфері розпізнавання голосу активно розвиваються завдяки використанню алгоритмів глибокого навчання, які значно перевершують класичні статистичні підходи.

Дослідження [17] фокусується на створенні цілісної архітектури «розумного дому», де голос виступає основним контролером кіберфізичного простору. Автори пропонують систему, що інтегрує мікроконтролери з хмарними сервісами розпізнавання мовлення для виконання фізичних дій, як-от керування освітленням чи побутовою технікою. Головною перевагою цього рішення є його висока ергономічність та орієнтованість на користувача: система дозволяє мінімізувати фізичну взаємодію з пристроями, що є критично важливим для людей з обмеженою рухливістю або в ситуаціях, коли руки зайняті. Використання CPS-підходу забезпечує тісний зв'язок між програмним кодом та фізичним середовищем, що робить відгук системи майже миттєвим. Однак суттєвим недоліком є висока залежність від стабільності інтернет-з’єднання, оскільки обробка мовлення часто виноситься на зовнішні сервіси. Крім того, архітектура демонструє слабку стійкість до випадкових звукових подразників; система може активуватися від розмов у фоні або телевізійного шуму, що призводить до помилкових спрацювань і знижує загальну надійність експлуатації в реальних побутових умовах.

У роботі [18] автори піднімають критичне питання безпеки, яке ігнорується в багатьох споживчих системах. Вони пропонують методи захисту від складних атак підміни голосу, включаючи запис, синтез та перетворення мовлення. Основна перевага рішення полягає у розробці алгоритмів, здатних виявляти «артефакти» в аудіосигналі, які невидимі для людського вуха, але притаманні

відтвореному або згенерованому звуку. Це значно підвищує рівень довіри до голосової біометрії в критичних інфраструктурах. Проте складність пропонованих методів аналізу створює певний бар’єр для їхнього впровадження. Недоліком є висока обчислювальна вартість: для глибокого аналізу спектральних характеристик у реальному часі потрібні значні ресурси, що робить систему важкою для інтеграції в дешеві IoT-модулі. Також існує ризик «хибнопозитивних» результатів, коли через низьку якість мікрофона або специфічні умови приміщення система може заблокувати доступ справжньому власнику, сприйнявши його голос за синтезований запис.

Робота [19] представляє класичний підхід до побудови систем розпізнавання мовлення через традиційні алгоритми машинного навчання. Автори детально описують етапи вилучення ознак, таких як кепстральні коефіцієнти Mel-частоти, для подальшої класифікації. Перевагою такої системи є її передбачуваність та легкість у налаштуванні для конкретних, вузькоспеціалізованих завдань. Вона не потребує гігантських обсягів даних для навчання, що дозволяє швидко розгорнути локальне рішення для розпізнавання обмеженого набору команд. Водночас головним недоліком є обмежена масштабованість. Такі моделі демонструють чудові результати в ідеальних лабораторних умовах, але їхня точність стрімко падає при появі сторонніх шумів, зміні акценту мовця або використанні іншого типу мікрофона. Відсутність здатності до глибокого узагальнення даних робить це рішення придатним лише для простих інтерфейсів з чітко визначеним словником, обмежуючи його застосування у складних інтелектуальних системах.

Робота [20] заглиблюється в оптимізацію автоматичного розпізнавання мовлення (ASR) шляхом порівняння різних моделей машинного навчання. Автори ставлять за мету знайти баланс між швидкістю обробки та точністю декодування сигналу. Перевагою дослідження є акцент на попередній обробці даних, що дозволяє ефективно фільтрувати шуми ще до етапу класифікації, підвищуючи загальну робастність системи. Це робить їхнє рішення більш життєздатним для використання в умовах промислових приміщень або вуличного шуму. Проте

недоліком залишається статичність моделей. Традиційні алгоритми машинного навчання, розглянуті в роботі, важко адаптуються до динамічних змін мови (нових сленгових слів або специфічних термінів) без повного перенавчання системи. Це створює труднощі в довгостроковій підтримці продукту, вимагаючи постійного втручання розробників для оновлення бази знань.

Дослідження [21] зосереджено на біометричному аспекті, тобто ідентифікації особи за голосом. Автори використовують алгоритми навчання для створення «голосового відбитка» користувача. Головною перевагою є можливість забезпечити високий рівень персоналізації: система не просто виконує команду, а знає, хто саме її віддав, що дозволяє налаштовувати права доступу. Це ідеальне рішення для багатокористувацьких систем «розумного офісу» або дому. Проте суттєвим недоліком є вразливість до фізіологічних змін людини. Застуда, втома або навіть емоційний стан можуть змінити частотні характеристики голосу настільки, що система перестане впізнавати легітимного користувача. Крім того, створення надійної моделі диктора потребує тривалого процесу збору зразків голосу в різних станах, що може бути незручним для кінцевого споживача.

Автори роботи [22] проводять масштабний аналіз переходу від класичних методів до глибокого навчання (Deep Learning). Вони розглядають архітектури CNN та RNN, які кардинально змінили ринок. Основна перевага глибоких мереж – це їхня неймовірна точність і здатність розуміти контекст, а не просто окремі звуки. Такі системи здатні самостійно вивчати складні патерни в даних, що дозволяє їм працювати з безперервним мовленням будь-якої складності. Головним недоліком, описаним в огляді, є «ресурсна ненажерливість». Глибокі моделі потребують спеціалізованих графічних процесорів (GPU) та величезних масивів розмічених даних для навчання, що робить їх розробку надзвичайно дорогою. Крім того, складність нейромереж робить їх «чорними скриньками»: розробникам складно зрозуміти, чому модель припустилася конкретної помилки, що ускладнює процес тонкого налагодження.

В роботі [23] проведений огляд, який охоплює історичний шлях розвитку технологій розпізнавання мовлення. Автори детально розбирають перехід від

прихованих Марковських моделей до сучасних наскрізних (end-to-end) систем. Перевагою цього дослідження є комплексний погляд на проблему: воно визначає не лише технічні досягнення, а й майбутні виклики, як-от підтримка рідкісних мов та діалектів. Це дає розробникам дорожню карту для створення універсальних систем. Основним недоліком поточної стадії розвитку, згідно з авторами, є відсутність справжньої «мовної інтуїції» у машин. Попри високу точність, системи все ще часто помиляються в омонімах або в іронічному контексті, оскільки вони спираються на статистичні ймовірності, а не на реальне розуміння змісту, що залишає простір для вдосконалення алгоритмів обробки природної мови (NLP).

## 1.3 Постановка задачі

Проведений аналіз наукових робіт показує, що основним напрямом розвитку систем розпізнавання голосу є використання глибоких нейронних мереж та embedding-представлень. Найбільш ефективними є підходи на основі x-vector та їх модифікацій, які забезпечують високу точність і стійкість до варіацій сигналу. Водночас актуальною залишається проблема зменшення обчислювальної складності, що є критично важливим для кіберфізичних систем. Це обґрунтовує доцільність розробки оптимізованих гібридних або легковагових моделей у межах даної роботи.

Отже, для досягнення мети магістерської роботи необхідно розв’язати наступні завдання:

- провести аналіз сучасного стану та тенденції розвитку технологій розпізнавання голосу в контексті кіберфізичних систем;
- змоделювати процес розпізнавання голосу в кіберфізичній системі;
- розробити метод розпізнавання голосу на основі згорткових і рекурентних нейронних мереж;
- провести експериментальне дослідження розробленого запропонованого рішення та оцінити його точність.

## 1.4 Висновки до першого розділу

У першому розділі проведено комплексний аналіз предметної галузі розпізнавання голосу людини, розглянуто існуючі комерційні та відкриті системи, а також виконано огляд сучасних наукових розробок.

Розгляд провідних комерційних систем показав їх високу ефективність, проте виявив значну залежність від стабільного інтернет-з’єднання та хмарної інфраструктури. Відкриті інструменти, такі як Kaldi та i-vector системи, хоч і забезпечують гнучкість, є складними у налаштуванні або виявляють високу чутливість до сторонніх акустичних шумів. Огляд сучасних досліджень підтвердив домінування методів глибокого навчання над класичними статистичними підходами.

Основними викликами для впровадження голосового інтерфейсу в кіберфізичні системи залишаються:

1. Необхідність роботи в реальному часі при обмежених ресурсах периферійних пристроїв.
2. Вразливість систем до зашумленого середовища та варіативності індивідуальної вимови.
3. Ризики безпеки, пов'язані зі спуфінг-атаками (підробкою голосу).

На основі проведеного аналізу доведено доцільність розробки кіберфізичної системи, яка поєднує локальну попередню обробку сигналу для фільтрації шумів із нейромережевими методами аналізу. Це дозволить забезпечити баланс між точністю ідентифікації та швидкістю реакції системи.

# 2 МОДЕЛЮВАННЯ КІБЕРФІЗИЧНОЇ СИСТЕМИ РОЗПІЗНАВАННЯ ГОЛОСУ ЛЮДИНИ

## 2.1 Структурна модель кіберфізичної системи розпізнавання голосу

Кіберфізична система розпізнавання голосу є інтегрованою багаторівневою системою, що поєднує фізичні пристрої збору даних, комунікаційну інфраструктуру та інтелектуальні алгоритми обробки. Її основною метою є перетворення акустичних сигналів (мовлення людини) у цифрову інформацію з подальшою інтерпретацією та виконанням відповідних дій (рисунок 2.1).


Рисунок 2.1 – Структурна схема кіберфізичної системи розпізнавання голосу

У загальному вигляді система складається з таких рівнів:

- фізичного;
- мережевого/комунікаційного;
- інтелектуального або кібернетичного;
- рівня виконавчих механізмів, який іноді розглядається як окремий або частина кібер-рівня.

Фізичний рівень кіберфізичної системи розпізнавання голосу виконує функцію збору первинної інформації із зовнішнього середовища, а саме акустичних сигналів мовлення користувача. Основу цього рівня складають мікрофонні системи, які можуть бути представлені як одиничними мікрофонами, так і складнішими мікрофонними решітками.

Використання мікрофонних решіток дозволяє суттєво підвищити якість захоплення звуку за рахунок просторової фільтрації сигналу, зокрема реалізації технологій формування променя, що дає змогу виділяти голос користувача навіть у шумному середовищі. Отриманий аналоговий сигнал надходить до модуля аналогово-цифрового перетворення, де здійснюється його дискретизація та квантування. Важливими характеристиками цього процесу є частота дискретизації та розрядність, які безпосередньо впливають на точність подальшої обробки мовлення.

Після оцифрування сигнал проходить етап попередньої обробки, що включає фільтрацію шумів, компенсацію акустичного ехо, нормалізацію рівня сигналу та виявлення мовної активності. Ці процедури дозволяють підвищити співвідношення сигнал/шум і зменшити обсяг даних, які необхідно передавати та обробляти.

Важливу роль на цьому рівні відіграють крайові обчислювальні пристрої, які виконують локальну обробку сигналу. Вони можуть здійснювати попередній аналіз, стиснення аудіоданих або навіть базове розпізнавання мовлення, що дозволяє зменшити затримки та навантаження на мережу.

Таким чином, фізичний рівень забезпечує підготовку якісного цифрового сигналу для подальшої передачі та обробки в системі.

Мережевий рівень кіберфізичної системи відповідає за надійну та ефективну передачу даних між фізичними компонентами системи та обчислювальними ресурсами, на яких виконується обробка інформації.

Передача аудіоданих може здійснюватися як через провідні канали зв’язку, такі як Ethernet або USB, так і через безпровідні технології, включаючи Wi-Fi, Bluetooth або мобільні мережі зв’язку. Вибір конкретної технології залежить від

вимог до швидкості передачі, затримки, енергоспоживання та мобільності системи.

У процесі передачі даних використовуються різноманітні протоколи, які забезпечують структуровану взаємодію між компонентами системи. Наприклад, протокол MQTT широко застосовується в IoT-системах завдяки своїй легковаговості та ефективності, тоді як HTTP або HTTPS використовуються для взаємодії з хмарними сервісами [26, 27]. Для передачі аудіо в реальному часі можуть застосовуватись потокові протоколи, такі як WebSocket або RTP, які забезпечують мінімальні затримки та стабільність потоку даних [28, 29].

Важливим аспектом мережевого рівня є також буферизація та керування потоками даних, що дозволяє уникнути втрат інформації при нестабільному з’єднанні.

Окрему увагу приділяють питанням безпеки, оскільки аудіодані можуть містити конфіденційну інформацію. Для цього застосовуються механізми шифрування, зокрема протокол TLS, а також системи аутентифікації пристроїв і контролю доступу.

Таким чином, мережевий рівень забезпечує не лише передачу даних, але й їхню цілісність, конфіденційність та доступність.

Інтелектуальний рівень є центральним елементом кіберфізичної системи розпізнавання голосу, оскільки саме на цьому рівні здійснюється аналіз аудіосигналу та перетворення його у текстову або семантичну інформацію. Обробка може виконуватися на різних обчислювальних платформах, включаючи хмарні сервери, туманні обчислювальні вузли або локальні edge-пристрої. Використання хмарних технологій забезпечує високу обчислювальну потужність і можливість застосування складних моделей машинного навчання, тоді як edge-обчислення дозволяють зменшити затримки та підвищити автономність системи.

Першим етапом обробки є перетворення аудіосигналу у форму, придатну для аналізу алгоритмами машинного навчання. Для цього використовуються спектральні представлення сигналу, такі як MFCC або спектрограми, які відображають частотні характеристики мовлення.

Далі ці дані подаються на вхід моделей автоматичного розпізнавання мовлення, які можуть бути реалізовані на основі різних підходів, включаючи нейронні мережі глибокого навчання, рекурентні мережі або трансформерні архітектури. Результатом роботи цих моделей є текстове представлення мовлення.

Наступним етапом є обробка природної мови, яка дозволяє інтерпретувати зміст отриманого тексту.

На цьому етапі виконуються задачі визначення наміру користувача та виділення ключових сутностей. Отримана інформація передається до модуля прийняття рішень, який на основі заданих правил або алгоритмів формує відповідну реакцію системи.

Таким чином, інтелектуальний рівень забезпечує перехід від сирого аудіосигналу до осмислених дій.

Рівень виконавчих механізмів призначений для реалізації дій у відповідь на розпізнані голосові команди користувача. Цей рівень безпосередньо взаємодіє з фізичними або програмними об’єктами, забезпечуючи виконання команд, сформованих на інтелектуальному рівні. У контексті систем розумного будинку це може бути керування освітленням, температурою або побутовими приладами. У промислових системах це може включати управління технологічними процесами або обладнанням.

Крім того, важливою складовою цього рівня є інтерфейси взаємодії з користувачем. Система може надавати зворотний зв’язок у вигляді синтезованого мовлення, текстових повідомлень або графічного інтерфейсу. Це дозволяє користувачу отримувати підтвердження виконання команд або додаткову інформацію.

Також на цьому рівні можуть реалізовуватися системи контролю доступу, де голос використовується як біометричний ідентифікатор. У разі необхідності система може генерувати сповіщення через різні канали зв’язку, такі як мобільні додатки, електронна пошта або повідомлення.

Таким чином, рівень виконавчих механізмів є завершальним етапом функціонування кіберфізичної системи, який забезпечує практичну реалізацію її

роботи та взаємодію з користувачем або навколишнім середовищем.

## 2.2 Модель процесу розпізнавання голосу в кіберфізичній системі

Процес розпізнавання голосу в кіберфізичній системі являє собою послідовність взаємопов’язаних етапів, у межах яких акустичний сигнал, що генерується користувачем, перетворюється у цифрові дані, аналізується за допомогою алгоритмів машинного навчання та трансформується у керуючі дії або інформаційні відповіді. Така модель відображає інтеграцію фізичних процесів із кібернетичними компонентами, що забезпечує інтелектуальну взаємодію людини з технічною системою.

На початковому етапі відбувається формування мовного сигналу, який поширюється у вигляді звукових хвиль у фізичному середовищі. Ці хвилі сприймаються мікрофонними пристроями, що входять до складу фізичного рівня системи. Мікрофони перетворюють акустичні коливання в аналоговий електричний сигнал, який далі підлягає оцифруванню за допомогою аналогово-цифрового перетворювача. У результаті формується цифровий сигнал, придатний для подальшої комп’ютерної обробки.

Після цього виконується попередня обробка сигналу, метою якої є підвищення якості даних і зменшення впливу зовнішніх перешкод. На цьому етапі здійснюється фільтрація шумів, нормалізація амплітуди сигналу, усунення ехо та визначення ділянок, що містять мовлення. Виявлення мовної активності дозволяє відокремити корисний сигнал від пауз або фонових шумів, що значно підвищує ефективність наступних етапів обробки. Оброблений сигнал може частково аналізуватися на крайових пристроях, що дає змогу скоротити обсяг переданих даних і зменшити затримки.

Наступним кроком є передача даних через мережеву інфраструктуру до обчислювального середовища, де відбувається основна обробка. Передача може здійснюватися у вигляді потокового аудіо або пакетів даних із використанням відповідних протоколів зв’язку.

Важливими характеристиками цього етапу є забезпечення низької затримки, цілісності даних та їх захищеності від несанкціонованого доступу.

На інтелектуальному рівні система виконує перетворення аудіосигналу у набір ознак, що характеризують мовлення. Зазвичай використовуються спектральні характеристики, які відображають розподіл енергії сигналу за частотами.

Отримані ознаки подаються на вхід моделі автоматичного розпізнавання мовлення, яка перетворює їх у текстову форму.

Сучасні системи використовують глибокі нейронні мережі, здатні враховувати часову структуру мовлення та контекст.

Після отримання текстового представлення виконується його інтерпретація за допомогою методів обробки природної мови. На цьому етапі визначається зміст висловлювання, виявляється намір користувача та виділяються ключові параметри команди.

Цей процес дозволяє системі перейти від простого розпізнавання слів до розуміння їхнього значення.

Далі результати аналізу надходять до модуля прийняття рішень, який визначає, яку саме дію необхідно виконати. Це може бути як виконання конкретної команди, так і формування відповіді користувачу.

Прийняття рішень може базуватися на заздалегідь визначених правилах, сценаріях або адаптивних алгоритмах, що враховують контекст використання системи.

Завершальним етапом є реалізація дії або формування зворотного зв’язку. Система може активувати виконавчі механізми, змінювати стан пристроїв або генерувати відповідь у вигляді синтезованого мовлення чи повідомлення.

Таким чином, процес розпізнавання голосу завершується повним циклом взаємодії між користувачем і системою.

Процес розпізнавання голосу в КФС доцільно представити як впорядковану систему множин і відображень між ними.

У загальному вигляді система може бути у вигляді 2.1, де кожна множина

відповідає певному етапу обробки інформації:

$$S = \langle X,A,F,M,T,I,D,Y \rangle, \tag{2.1}$$

де $X$ – множина вхідних сигналів;

A – множина оцифрованих сигналів;

$F$ – множина попередньо оброблених сигналів;

$M$ – множина ознак;

$T$ – множина текстових представлень;

$I$ – множина інтерпретацій (семантика);

$D$ – множина рішень;

$Y$ – множина вихідних дій.

Множина вхідних сигналів $X$ описує всі можливі вхідні акустичні сигнали, що надходять до системи з навколишнього середовища (формула 2.2):

$$X = \{x(t)\}, \tag{2.2}$$

де $x(t)$ – неперервна функція часу, яка відображає зміну звукового тиску, створеного мовленням користувача.

Ці сигнали мають аналогову природу та характеризуються широким спектром частот, амплітуд і часових залежностей.

Важливою особливістю цієї множини є її стохастичний характер, оскільки реальні мовні сигнали завжди супроводжуються шумами, реверберацією та іншими спотвореннями.

До складу $X$ можуть входити як чисті мовні сигнали, так і сигнали з різним рівнем завад, що обумовлює необхідність подальшої обробки.

Таким чином, множина $X$ є початковою точкою всього процесу та визначає якість функціонування системи в цілому.

Множина оцифрованих сигналів A містить дискретизовані сигнали,

отримані в результаті аналого-цифрового перетворення (формула 2.3):

$$A = \{a[n]\}, \tag{2.3}$$

де $a[n]$ – послідовність відліків, які представляють значення сигналу у дискретні моменти часу.

Цей перехід від неперервного представлення до дискретного є необхідним для подальшої цифрової обробки.

Процес формування множини A визначається параметрами дискретизації, такими як частота дискретизації та розрядність квантування. Від правильного вибору цих параметрів залежить точність відтворення сигналу та ефективність подальших алгоритмів. Множина A фактично виступає мостом між фізичним та кібернетичним рівнями системи.

Множина попередньо оброблених сигналів F включає сигнали, які пройшли етап попередньої обробки (формула 2.4):

$$F = \{f[n]\}, \tag{2.3}$$

де $f[n]$ – результат застосування до дискретного сигналу різноманітних алгоритмів, спрямованих на покращення якості даних.

До алгоритмів, спрямованих на покращення якості даних, належать фільтрація шумів, нормалізація, компенсація ехо та виділення мовної активності [32 – 34].

Цей етап є критично важливим, оскільки саме тут формується сигнал, придатний для виділення інформативних ознак. Якщо попередня обробка виконана неякісно, це призводить до накопичення помилок на наступних етапах.

До алгоритмів, спрямованих на покращення якості аудіоданих у системах розпізнавання голосу, належить комплекс методів попередньої обробки сигналу, кожен із яких виконує специфічну функцію підвищення інформативності

мовлення та зменшення впливу небажаних факторів.

Одним із ключових напрямів є фільтрація шумів, яка полягає у видаленні або пригніченні компонентів сигналу, що не несуть корисної інформації. У реальних умовах мовлення майже завжди супроводжується фоновими шумами, такими як звуки довкілля, робота техніки або інші голоси.

Для боротьби з цим використовуються як класичні методи, наприклад лінійна фільтрація у частотній області, так і більш складні адаптивні алгоритми.

Серед них особливе місце займає спектральне віднімання, при якому оцінюється спектр шуму і віднімається від спектра сигналу, а також методи на основі вейвлет-перетворень і нейронних мереж, які здатні відокремлювати мовлення від шуму навіть у складних умовах. Результатом є сигнал із покращеним співвідношенням сигнал/шум, що суттєво підвищує точність подальшого розпізнавання.

Не менш важливим є процес нормалізації сигналу, який забезпечує приведення амплітуди аудіосигналу до певного стандартного рівня. У природних умовах гучність мовлення може суттєво варіюватися залежно від відстані до мікрофона, індивідуальних особливостей мовця або характеристик обладнання.

Нормалізація дозволяє усунути ці варіації шляхом масштабування сигналу, що забезпечує стабільність подальших обчислень. У більш складних реалізаціях застосовується не лише амплітудна нормалізація, але й методи, які враховують статистичні властивості сигналу, наприклад, приведення до нульового середнього значення та одиничної дисперсії. Це особливо важливо для алгоритмів машинного навчання, які чутливі до масштабу вхідних даних.

Окрему роль відіграє компенсація ехо, яка спрямована на усунення повторних відбиттів звукового сигналу. Ехо виникає внаслідок відбиття звуку від поверхонь у приміщенні або через повторне захоплення сигналу, який відтворюється динаміками системи. Це призводить до накладання кількох копій сигналу з різними затримками, що значно ускладнює його аналіз.

Для вирішення цієї проблеми застосовуються адаптивні фільтри, які оцінюють імпульсну характеристику середовища та віднімають відбитий сигнал із

вхідного. Сучасні методи можуть враховувати змінні умови середовища та динамічно підлаштовувати свої параметри, що забезпечує ефективне придушення ехо навіть у складних акустичних умовах.

Важливим етапом попередньої обробки є також виділення мовної активності, яке полягає у визначенні ділянок сигналу, що містять мовлення. У типовому аудіосигналі значна частина часу може припадати на паузи або фонові шуми, які не несуть корисної інформації.

Алгоритми виявлення мовної активності аналізують такі характеристики, як енергія сигналу, спектральний склад або статистичні параметри, щоб відокремити мовлення від тиші. У більш складних системах використовуються моделі машинного навчання, які здатні точно визначати наявність мовлення навіть у шумному середовищі. Це дозволяє не лише зменшити обсяг оброблюваних даних, але й підвищити швидкодію системи та точність розпізнавання.

Сукупність зазначених методів формує ефективний механізм попередньої обробки аудіосигналу, який забезпечує підготовку якісних даних для наступних етапів аналізу. Кожен із методів виконує свою функцію, але їх спільне застосування дозволяє досягти значно кращих результатів, ніж використання окремих підходів, що є критично важливим для сучасних кіберфізичних систем розпізнавання голосу.

Таким чином, множина $F$ є очищеним та підготовленим представленням аудіосигналу.

Множина ознак M містить вектори ознак, які описують суттєві характеристики мовного сигналу (формула 2.4):

$$M = \{m_k\}, \tag{2.4}$$

де $m_k$ – багатовимірний вектор, що формується шляхом перетворення сигналу у спектральну або іншу інформативну область. Найчастіше використовуються такі представлення, як MFCC або спектрограми.

Особливістю цієї множини є значне зменшення розмірності даних при

збереженні ключової інформації про мовлення. Саме на цьому етапі відбувається перехід від так званих сирих сигналів до структурованих даних, які можуть ефективно оброблятися алгоритмами машинного навчання. Сирий аудіосигнал, який надходить після попередньої обробки, зазвичай являє собою довгу послідовність відліків у часовій області. Такий сигнал містить велику кількість даних, значна частина яких є надлишковою або не має прямого відношення до розпізнавання мовлення. Наприклад, сусідні відліки сильно корельовані між собою, а інформація про мовлення розподілена у складній формі, яка не є безпосередньо придатною для аналізу.

Процес зменшення розмірності починається з переходу від часової області до частотної або часово-частотної. Це дозволяє представити сигнал у вигляді спектральних характеристик, які краще відображають фізичну природу мовлення. Мовний сигнал формується внаслідок коливань голосових зв’язок і резонансів голосового тракту, тому його ключові властивості проявляються саме у частотній структурі. Використовуючи перетворення, такі як короткочасне перетворення Фур’є, сигнал розбивається на короткі фрагменти (фрейми), для кожного з яких обчислюється спектр [35]. Це дозволяє отримати інформацію про зміну частотного складу мовлення у часі.

Далі застосовується додаткове узагальнення та стиск інформації. Одним із найпоширеніших підходів є використання кепстральних коефіцієнтів, які фактично представляють спектр у ще більш компактній формі. При цьому відбувається відбір лише найбільш значущих компонентів, які відповідають загальній формі спектра, тоді як дрібні деталі та шумові складові відкидаються. Таким чином, із тисяч початкових відліків формується відносно невеликий вектор ознак, який зберігає основні характеристики мовлення, такі як тембр, артикуляція та фонетичні особливості.

Важливим аспектом цього етапу є не лише зменшення обсягу даних, але й підвищення їх інформативності. Ознаки формуються таким чином, щоб бути максимально стійкими до змін умов запису, шумів та індивідуальних особливостей мовців, одночасно зберігаючи відмінності між різними звуками та

словами. Це досягається завдяки використанню психоакустичних моделей, які враховують особливості сприйняття звуку людиною, наприклад нелінійність шкали частот. У результаті ознаки краще відповідають тим характеристикам, які є важливими для розпізнавання.

Ще одним важливим моментом є те, що структуроване представлення даних дозволяє ефективніше застосовувати алгоритми машинного навчання. Нейронні мережі та інші моделі працюють значно краще, коли вхідні дані мають фіксовану розмірність і зрозумілу структуру. Вектори ознак можна розглядати як точки у багатовимірному просторі, де відстані між ними відображають схожість або відмінність мовних елементів. Це спрощує задачу класифікації та дозволяє моделям швидше навчатися і досягати вищої точності.

Таким чином, етап формування множини ознак є ключовим у всьому процесі розпізнавання голосу, оскільки саме тут відбувається перехід від сирого сигналу до компактного, інформативного та структурованого представлення даних. Це представлення одночасно зменшує обчислювальні витрати та підвищує ефективність подальшого аналізу, що є критично важливим для роботи кіберфізичних систем у реальному часі.

Множина $M$ є вхідною для моделей розпізнавання мовлення.

Множина текстових представлень T складається з текстових результатів розпізнавання мовлення (формула 2.5):

$$T = \{t_i\}, \tag{2.5}$$

де $t_i$ – послідовність символів або слів, яка відповідає вимовленій користувачем фразі.

Цей етап є результатом роботи моделей автоматичного розпізнавання голосових сигналів. Після формування векторів ознак система передає їх на вхід моделі, яка навчена встановлювати відповідність між акустичними характеристиками мовлення та мовними одиницями, такими як фонеми, склади або слова. Цей процес базується на складних статистичних або нейромережевих

залежностях, що дозволяють моделі враховувати як локальні особливості сигналу, так і його часову структуру.

Сучасні моделі автоматичного розпізнавання голосу здатні працювати з послідовностями змінної довжини та враховувати контекст, що є критично важливим для коректного розпізнавання. На відміну від ранніх підходів, де використовувалися жорстко задані правила та прості ймовірнісні моделі, сучасні системи застосовують глибокі нейронні мережі, які автоматично виділяють складні залежності у даних. Це дозволяє значно підвищити точність розпізнавання навіть у складних умовах, наприклад при наявності шумів або різних акцентів.

Процес розпізнавання включає зіставлення послідовності ознак із найбільш ймовірною послідовністю мовних елементів. При цьому модель оцінює різні варіанти інтерпретації сигналу та обирає той, що має найвищу ймовірність. У багатьох системах додатково використовується мовна модель, яка враховує статистичні закономірності мови і допомагає вибрати найбільш логічний варіант тексту. Це особливо важливо у випадках, коли акустичний сигнал є неоднозначним.

Результатом цього етапу є текстове представлення мовлення, яке може містити як окремі слова, так і повні речення. Якість цього результату залежить від багатьох факторів, зокрема якості попередніх етапів обробки, характеристик навчальних даних і складності самої моделі. Незважаючи на можливі помилки, саме цей етап забезпечує основу для подальшого семантичного аналізу, оскільки перетворює безперервний аудіосигнал у дискретну форму, зрозумілу для алгоритмів обробки природної мови.

Таким чином, етап автоматичного розпізнавання мовлення виконує функцію інтерпретації акустичних даних у текстову форму, що є необхідною умовою для подальшої інтелектуальної обробки та прийняття рішень у кіберфізичній системі.

Якість елементів множини $T$ залежить від точності моделі, якості вхідних даних та складності мовлення. Помилки на цьому етапі можуть проявлятися у

вигляді неправильних слів або пропусків. Незважаючи на це, саме множина $T$ є основою для подальшого семантичного аналізу.

Множина інтерпретацій $I$ містить структуровані інтерпретації тексту, отриманого після розпізнавання мовлення (формула 2.6):

$$I = \{i_j\}, \tag{2.6}$$

де $i_j$ – елемент, що відображає намір користувача, а також пов’язані з ним параметри (сутності). Це може бути, наприклад, команда включення пристрою або запит на отримання інформації.

Цей етап передбачає використання методів обробки природної мови, які дозволяють перейти від формального тексту до його змістовного розуміння [37, 38].

Якщо на попередньому етапі система лише трансформує акустичний сигнал у послідовність слів, то на цьому рівні вона намагається інтерпретувати значення цієї послідовності, враховуючи контекст, структуру мови та намір користувача.

Отриманий текст сам по собі є лише набором символів або слів, які не містять явної інформації про те, яку дію необхідно виконати. Тому першим кроком є лінгвістичний аналіз, що включає розбиття тексту на складові елементи, визначення частин мови, синтаксичну структуру речення та зв’язки між словами. Це дозволяє системі зрозуміти граматичну організацію висловлювання і виділити ключові компоненти, які несуть основне смислове навантаження.

Наступним етапом є семантичний аналіз, у ході якого визначається значення висловлювання. Тут система намагається встановити, що саме мав на увазі користувач, тобто визначити його намір. Наприклад, одна і та сама фраза може мати різні значення залежно від контексту, інтонації або попередніх взаємодій. Для вирішення цієї задачі застосовуються моделі, які враховують контекст і здатні узагальнювати інформацію, виділяючи сутності, такі як об’єкти, дії, параметри або часові характеристики.

Особливу роль відіграє процес виділення сутностей, який дозволяє структурувати інформацію, що міститься у тексті. Наприклад, у команді користувача можуть бути присутні назви пристроїв, значення параметрів або часові інтервали. Виділення цих елементів дозволяє перетворити неструктурований текст у формалізоване представлення, яке може бути використане для прийняття рішень. Таким чином, текст трансформується у набір параметрів і команд, зрозумілих системі.

Крім того, сучасні методи обробки природної мови враховують контекст взаємодії, що дозволяє системі працювати в режимі діалогу. Це означає, що інтерпретація поточного запиту може залежати від попередніх команд або стану системи. Такий підхід значно підвищує природність взаємодії та робить систему більш гнучкою і адаптивною.

У результаті виконання цього етапу формується структуроване представлення змісту висловлювання, яке включає намір користувача та пов’язані з ним параметри. Це представлення є основою для подальшого прийняття рішень і виконання дій. Таким чином, методи обробки природної мови забезпечують ключовий перехід від синтаксичного рівня до семантичного, що є необхідною умовою для створення інтелектуальних кіберфізичних систем.

Множина $I$ відіграє ключову роль у забезпеченні інтелектуальної складової системи, оскільки саме тут відбувається інтерпретація намірів користувача.

Множина рішень $D$ описує можливі рішення, які приймає система на основі інтерпретації (формула 2.7):

$$D = \{d_l\}, \tag{2.7}$$

де $d_l$ – елемент, що є результатом аналізу наміру користувача та може представляти конкретну команду або сценарій дій.

Формування множини $D$ базується на логічних правилах, базах знань або алгоритмах прийняття рішень. У складніших системах можуть використовуватися

адаптивні або навчальні підходи.

Таким чином, множина $D$ виступає як проміжна ланка між аналізом інформації та її практичним застосуванням.

Множина вихідних дій $Y$ включає всі можливі дії або відповіді системи (формула 2.8):

$$Y = \{y_g\}, \tag{2.8}$$

де $y_g$ – елемент, що може бути як керуючим сигналом для фізичного пристрою, так і інформаційним повідомленням для користувача.

Це завершальний етап функціонування системи. Важливою характеристикою цієї множини є її залежність від прикладної області. Наприклад, у системах розумного будинку це можуть бути команди керування пристроями, тоді як у інформаційних системах – текстові або голосові відповіді.

Таким чином, множина $Y$ реалізує взаємодію системи з зовнішнім середовищем.

## 2.3 Формалізація процесу розпізнавання голосу

Процес розпізнавання голосу в кіберфізичній системі може бути формально представлений як послідовна композиція функціональних відображень, кожне з яких відповідає окремому етапу перетворення інформації. Такий підхід дозволяє розглядати систему не просто як набір компонентів, а як впорядковану математичну структуру, у якій результат кожного етапу є вхідними даними для наступного.

Загальна модель може бути записана у вигляді композиції функцій вигляду 2.9:

$$Y = f_7(f_6(f_5(f_4(f_3(f_2(f_1(x)))))))), \tag{2.9}$$

де $f_i$ - функція, що реалізує певний функціональний блок системи.

Цей вираз відображає поетапне перетворення вхідного сигналу $x \in X$ у вихідну дію $y \in Y$.

На першому етапі виконується відображення $f_1$: $X \to A$, яке відповідає процесу аналого-цифрового перетворення сигналу. Вхідний елемент $x(t)$, що є неперервним акустичним сигналом, перетворюється у дискретну послідовність $a[n]$.

Це відображення включає операції дискретизації та квантування, які забезпечують можливість подальшої цифрової обробки. Функція $f_1$ є критичною, оскільки саме на цьому етапі визначається точність представлення сигналу в цифровій формі.

Наступне відображення $f_2$: $A \to F$ реалізує попередню обробку сигналу. Воно включає фільтрацію шумів, нормалізацію амплітуди, компенсацію ехо та виділення мовної активності. У результаті формується сигнал $f$ $[n]$, який має значно кращі характеристики для подальшого аналізу. Функція $f_2$ виконує роль очищення та підготовки даних, зменшуючи вплив зовнішніх завад і покращуючи співвідношення сигнал/шум.

Третє відображення $f_3$: $F \to M$ відповідає етапу виділення ознак. На цьому кроці сигнал перетворюється у компактне представлення у вигляді векторів ознак $m_k$. Це можуть бути спектральні характеристики, які відображають частотну структуру мовлення. Основною метою функції $f_3$ є зменшення розмірності даних при збереженні максимальної кількості корисної інформації, необхідної для розпізнавання.

Четверте відображення $f_4$: $M \to T$ реалізує процес автоматичного розпізнавання мовлення. Воно базується на використанні моделей машинного навчання, які аналізують вектори ознак і перетворюють їх у текстове представлення. Результатом є елемент $t_i$, що відповідає розпізнаній фразі. Це один із найскладніших етапів, оскільки він потребує врахування часових залежностей, варіативності мовлення та контексту.

П’яте відображення $f_5 : T \to I$ виконує семантичний аналіз тексту. На цьому етапі система переходить від синтаксичного рівня до змістовного, визначаючи намір користувача та виділяючи ключові сутності. Функція $f_5$ реалізує алгоритми обробки природної мови, що дозволяють інтерпретувати текст як команду або запит.

Шосте відображення $f_6 : I \to D$ відповідає процесу прийняття рішення. На основі інтерпретації формується конкретне рішення $d_l$, яке визначає подальшу поведінку системи. Це відображення може включати логічні правила, сценарії або адаптивні алгоритми, що враховують контекст і стан системи.

Завершальне відображення $f_7 : D \to Y$ реалізує виконання дії або формування відповіді. На цьому етапі система генерує керуючий сигнал або інформаційне повідомлення, яке передається користувачу або виконавчим механізмам. Результатом є елемент $y_g$, що відображає реакцію системи на вхідний сигнал.

Таким чином, композиція функцій виду 2.9 утворює єдиний безперервний процес обробки інформації, у якому кожен етап логічно пов’язаний із попереднім і наступним.

Важливою властивістю цієї моделі є те, що похибки або втрати інформації на будь-якому етапі можуть впливати на кінцевий результат, що підкреслює необхідність оптимізації кожного відображення окремо.

Крім того, така формалізація дозволяє розглядати систему з точки зору математичного аналізу, що відкриває можливості для оцінювання ефективності, стійкості до шумів і масштабованості. Вона також спрощує проєктування системи, оскільки кожне відображення може бути реалізоване як окремий модуль із чітко визначеними входами та виходами.

## 2.4 Математична модель сигналу та виділення ознак

У кіберфізичних системах розпізнавання голосу математичне моделювання сигналу є основою для подальшої ефективної обробки та аналізу мовлення. Акустичний сигнал, що генерується мовним апаратом людини, можна розглядати як результат складного фізичного процесу, який включає коливання голосових зв’язок та резонансні властивості голосового тракту.

З математичної точки зору цей сигнал доцільно представити як неперервну функцію часу x(t), яка описує зміну звукового тиску. Однак для практичної реалізації в цифрових системах цей сигнал підлягає дискретизації, у результаті чого формується послідовність відліків $x[n]$.

Однією з ключових моделей мовного сигналу є модель джерело–фільтр, згідно з якою мовлення розглядається як результат згортки збуджуючого сигналу (джерела) з імпульсною характеристикою голосового тракту (фільтра) [39, 40].

У дискретному вигляді це можна записати наступним чином, формула 2.10:

$$x[n] = e[n] * h[n], \tag{2.10}$$

де $e[n]$ – сигнал збудження (наприклад, періодичний для голосних звуків або шумоподібний для приголосних);
$h[n]$ – імпульсна характеристика голосового тракту.

Така модель дозволяє розділити властивості джерела та артикуляційні особливості мовлення, що є важливим для подальшого аналізу.

Оскільки мовний сигнал є нестаціонарним, його обробка виконується у коротких часових інтервалах, у межах яких сигнал вважається квазістаціонарним.

Для цього сигнал розбивається на фрейми тривалістю 20 – 30 мс із частковим перекриттям. Кожен фрейм множиться на віконну функцію, наприклад вікно Хеммінга, що дозволяє зменшити ефекти розривів на межах. У результаті формується послідовність сегментів, придатних для подальшого спектрального аналізу.

Наступним етапом є перехід до частотного представлення сигналу. Для цього використовується дискретне перетворення Фур'є (формула 2.11), яке дозволяє отримати спектр сигналу:

$$X(k) = \sum_{n=0}^{N-1} \mathrm{x}[\mathrm{n}] * e^{-j2\pi kn/N}, \quad (2.11)$$

Спектральне представлення є більш інформативним для мовлення, оскільки відображає розподіл енергії сигналу за частотами. Саме у частотній області проявляються форманти, тобто характерні піки спектра, що визначають звучання голосних звуків.

На основі спектрального представлення виконується виділення ознак, яке є центральним етапом цього пройесу. Одним із найпоширеніших підходів є використання мел-частотних кепстральних коефіцієнтів. Цей метод включає перетворення частотної шкали до мел-шкали, яка відповідає особливостям сприйняття звуку людиною, застосування логарифмування енергії та подальше кепстральне перетворення. У результаті формується компактний вектор ознак, який описує форму спектра і є стійким до шумів та варіацій сигналу.

З математичної точки зору процес виділення ознак можна розглядати наступним чином (формула 2.12) :

$$\Phi = x[n] \rightarrow m_k, \quad (2.10)$$

де $m_k$ – вектор ознак.

Це відображення є нелінійним і спрямоване на зменшення розмірності даних при збереженні їх інформативності. У результаті замість великої кількості сирих відліків отримується компактне представлення, яке зручно використовувати для навчання моделей.

Додатково до основних ознак можуть використовуватися похідні

характеристики, такі як дельта- та дельта-дельта коефіцієнти, які відображають динаміку зміни сигналу у часі. Це дозволяє враховувати не лише статичні властивості мовлення, але й його часову структуру, що є важливим для розпізнавання.

Виділення ознак є одним із ключових етапів обробки мовного сигналу, оскільки саме на цьому рівні відбувається перехід від сирих аудіоданих до компактного, інформативного та структурованого представлення, придатного для ефективного аналізу алгоритмами машинного навчання. Основна мета цього процесу полягає не лише у зменшенні розмірності даних, але й у виділенні тих характеристик сигналу, які найбільш точно відображають особливості мовлення і дозволяють розрізняти різні звуки, слова та інтонації. Сирий сигнал, представлений у часовій області, містить велику кількість надлишкової інформації та є складним для безпосереднього аналізу, тому його перетворюють у форму, яка краще відповідає фізичній природі мовлення та особливостям його сприйняття.

Процес виділення ознак починається з аналізу коротких сегментів сигналу, які вважаються квазістаціонарними. Для кожного такого сегмента формується спектральне представлення, що дозволяє оцінити розподіл енергії сигналу за частотами. Оскільки саме частотна структура визначає фонетичні характеристики мовлення, подальша обробка зосереджена на узагальненні цієї інформації. Для цього використовується перетворення частотної шкали у нелінійну шкалу, яка відображає особливості слухового сприйняття людини, де низькі частоти мають більшу роздільну здатність, ніж високі. Після цього виконується логарифмування енергії сигналу, що дозволяє зменшити вплив різких змін амплітуди та наблизити представлення до перцептивних характеристик слуху.

Наступним кроком є отримання компактного набору коефіцієнтів, які описують форму спектра. При цьому відсікаються дрібні коливання та шумові компоненти, залишаючи лише узагальнену інформацію про сигнал. У результаті кожен фрагмент мовлення описується невеликим вектором ознак, який містить суттєві параметри, пов’язані з артикуляційними властивостями мовлення. Таке

представлення значно зменшує обсяг даних і водночас підвищує їх інформативність для задачі розпізнавання.

Важливою особливістю є те, що ознаки формуються таким чином, щоб бути стійкими до змін умов запису, таких як шум, відстань до мікрофона або індивідуальні відмінності мовців. Це досягається шляхом використання нормалізації та перетворень, які згладжують варіації, не пов’язані зі змістом мовлення. Додатково можуть враховуватися часові зміни сигналу шляхом обчислення похідних ознак, що дозволяє відобразити динаміку мовлення і покращити розпізнавання послідовностей звуків.

У підсумку, виділення ознак формує багатовимірний простір, у якому кожен мовний фрагмент представлений як точка з певними характеристиками. У цьому просторі схожі звуки розташовуються ближче один до одного, тоді як різні розташовуються далі, що значно спрощує задачу класифікації.

Це багатовимірне представлення можна інтерпретувати як геометричний простір ознак, у якому кожен мовний фрагмент кодується у вигляді вектора з фіксованою кількістю параметрів. Кожна координата такого вектора відповідає певній характеристиці сигналу, наприклад енергії в окремих частотних діапазонах або узагальненим спектральним властивостям. У результаті формується простір, у якому структура даних набуває більш впорядкованого вигляду порівняно з сирим сигналом.

У цьому просторі природним чином виникає кластеризація: фрагменти мовлення, що відповідають однаковим або подібним звукам, утворюють компактні групи. Це пояснюється тим, що однакові фонетичні одиниці мають подібні спектральні характеристики, а отже їхні вектори ознак розташовуються близько один до одного відповідно до обраної метрики відстані. Найчастіше використовується евклідова відстань або косинусна міра подібності, які дозволяють кількісно оцінити схожість між різними фрагментами мовлення. Водночас звуки, що суттєво відрізняються за артикуляцією або акустичними властивостями, формують віддалені кластери.

Таке геометричне трактування є надзвичайно важливим для алгоритмів машинного навчання, оскільки більшість із них фактично працюють із відстанями або границями між класами у цьому просторі. Наприклад, класифікаційні моделі намагаються побудувати роздільні поверхні, які відокремлюють різні групи точок. Чим краще сформовані ознаки, тим чіткіше розділені ці групи і тим простішою стає задача навчання моделі. У ідеальному випадку ознаки забезпечують таке представлення, при якому дані різних класів можна розділити навіть простими лінійними методами.

Крім того, важливою характеристикою такого простору є його стійкість до варіацій. Добре сформовані ознаки забезпечують те, що зміни, пов’язані з шумом або індивідуальними особливостями мовця, призводять лише до незначних зсувів точок у просторі, не порушуючи загальної структури кластерів. Це означає, що система здатна узагальнювати та правильно класифікувати нові, раніше не зустрічені дані.

Таким чином, багатовимірний простір ознак є не просто зручним представленням даних, а фундаментом для роботи всієї системи розпізнавання. Саме його геометричні властивості визначають складність задачі класифікації, швидкість навчання моделей і їхню здатність до узагальнення, що в кінцевому підсумку впливає на точність і надійність роботи кіберфізичної системи.

Саме тому якість сформованих ознак безпосередньо впливає на ефективність усієї системи розпізнавання голосу, визначаючи точність, швидкодію та стійкість до зовнішніх впливів.

Таким чином, математична модель сигналу та процес виділення ознак забезпечують перехід від фізичного акустичного процесу до формалізованого представлення даних. Це представлення є основою для роботи алгоритмів машинного навчання, оскільки дозволяє ефективно аналізувати мовлення, зменшуючи обчислювальні витрати та підвищуючи точність розпізнавання.

## 2.5 Висновки

За результатами проведених досліджень в другому розділі розроблено структурну модель КФС, яка представлена як багаторівнева ієрархія, що включає фізичний рівень (збір акустичних даних), мережевий (передача інформації), інтелектуальний (ML-моделі) та рівень виконавчих механізмів. Це забезпечує цілісне розуміння взаємодії між апаратними компонентами та алгоритмічною частиною.

Здійснено формалізацію процесу розпізнавання голосу через представлення його як впорядкованої системи множин і послідовної композиції функціональних відображень. Такий підхід дозволив математично описати кожен етап перетворення інформації, від аналогового сигналу до прийняття рішення та виконання дії.

Описано математичну модель мовного сигналу на основі моделі джерело-фільтр, де мовлення розглядається як згортка сигналу збудження з імпульсною характеристикою голосового тракту. Визначено параметри віконної обробки (вікно Хеммінга, фрейми 20–30 мс) та використання дискретного перетворення Фур’є для переходу у частотну область. Обґрунтовано метод виділення ознак із використанням мел-частотних кепстральних коефіцієнтів, що дозволяє сформувати компактний багатовимірний простір ознак. Це забезпечує перехід від надлишкових сирих даних до структурованих векторів, стійких до шумів та індивідуальних особливостей мовця, що критично важливо для ефективного навчання моделей машинного навчання.

# 3 МЕТОДИ ТА АЛГОРИТМИ МАШИННОГО НАВЧАННЯ ДЛЯ РОЗПІЗНАВАННЯ ГОЛОСУ

## 3.1 Класифікація підходів до розпізнавання голосу на рівні мовлення

Розпізнавання голосу на рівні мовлення є складною задачею, що поєднує обробку сигналів, лінгвістику та методи машинного навчання. У контексті кіберфізичних систем ця задача набуває додаткової складності через необхідність роботи в реальному часі, обмежені обчислювальні ресурси та вплив зовнішніх факторів, таких як шум або змінні умови середовища.

Саме тому вибір підходів до розпізнавання голосу на рівні мовлення та відповідних алгоритмів машинного навчання є критично важливим етапом проєктування системи.

Існуючі підходи до розпізнавання голосу на рівні мовлення можна класифікувати за кількома ознаками, серед яких найбільш важливими є принцип побудови моделі, тип використовуваних даних та рівень абстракції обробки інформації.

З точки зору історичного розвитку, виділяють класичні статистичні методи, нейромережеві підходи та сучасні гібридні або наскрізні (end-to-end) моделі. Кожен із цих підходів має свої переваги та обмеження, що визначають доцільність їх використання в конкретних умовах.

Класичні методи розпізнавання голосу на рівні мовлення базуються на статистичних моделях і припущенні про ймовірнісну природу мовного сигналу. У таких підходах задача розпізнавання формулюється як пошук найбільш ймовірної послідовності слів або фонем за заданим акустичним сигналом. Для цього використовуються моделі, що описують як акустичні властивості мовлення, так і мовні закономірності.

Основною перевагою цих методів є їхня інтерпретованість та відносно невисокі вимоги до обчислювальних ресурсів. Однак вони мають обмежену здатність моделювати складні нелінійні залежності, що характерні для реального мовлення.

З розвитком обчислювальних технологій значного поширення набули нейромережеві методи, які дозволяють автоматично навчатися складним закономірностям на основі великих обсягів даних [41]. На відміну від класичних підходів, нейронні мережі не потребують явного задання правил або моделей, а формують внутрішнє представлення даних у процесі навчання. Це дозволяє значно підвищити точність розпізнавання, особливо у складних умовах. Різні типи нейронних мереж, такі як згорткові або рекурентні, орієнтовані на обробку різних аспектів сигналу, зокрема просторових або часових залежностей.

Сучасні підходи до розпізнавання голосу на рівні мовлення все частіше використовують наскрізні моделі, які поєднують усі етапи обробки в єдину архітектуру. У таких системах відбувається пряме відображення вхідного аудіосигналу у текст без явного розділення на окремі етапи, такі як виділення ознак чи побудова мовної моделі [42]. Це спрощує архітектуру системи та дозволяє досягти високих показників точності, однак потребує значних обчислювальних ресурсів і великих навчальних вибірок.

Окрім класифікації за типом моделей, методи машинного навчання для розпізнавання мовлення можна поділити за способом навчання. Виділяють навчання з учителем, без учителя та напівкероване навчання. У більшості практичних систем використовується навчання з учителем, де модель навчається на розмічених даних, що містять відповідності між аудіосигналом і текстом. Водночас у сучасних дослідженнях активно розвиваються підходи, що дозволяють використовувати нерозмічені дані, що є важливим для масштабування систем.

Важливо також враховувати, що ефективність алгоритмів машинного навчання значною мірою залежить від якості вхідних ознак [43]. Як було показано в попередньому розділі, процес виділення ознак відіграє ключову роль у формуванні інформативного представлення сигналу. У класичних системах цей етап виконується окремо, тоді як у сучасних нейромережевих підходах він може бути інтегрований у саму модель.

З точки зору застосування в кіберфізичних системах, важливими критеріями

вибору методів є не лише точність, але й швидкодія, енергоефективність та здатність працювати в умовах обмежених ресурсів. Наприклад, для edge-пристроїв доцільно використовувати легкі моделі або гібридні підходи, тоді як у хмарних системах можна застосовувати більш складні архітектури.

Таблиця 3.1 – Порівняльна характеристика підходів до розпізнавання голосу на рівні мовлення

| Критерій | Класичні статистичні методи | Нейромережеві методи | Сучасні end-to-end підходи |
|---|---|---|---|
| Принцип роботи | Ймовірнісне моделювання | Навчання нелінійних залежностей | Наскрізне відображення сигнал в текст |
| Необхідність виділення ознак | Обов’язкова | Переважно потрібна | Часто інтегрована |
| Точність розпізнавання | Середня | Висока | Дуже висока |
| Стійкість до шумів | Обмежена | Висока | Дуже висока |
| Вимоги до ресурсів | Низькі | Середні–високі | Високі |
| Потреба у даних | Помірна | Велика | Дуже велика |
| Інтерпретованість | Висока | Середня | Низька |
| Застосування в КФС | Обмежене | Широке | Переважно хмарні системи |

Таким чином, класифікація підходів до розпізнавання голосу на рівні мовлення дозволяє систематизувати існуючі методи та визначити їхню придатність для конкретних умов використання. Подальший аналіз цих підходів дає можливість обґрунтовано обрати алгоритми, які забезпечать оптимальне

поєднання точності, швидкодії та ресурсної ефективності.

У результаті аналізу можна зробити висновок, що сучасні методи машинного навчання значно перевершують класичні підходи за точністю та адаптивністю, однак потребують більших ресурсів. Вибір конкретного підходу має здійснюватися з урахуванням особливостей кіберфізичної системи, включаючи обчислювальні можливості, вимоги до затримки та умови експлуатації.

Саме тому доцільним є використання комбінованих або оптимізованих моделей, які забезпечують баланс між ефективністю та ресурсними витратами.

## 3.2 Класичні статистичні методи розпізнавання голосу на рівні мовлення та їх модифікації

Класичні статистичні методи розпізнавання голосу на рівні мовлення є одним із перших і фундаментальних підходів, що заклали основу для сучасних систем автоматичного розпізнавання мовлення. Вони базуються на ймовірнісному моделюванні мовного сигналу та припущенні, що процес генерації мовлення можна описати як стохастичну послідовність прихованих станів і відповідних їм спостережень. Основною метою таких методів є знаходження найбільш імовірної послідовності мовних одиниць (фонем, складів або слів) за заданою послідовністю акустичних ознак.

### 3.2.1 Прихована марковська модель розпізнавання голосу на рівні мовлення

Однією з ключових моделей у цьому підході є прихована марковська модель (HMM). Вона описує мовний сигнал як марковський процес із прихованими станами, які не спостерігаються безпосередньо, але впливають на формування спостережуваних даних [51].

У контексті розпізнавання голосу на рівні мовлення такими станами можуть бути фонеми або їх частини, тоді як спостереженнями виступають вектори ознак,

отримані на попередньому етапі обробки сигналу (рисунок 3.1). Основна ідея НММ полягає в тому, що поточний стан залежить лише від попереднього, що значно спрощує математичний опис процесу.


Рисунок 3.1 – Класична система розпізнавання голосу на рівні мовлення

Формально НММ визначається набором параметрів, які включають ймовірності переходів між станами, розподіли ймовірностей спостережень у кожному стані та початкові ймовірності станів. Задача розпізнавання мовлення у цьому випадку зводиться до знаходження такої послідовності станів, яка максимізує ймовірність спостережуваної послідовності ознак. Для цього застосовуються спеціальні алгоритми, такі як алгоритм Вітербі, який дозволяє ефективно знайти найбільш імовірний шлях у моделі.

Важливим компонентом HMM є модель спостережень, яка описує, як саме формуються ознаки в кожному стані. Для цього традиційно використовуються гаусові суміші (GMM). GMM дозволяє апроксимувати складні розподіли ймовірностей за допомогою лінійної комбінації кількох гаусових компонент [52]. Це дає змогу гнучко моделювати варіативність мовного сигналу, враховуючи різні варіанти вимови, акценти та інші фактори.

Поєднання HMM і GMM утворює класичну архітектуру систем розпізнавання мовлення, яка протягом тривалого часу була стандартом у цій галузі. У такій системі HMM відповідає за моделювання часової структури мовлення, тоді як GMM відповідає за опис акустичних властивостей сигналу. Такий розподіл функцій дозволяє ефективно обробляти послідовні дані та враховувати як часові, так і спектральні характеристики мовлення.

Процес навчання моделей HMM і GMM здійснюється на основі статистичних методів, зокрема алгоритму максимальної правдоподібності. Для оцінювання параметрів використовується алгоритм Баум–Велша, який є спеціальним випадком алгоритму очікування-максимізації. Він дозволяє ітеративно уточнювати параметри моделі на основі навчальних даних. Важливою особливістю цього процесу є необхідність наявності великої кількості розмічених даних, що містять відповідності між аудіосигналами та мовними одиницями.

Незважаючи на свою ефективність, класичні методи мають ряд обмежень. Одним із основних є припущення про незалежність ознак, яке не завжди виконується для реального мовлення.

Крім того, HMM має обмежену здатність моделювати довгострокові залежності, оскільки враховує лише попередній стан. Це може призводити до втрати важливої контекстної інформації. GMM, у свою чергу, не здатні ефективно моделювати складні нелінійні залежності між ознаками.

Для подолання цих обмежень було запропоновано ряд модифікацій і розширень класичних моделей. Зокрема, використовуються контекстно-залежні моделі, які враховують вплив сусідніх фонем на поточний стан [53]. Також

застосовуються адаптивні методи, що дозволяють підлаштовувати модель під конкретного користувача або умови середовища.

Іншим напрямом розвитку є використання дискримінативних критеріїв навчання, які спрямовані на безпосереднє покращення точності класифікації, а не лише на максимізацію правдоподібності.

З розвитком технологій машинного навчання класичні методи поступово витісняються нейромережевими підходами, однак вони й досі залишаються актуальними в ряді застосувань. Зокрема, вони використовуються у системах з обмеженими обчислювальними ресурсами, а також як базові моделі або складові більш складних гібридних систем. Крім того, їхня відносна простота та інтерпретованість роблять їх зручними для аналізу та дослідження.

Таким чином, класичні статистичні методи, зокрема HMM і GMM, відіграли ключову роль у розвитку систем розпізнавання мовлення. Вони забезпечують ефективне моделювання часових і спектральних характеристик сигналу, хоча й мають певні обмеження. Розуміння принципів їхньої роботи є важливим для подальшого аналізу сучасних методів і вибору оптимальних рішень для кіберфізичних систем.

### 3.2.2 Нейромережевий підхід до розпізнавання голосу на рівні мовлення

Нейромережевий підхід до розпізнавання голосу на рівні мовлення, заснований на поєднанні глибоких нейронних мереж і прихованих марковських моделей, є еволюційним розвитком класичних статистичних методів [54].

Його основна ідея полягає у заміні традиційних гаусових сумішей на нейронну мережу, яка здатна значно точніше оцінювати ймовірності акустичних станів на основі вхідних ознак. Такий підхід дозволяє зберегти переваги HMM у моделюванні часової структури мовлення, одночасно підвищуючи якість акустичного моделювання.

На початковому етапі система отримує аудіосигнал із мікрофона, який є неперервним у часі та містить як корисну інформацію, так і різноманітні завади.

Для підготовки сигналу до подальшої обробки виконується попередня обробка, що включає фільтрацію шумів, нормалізацію амплітуди, компенсацію ехо та виділення мовної активності. Цей етап є критично важливим, оскільки він безпосередньо впливає на якість ознак і, відповідно, на точність розпізнавання (рисунок 3.2).


Рисунок 3.2 – Структурна схема нейромережевого підходу розпізнавання голосу на рівні мовлення

Після цього сигнал переходить до етапу виділення ознак, де він перетворюється у компактне числове представлення. Найчастіше використовуються спектральні або кепстральні коефіцієнти, які відображають частотну структуру мовлення. У результаті формується послідовність векторів ознак, кожен із яких описує короткий фрагмент сигналу. Саме ці вектори є вхідними даними для нейронної мережі.

Далі виконується передача даних між рівнями кіберфізичної системи. У більшості випадків обчислювально складні операції, пов’язані з роботою нейронної мережі, виконуються на віддалених обчислювальних ресурсах, таких як fog або cloud-сервери. Це дозволяє зменшити навантаження на edge-пристрої та забезпечити використання більш потужних моделей.

На рівні обчислень ключову роль відіграє глибока нейронна мережа. Вона приймає на вхід вектори ознак і виконує їх нелінійне перетворення через кілька шарів. У результаті мережа формує оцінки ймовірностей того, що кожен фрагмент сигналу відповідає певному акустичному стану. На відміну від GMM, нейронна мережа здатна моделювати складні залежності між ознаками, що дозволяє значно підвищити точність системи.

Отримані ймовірності передаються до HMM-декодера, який відповідає за визначення найбільш імовірної послідовності станів. Для цього використовується алгоритм Вітербі, який знаходить оптимальний шлях у просторі станів з урахуванням як акустичних ймовірностей, так і ймовірностей переходів між станами [57]. Таким чином, HMM забезпечує врахування часової структури мовлення, що є важливим для коректного розпізнавання.

Результатом роботи декодера є текстове представлення мовлення, яке може містити окремі слова або повні фрази. Далі цей результат передається до виконавчого рівня кіберфізичної системи, де на його основі формуються відповідні дії або команди. Це може бути керування пристроями, відображення інформації або взаємодія з користувачем.

Таким чином, нейромережевий підхід DNN-HMM забезпечує ефективне поєднання точності та структурованості [58, 59]. Нейронна мережа відповідає за високоточне акустичне моделювання, тоді як HMM відповідає за часову організацію процесу розпізнавання. Завдяки цьому досягається баланс між складністю моделі, якістю результату та можливістю практичної реалізації в кіберфізичних системах. У цьому випадку нейронна мережа використовується для оцінювання ймовірностей станів, що значно підвищує точність системи. При цьому HMM залишається відповідальним за часову динаміку.

### 3.2.3 Рекурентні нейронні мережі у задачах розпізнавання голосу на рівні мовлення

Рекурентні нейронні мережі (RNN) є важливим класом моделей машинного навчання, спеціально призначених для обробки послідовних даних, до яких належить і мовний сигнал. На відміну від традиційних нейронних мереж прямого поширення, рекурентні моделі мають внутрішню пам’ять, що дозволяє враховувати попередні стани під час обробки поточного елемента послідовності [60, 61]. Це є критично важливим для задач розпізнавання мовлення, оскільки звукові сигнали мають виражену часову структуру, а значення окремих елементів часто залежить від контексту.

Базова рекурентна нейронна мережа будується таким чином, що на кожному кроці часу вона приймає на вхід не лише поточний вектор ознак, але й прихований стан із попереднього кроку.

У результаті формується новий прихований стан, який містить інформацію як про поточний сигнал, так і про попередній контекст. Така архітектура дозволяє моделі накопичувати інформацію про послідовність і використовувати її для прийняття рішень. У задачі розпізнавання мовлення це означає, що система може враховувати залежності між звуками, що значно покращує точність порівняно з моделями, які аналізують кожен фрагмент незалежно.

Однак класичні RNN мають суттєве обмеження, пов’язане з проблемою зникнення або вибуху градієнтів під час навчання. Це ускладнює запам’ятовування довгострокових залежностей, які часто присутні у мовленні.

Для подолання цієї проблеми були розроблені більш складні архітектури, зокрема мережі з довгою короткочасною пам’яттю (LSTM). Вони мають спеціальні механізми керування інформаційними потоками, які дозволяють зберігати важливу інформацію протягом тривалого часу та відкидати несуттєву [62, 63].

LSTM-мережі містять так звані гейти (вхідний, вихідний та гейт забування), які регулюють, яка інформація повинна бути збережена, оновлена або видалена.

Завдяки цьому модель може ефективно працювати з довгими послідовностями та враховувати контекст на значних часових інтервалах.

У задачах розпізнавання мовлення це дозволяє коректно інтерпретувати звуки, значення яких залежить від попередніх або наступних елементів, наприклад у випадку коартикуляції або швидкої мови.

У практичних системах розпізнавання голосу на рівні мовлення часто використовуються двонаправлені рекурентні мережі, які аналізують сигнал як у прямому, так і у зворотному часовому напрямку. Це дозволяє враховувати не лише попередній, але й майбутній контекст, що особливо корисно для підвищення точності розпізнавання. У такому випадку модель має доступ до повної інформації про послідовність і може більш точно визначати межі та значення мовних одиниць.


Рисунок 3.3 – Структурна схема підходу з використанням рекурентних нейронних мереж для розпізнавання голосу на рівні мовлення

Ще одним важливим аспектом є інтеграція RNN/LSTM у загальну архітектуру системи. Вони можуть використовуватися як окремі моделі для безпосереднього розпізнавання мовлення або як складова частина більш складних систем, наприклад у поєднанні з іншими нейронними мережами чи декодерами. У деяких підходах рекурентні мережі виконують роль акустичної моделі, замінюючи традиційні компоненти, а в інших працюють у складі наскрізних систем.

З точки зору кіберфізичних систем, рекурентні моделі мають як переваги, так і обмеження. До переваг належить висока точність і здатність враховувати часовий контекст, що є критично важливим для обробки мовлення.

Водночас їх використання пов’язане з підвищеними обчислювальними витратами та складністю реалізації, особливо на edge-пристроях. Це обумовлює необхідність оптимізації моделей або їх часткового винесення на більш потужні обчислювальні ресурси.

Таким чином, рекурентні нейронні мережі, зокрема LSTM, є потужним інструментом для розпізнавання мовлення, який дозволяє ефективно моделювати часові залежності сигналу. Вони значно перевершують класичні підходи за здатністю враховувати контекст і забезпечують високу якість розпізнавання, що робить їх важливим компонентом сучасних кіберфізичних систем.

### 3.2.4 Методи розпізнавання голосу на рівні мовлення на основі end-to-end архітектур

Сучасні підходи до розпізнавання голосу на рівні мовлення все частіше базуються на концепції end-to-end, яка передбачає побудову єдиної моделі, що безпосередньо відображає вхідний аудіосигнал у текстовий результат без явного розділення на окремі етапи, характерні для класичних систем [65, 66].

На відміну від традиційних підходів, де обробка сигналу, виділення ознак, акустичне моделювання та декодування реалізуються як окремі модулі, у end-to-end архітектурі всі ці функції інтегруються в одну нейронну мережу, яка

навчається виконувати повний цикл перетворення (рисунок 3.4).

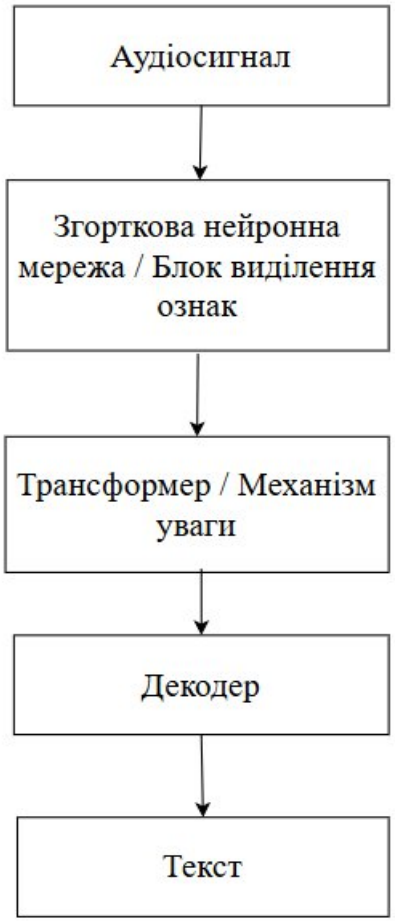



Рисунок 3.4 – Структурна схема підходу на основі end-to-end архітектур для розпізнавання голосу

На початковому етапі система отримує сирий аудіосигнал або його базове спектральне представлення. У деяких реалізаціях модель може працювати безпосередньо з часовими відліками сигналу, однак частіше використовується попередньо обчислений спектр або інші часово-частотні характеристики.

Ці дані подаються на вхід нейронної мережі, яка самостійно виконує функцію виділення ознак. Для цього використовуються згорткові шари або інші механізми, що дозволяють автоматично виявляти локальні закономірності у сигналі, такі як характерні частотні патерни.

Далі обробка переходить до більш глибоких шарів моделі, які відповідають за формування узагальненого представлення мовлення. У сучасних архітектурах

часто використовуються механізми уваги або трансформери, які дозволяють моделі враховувати залежності між різними частинами сигналу незалежно від їх відстані у часі. Це є суттєвою перевагою порівняно з рекурентними мережами, оскільки дозволяє ефективно працювати з довгими послідовностями та складними контекстами.

Ключовим елементом end-to-end моделей є декодер, який перетворює внутрішнє представлення сигналу у послідовність символів або слів [67]. На цьому етапі можуть використовуватися різні підходи, зокрема механізми послідовного генерування або спеціальні функції втрат, які дозволяють узгодити довжину вхідної та вихідної послідовностей.

У результаті модель формує текстове представлення мовлення без необхідності використання окремого HMM-декодера чи мовної моделі, хоча в деяких випадках додаткові мовні моделі можуть інтегруватися для покращення результату.

Однією з головних переваг end-to-end підходу є його здатність до автоматичного навчання всіх етапів обробки [68]. Модель оптимізується як єдине ціле, що дозволяє уникнути накопичення помилок між окремими модулями, характерного для класичних систем. Крім того, відсутність необхідності ручного проєктування ознак спрощує процес розробки та дозволяє адаптувати систему до нових умов або мов.

Водночас такий підхід має і певні обмеження. Основним із них є висока вимогливість до обчислювальних ресурсів і обсягу навчальних даних [69]. Для ефективного навчання end-to-end моделей необхідні великі корпуси розміченого мовлення, а сама модель потребує значних обчислювальних потужностей. Це може ускладнювати їх використання у кіберфізичних системах, особливо на edge-рівні, де ресурси обмежені.

З точки зору архітектури кіберфізичної системи, end-to-end моделі зазвичай реалізуються на рівні хмарних або туманних обчислень, де доступні необхідні ресурси. Edge-пристрої при цьому виконують лише базову обробку сигналу та передачу даних. Такий розподіл дозволяє використовувати переваги сучасних

моделей, не перевантажуючи локальні пристрої [70].

Таким чином, end-to-end підхід є сучасним напрямом розвитку систем розпізнавання мовлення, який забезпечує високу точність і гнучкість за рахунок інтеграції всіх етапів обробки в єдину модель. Незважаючи на високі вимоги до ресурсів, він відкриває нові можливості для створення інтелектуальних систем і є перспективним для подальшого розвитку кіберфізичних технологій.

## 3.3 Метод розпізнавання голосу на основі згорткових і рекурентних нейронних мереж

У процесі проєктування кіберфізичної системи розпізнавання голосу важливим етапом є вибір алгоритмічного підходу, який забезпечує необхідний баланс між точністю, швидкодією та складністю реалізації.

З урахуванням обмежень, характерних для кіберфізичних систем, таких як обмежені обчислювальні ресурси на периферійних пристроях, вимоги до роботи в реальному часі та необхідність стійкості до шумів, доцільним є використання нейромережевого підходу на основі комбінації згорткових і рекурентних нейронних мереж (CNN – LSTM).

Даний підхід дозволяє ефективно обробляти як спектральні, так і часові характеристики мовного сигналу, забезпечуючи при цьому відносну простоту реалізації порівняно з більш складними гібридними або трансформерними моделями.

Запропонований метод базується на ідеї послідовного перетворення аудіосигналу у текстове представлення шляхом використання єдиної нейромережевої архітектури, яка виконує функції як виділення ознак, так і їх інтерпретації (формула 3.1):

$$Y = f_{dec}(f_{LSTM}(f_{CNN}(f_{spec}(X)))), \tag{3.1}$$

де $f_{dec}$ – функція, що переводить сигнал у спектральну форму;

$f_{LSTM}$ – функція, яка відповідає за розпізнавання часового контексту мовлення;

$f_{CNN}$ – функція, яка відповідає за формування акустичних ознак;

$f_{spec}$ – функція, яка перетворює результат у текст.

На відміну від класичних підходів, де ці етапи реалізуються окремо, у даному випадку значна частина обробки інтегрується в саму модель, що спрощує загальну архітектуру системи (рисунок 3.5).

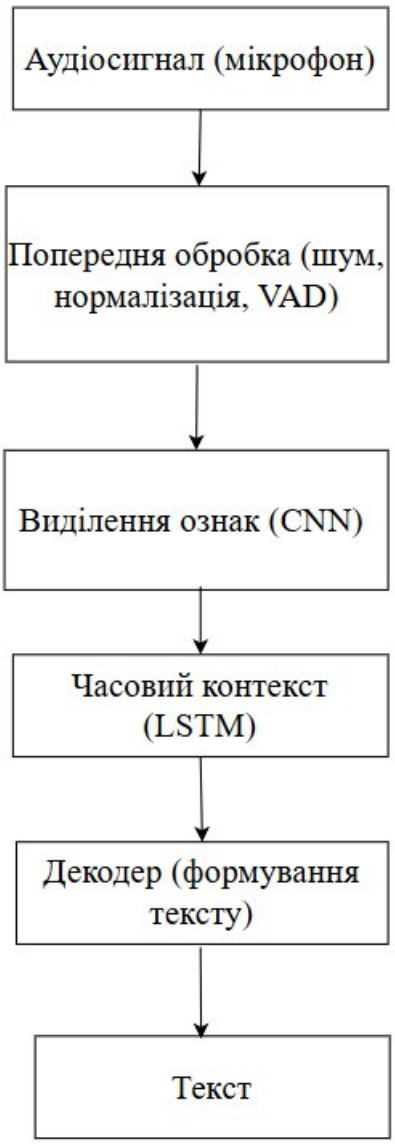



Рисунок 3.5 – Структурна схема запропонованого методу на основі згорткових і рекурентних нейронних мереж

На початковому етапі система отримує аудіосигнал із мікрофона, який може містити як корисну мовну інформацію, так і різноманітні шуми. З метою підвищення якості даних виконується попередня обробка сигналу, що включає фільтрацію шумів, нормалізацію амплітуди та виділення мовної активності. Цей етап є необхідним для забезпечення стабільності роботи моделі, особливо в умовах реального середовища.

Після попередньої обробки сигнал перетворюється у спектральне представлення, наприклад у вигляді спектрограми або мел-спектрограми. Таке представлення дозволяє відобразити зміну частотного складу сигналу в часі та є зручним для подальшої обробки нейронними мережами. Саме ці дані подаються на вхід згорткової нейронної мережі.

Згорткова нейронна мережа виконує роль автоматичного виділення ознак. Вона аналізує локальні структури спектрограми, виявляючи характерні патерни, що відповідають певним звукам або їх поєднанням. Завдяки використанню згорткових фільтрів модель здатна враховувати просторові залежності між частотами, а також зменшувати вплив шумів і незначних варіацій сигналу. У результаті формується узагальнене представлення даних, яке містить найбільш інформативні характеристики мовлення.

Далі отримані ознаки передаються до рекурентної нейронної мережі типу LSTM. Основною функцією цього блоку є врахування часових залежностей у мовленні. Оскільки мовний сигнал є послідовністю, де кожен елемент залежить від попередніх, використання LSTM дозволяє моделі запам’ятовувати контекст і коректно інтерпретувати звуки у межах фрази. Завдяки механізмам керування пам’яттю LSTM здатна ефективно працювати з довгими послідовностями та уникати проблем, характерних для звичайних рекурентних мереж.

На наступному етапі формується вихід моделі у вигляді ймовірностей належності до певних символів або класів.

Для перетворення цих ймовірностей у текст використовується декодер, який може базуватися на простих алгоритмах вибору найбільш імовірної послідовності. У спрощених реалізаціях це може бути покрокове визначення

символів, тоді як у більш складних відбувається застосування спеціалізованих алгоритмів узгодження послідовностей.

Отриманий текстовий результат передається до виконавчого рівня кіберфізичної системи, де використовується для формування команд або взаємодії з користувачем. Це може бути керування пристроями, виконання голосових команд або відображення інформації.

Однією з ключових переваг запропонованого підходу є його відносна простота та універсальність.

На відміну від складних гібридних моделей, він не потребує використання окремих статистичних компонентів, таких як HMM, що значно спрощує реалізацію. При цьому модель забезпечує достатньо високу точність розпізнавання завдяки використанню сучасних методів глибокого навчання.

З точки зору кіберфізичної системи, важливою є можливість розподілу обчислень між різними рівнями. Попередня обробка сигналу та формування спектрограми можуть виконуватися на edge-пристрої, що дозволяє зменшити обсяг передаваних даних. Основна нейромережева обробка може здійснюватися на більш потужних обчислювальних вузлах, що забезпечує необхідну продуктивність без перевантаження локальних ресурсів.

## 3.4 Функція втрат у задачах розпізнавання голосу

Функція втрат є ключовим елементом будь-якої моделі машинного навчання, оскільки саме вона визначає, наскільки точно модель виконує поставлену задачу та у якому напрямку необхідно коригувати її параметри під час навчання. У задачах розпізнавання голосу, залежно від постановки проблеми, функція втрат може мати різну форму, оскільки сама задача може включати як розпізнавання мовлення (перетворення аудіо в текст), так і ідентифікацію або верифікацію мовця.

У задачах розпізнавання голосу функція втрат відіграє визначальну роль, оскільки саме вона формалізує критерій якості роботи моделі та задає напрямок її

навчання. На відміну від розпізнавання мовлення, де основною метою є перетворення аудіосигналу у текст, у задачах розпізнавання голосу ключовим є визначення особи за її голосовими характеристиками. Це означає, що модель повинна навчитися виділяти унікальні ознаки голосу, які дозволяють відрізнити одного мовця від іншого, незалежно від змісту сказаного.

Особливістю таких задач є те, що голос людини характеризується складною сукупністю параметрів, включаючи тембр, висоту, інтонаційні особливості, артикуляцію та інші акустичні характеристики. Ці параметри можуть змінюватися залежно від емоційного стану, умов запису або якості мікрофона, що ускладнює задачу розпізнавання. Саме тому функція втрат повинна бути побудована таким чином, щоб забезпечити стійкість моделі до таких варіацій та одночасно підсилювати відмінності між різними мовцями.

У загальному випадку функція втрат визначає відхилення між прогнозованим результатом моделі та еталонним значенням.

Формально це можна записати у вигляді формули 3.2:

$$R = R\left(Y, \acute{Y}\right), \tag{3.2}$$

де $Y$ – істинна відповідь;

$\acute{Y}$ - результат, отриманий моделлю.

У загальному вигляді функція втрат визначає відстань або невідповідність між еталонним представленням голосу та тим, яке формує модель. На практиці це означає, що модель навчається відображати аудіосигнал у деякий вектор ознак, у якому схожі голоси розташовані близько один до одного, а різні розташовані на більшій відстані. Таким чином, задача розпізнавання голосу перетворюється на задачу навчання простору ознак із певною геометричною структурою.

Одним із базових підходів є використання функції крос-ентропії, коли задача формулюється як багатокласова класифікація. У цьому випадку кожен мовец розглядається як окремий клас, а модель навчається передбачати ймовірність належності голосу до певної особи. Такий підхід є відносно простим

у реалізації та добре працює, коли кількість мовців у системі є фіксованою. Однак його недоліком є обмежена здатність до узагальнення на нових користувачів, які не були присутні у навчальній вибірці.

Більш гнучким підходом є використання метричних функцій втрат, які безпосередньо оптимізують відстані між ознаками голосів. У цьому випадку модель не просто класифікує сигнал, а формує компактне представлення голосу у вигляді вектора фіксованої розмірності. Далі порівняння голосів здійснюється за допомогою певної метрики, наприклад евклідової відстані або косинусної подібності.

Однією з найбільш поширених є триплетна функція втрат, яка базується на одночасному розгляді трьох прикладів: опорного, позитивного та негативного. Опорний і позитивний приклади відповідають одному мовцю, тоді як негативний відповідає іншому. Мета навчання полягає у тому, щоб відстань між опорним і позитивним прикладами була меншою, ніж відстань між опорним і негативним, з певним запасом. Це дозволяє формувати добре структурований простір ознак, у якому голоси одного користувача утворюють компактні кластери.

Іншим підходом є контрастивна функція втрат, яка працює з парами прикладів і навчає модель зближувати ознаки схожих голосів та віддаляти різні. Такий підхід є ефективним у задачах верифікації, де потрібно відповісти на питання, чи належить голос певній особі.

У сучасних системах також використовуються вдосконалені функції втрат, наприклад варіанти з додатковими кутовими або маржинальними обмеженнями. Вони дозволяють покращити роздільність класів у просторі ознак і підвищити точність розпізнавання, особливо у складних умовах.

З точки зору кіберфізичних систем, важливим є те, що вибір функції втрат впливає не лише на точність, але й на обчислювальну складність моделі. Наприклад, метричні підходи можуть вимагати більш складної організації навчального процесу, але забезпечують кращу масштабованість та гнучкість при додаванні нових користувачів. Це є важливим фактором для систем, які повинні адаптуватися до змін у середовищі експлуатації.

Таким чином, функція втрат у задачах розпізнавання голосу визначає структуру простору ознак, у якому працює модель, та безпосередньо впливає на її здатність розрізняти мовців. Вибір конкретного типу функції втрат залежить від постановки задачі, вимог до системи та умов її використання. У сучасних підходах перевага надається метричним функціям втрат, які забезпечують кращу узагальнювальну здатність та високу точність ідентифікації голосу.

У задачах розпізнавання голосу вибір функції втрат безпосередньо визначає якість формування простору ознак, у якому відбувається ідентифікація або верифікація мовця. Саме тому аналіз її переваг та обмежень є важливим для обґрунтування запропонованого підходу.

Однією з ключових переваг сучасних функцій втрат, особливо метричних, є їх здатність формувати інформативний та добре структурований простір ознак. У такому просторі голосові характеристики одного мовця групуються у компактні кластери, тоді як ознаки різних мовців розташовуються на значній відстані. Це забезпечує високу точність розпізнавання навіть у випадках, коли вхідні дані містять варіації, пов’язані зі зміною інтонації, темпу мовлення або умов запису. Така властивість є особливо важливою для реальних кіберфізичних систем, де середовище є неконтрольованим.

Ще однією важливою перевагою є узагальнювальна здатність моделей, навчених із використанням метричних функцій втрат. На відміну від класичних підходів на основі класифікації, де система обмежена набором мовців із навчальної вибірки, метричні методи дозволяють працювати з новими користувачами без необхідності повного перенавчання моделі. Достатньо сформувати еталонний вектор ознак нового мовця, після чого система може виконувати його ідентифікацію або верифікацію шляхом порівняння з іншими векторами. Це робить такі підходи більш гнучкими та масштабованими.

Крім того, функції втрат у задачах розпізнавання голосу сприяють підвищенню стійкості системи до шумів та спотворень сигналу. Оскільки модель навчається виділяти саме індивідуальні характеристики голосу, вона частково ігнорує нерелевантні фактори, такі як фоновий шум або технічні особливості

запису. Це особливо важливо для кіберфізичних систем, які працюють у реальних умовах, де рівень завад може бути значним.

Водночас використання таких функцій втрат має і певні обмеження. Одним із основних є складність процесу навчання. Метричні підходи, зокрема триплетна функція втрат, вимагають спеціальної організації навчальних даних, зокрема формування пар або триплетів прикладів. Це ускладнює процес підготовки даних і може збільшувати час навчання моделі. Крім того, ефективність таких методів значною мірою залежить від якості вибору цих прикладів.

Ще одним обмеженням є підвищені вимоги до обчислювальних ресурсів. Формування простору ознак і обчислення відстаней між векторами може бути обчислювально затратним, особливо при великій кількості користувачів. У контексті кіберфізичних систем це може створювати додаткове навантаження на обчислювальні вузли, особливо якщо обробка виконується в реальному часі.

Також слід враховувати, що точність розпізнавання голосу може знижуватися в умовах значних змін голосу однієї і тієї ж людини, наприклад через хворобу, втому або зміну емоційного стану. Хоча сучасні функції втрат частково компенсують ці фактори, вони не завжди можуть повністю усунути їх вплив.

У запропонованій кіберфізичній системі функція втрат використовується на етапі навчання нейромережевої моделі, яка формує векторне представлення голосу користувача. Саме на цьому етапі визначається структура простору ознак, у якому в подальшому буде здійснюватися розпізнавання. Після завершення навчання модель переходить у режим експлуатації, де функція втрат безпосередньо не використовується, але її вплив зберігається у вигляді сформованих параметрів моделі.

У процесі роботи системи вхідний аудіосигнал перетворюється у вектор ознак, який порівнюється з еталонними векторами, що зберігаються в базі даних. Таким чином, функція втрат опосередковано визначає якість цього порівняння, оскільки саме вона забезпечує компактність і роздільність відповідних кластерів у просторі ознак.

З точки зору архітектури кіберфізичної системи, етап навчання моделі, де

використовується функція втрат, доцільно реалізовувати на рівні хмарних або туманних обчислень, де доступні необхідні ресурси. Водночас етапи формування ознак та їх порівняння можуть виконуватися на периферійних пристроях (edge-рівні), що дозволяє зменшити затримки та підвищити швидкодію системи.

Таким чином, функція втрат є невід'ємною складовою системи розпізнавання голосу, яка визначає ефективність навчання моделі та якість її подальшої роботи. Вона забезпечує формування інформативного простору ознак, у якому можливе точне та стійке розрізнення мовців, що є критично важливим для функціонування кіберфізичних систем.

## 3.5 Висновки

У межах третього розділу проведено класифікацію та порівняльний аналіз підходів до розпізнавання голосу, що дозволило встановити суттєву перевагу нейромережевих методів над класичними статистичними моделями HMM-GMM за показниками точності та стійкості до зовнішніх шумів.

Досліджено рекурентні архітектури, зокрема моделі LSTM, які завдяки механізмам керування пам'яттю забезпечують ефективне моделювання часового контексту та стабільне розпізнавання послідовностей на тривалих інтервалах.

На основі проведеного аналізу обґрунтовано вибір гібридного методу розпізнавання CNN–LSTM у поєднанні з функцією втрат CTC, що дозволяє інтегрувати процеси автоматичного виділення ознак і часового аналізу в єдину наскрізну систему без необхідності жорсткого вирівнювання даних. Визначено, що для оптимізації функціонування кіберфізичної системи розпізнавання голосу людини доцільним є розподіл обчислювального навантаження між периферійними пристроями для попередньої обробки сигналу та потужними хмарними вузлами для виконання глибокого нейромережевого аналізу.

# 4 ПРОЕКТУВАННЯ ТА ДОСЛІДЖЕННЯ КІБЕРФІЗИЧНОЇ СИСТЕМИ РОЗПІЗНАВАННЯ ГОЛОСУ

## 4.1 Архітектура програмної реалізації системи

Архітектура програмної реалізації кіберфізичної системи розпізнавання голосу визначає принципи організації її складових, взаємодію між ними, а також розподіл функціональних задач між різними рівнями системи.

Основною метою проєктування архітектури є забезпечення ефективної, масштабованої та стійкої роботи системи в умовах реального середовища, з урахуванням обмежень обчислювальних ресурсів і вимог до швидкодії.

Запропонована кіберфізична система розпізнавання голосу реалізується за багаторівневим принципом, що включає чотири основні рівні: фізичний рівень, який виконує функцію збору первинних акустичних сигналів та їхню попередню цифрову обробку за допомогою крайових обчислювальних пристроїв; мережевий рівень, що забезпечує надійну та ефективну передачу даних між компонентами системи за допомогою провідних і безпровідних технологій та протоколів зв’язку; інтелектуальний (або кібернетичний) рівень, який є центральним елементом для аналізу аудіосигналу та його перетворення у текстову або семантичну інформацію на основі моделей машинного навчання; а також рівень виконавчих механізмів, призначений для практичної реалізації дій у відповідь на розпізнані команди та забезпечення зворотного зв’язку з користувачем.

Такий підхід дозволяє оптимально розподілити обчислювальне навантаження та забезпечити гнучкість системи.

На фізичному рівні відбувається безпосередній збір аудіоданих за допомогою мікрофонного пристрою. У ролі edge-вузла може виступати вбудована система або одноплатний комп’ютер, який виконує первинну обробку сигналу. До функцій цього рівня належать дискретизація сигналу, базова фільтрація шумів, нормалізація та виділення сегментів мовлення. Реалізація цих операцій на периферійному рівні дозволяє зменшити обсяг передаваних даних та підвищити ефективність системи.

Після первинної обробки дані передаються на рівень обробки, де реалізовано основні алгоритми розпізнавання голосу. На цьому рівні виконується перетворення сигналу у спектральне представлення, формування ознак та їх подальша обробка нейромережевою моделлю типу CNN–LSTM. Саме тут відбувається формування векторного представлення голосу користувача, яке використовується для подальшої ідентифікації. Обчислення можуть виконуватися як на локальному сервері, так і в хмарному середовищі, залежно від вимог до продуктивності та доступних ресурсів.

Прикладний рівень відповідає за взаємодію системи з користувачем або іншими компонентами кіберфізичного середовища. На цьому етапі результати розпізнавання інтерпретуються та використовуються для прийняття рішень. Це може бути, наприклад, підтвердження особи користувача, надання доступу до системи або виконання певних команд. Таким чином, прикладний рівень реалізує логіку роботи системи відповідно до поставлених задач.

Важливою особливістю архітектури є організація потоків даних між рівнями. Аудіосигнал, отриманий на фізичному рівні, проходить послідовну обробку, перетворюючись із сирого сигналу у компактне представлення у вигляді векторів ознак. Ці вектори передаються між модулями системи, що дозволяє мінімізувати затримки та обсяг передаваних даних. У зворотному напрямку передаються результати розпізнавання, які використовуються для керування системою.

З програмної точки зору система реалізується у вигляді набору взаємодіючих модулів. Основними модулями є: модуль збору та попередньої обробки аудіо, модуль формування ознак, модуль нейромережевої обробки та модуль прийняття рішень. Кожен із цих модулів виконує чітко визначену функцію, що забезпечує модульність системи та спрощує її масштабування і модернізацію.

Окрему увагу приділено питанням розподілу обчислень між edge- та обчислювальним рівнями. Виконання частини обробки на периферійних пристроях дозволяє зменшити навантаження на центральні вузли та знизити

затримки, що є критично важливим для систем реального часу. Водночас складніші обчислення, пов’язані з роботою нейронної мережі, доцільно виконувати на більш потужних обчислювальних ресурсах.

Архітектура кіберфізичної системи розпізнавання голосу представлена на рисунку 4.1.

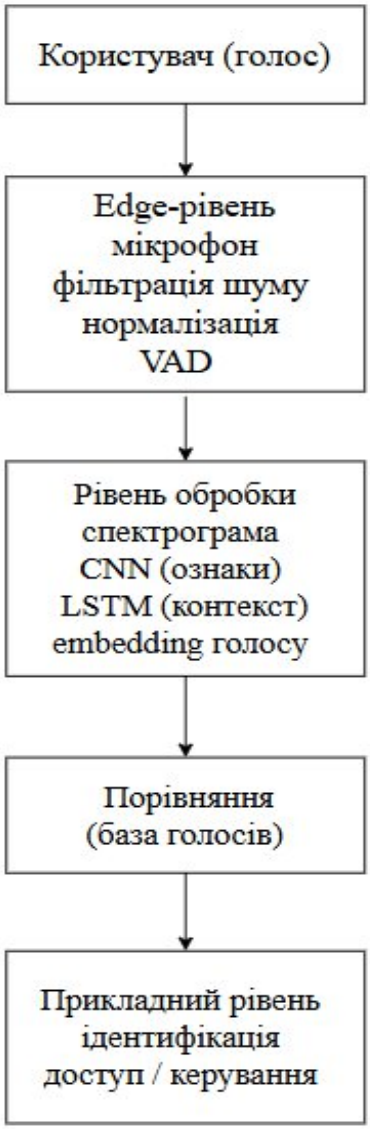



Рисунок 4.1 – Архітектура кіберфізичної системи розпізнавання голосу

Таким чином, запропонована архітектура забезпечує ефективну організацію процесу розпізнавання голосу, поєднуючи розподілену обробку даних, модульність та адаптивність. Вона дозволяє реалізувати систему, яка є достатньо гнучкою для використання у різних умовах та може бути масштабована залежно

від вимог конкретного застосування.

## 4.2 Вибір програмних засобів та технологій реалізації системи

Розробка кіберфізичної системи розпізнавання голосу потребує обґрунтованого вибору програмних засобів і технологій, які забезпечують ефективну реалізацію алгоритмів обробки сигналів, навчання нейронних мереж та інтеграцію компонентів системи.

При виборі інструментів враховуються такі критерії, як продуктивність, гнучкість, доступність бібліотек для обробки аудіо, підтримка машинного навчання, а також можливість роботи на різних рівнях архітектури, від edge-пристроїв до хмарних обчислювальних середовищ.

Основною мовою програмування для реалізації системи обрано Python. Такий вибір обумовлений широким використанням цієї мови у сфері обробки даних і машинного навчання, наявністю великої кількості спеціалізованих бібліотек, а також простотою інтеграції різних компонентів системи. Python забезпечує швидку розробку прототипів і дозволяє ефективно реалізовувати складні алгоритми без значних витрат часу на низькорівневу оптимізацію.

Для реалізації алгоритмів машинного навчання та побудови нейронної моделі використовується одна з сучасних бібліотек глибокого навчання, таких як TensorFlow або PyTorch.

Обидві платформи забезпечують підтримку побудови складних нейронних архітектур, включаючи згорткові та рекурентні мережі, а також мають інструменти для оптимізації навчання та роботи з великими обсягами даних. Використання таких бібліотек дозволяє реалізувати модель типу CNN – LSTM та ефективно виконувати її навчання і тестування.

Для обробки аудіосигналів застосовуються спеціалізовані бібліотеки, зокрема librosa та sciру. Вони надають інструменти для виконання таких операцій, як завантаження аудіофайлів, обчислення спектрограм, фільтрація сигналів та виділення ознак. Зокрема, бібліотека librosa широко використовується для

формування мел-спектрограм, які є стандартним представленням аудіоданих у задачах розпізнавання голосу.

Для роботи з числовими даними використовується бібліотека NumPy, яка забезпечує ефективні операції над багатовимірними масивами. Це є важливим для реалізації алгоритмів обробки сигналів і підготовки даних для нейронної мережі. Додатково можуть використовуватися бібліотеки pandas для організації даних та matplotlib для візуалізації результатів, зокрема побудови графіків навчання моделі.

З точки зору реалізації нейромережевої моделі, важливим є використання графічних процесорів (GPU), які дозволяють значно прискорити процес навчання. Сучасні бібліотеки глибокого навчання підтримують роботу з GPU, що забезпечує ефективне виконання обчислень, особливо при роботі з великими наборами даних.

У контексті кіберфізичної системи важливим є також розподіл програмного забезпечення між різними рівнями. На edge-рівні реалізуються легкі модулі збору та попередньої обробки аудіосигналу. Для цього можуть використовуватися бібліотеки, що забезпечують роботу з аудіопристроями у реальному часі.

Основна обробка, пов’язана з нейронною мережею, виконується на більш потужних обчислювальних вузлах, де доступні необхідні ресурси для навчання та виконання моделі. Для організації взаємодії між компонентами системи можуть використовуватися мережеві протоколи та інтерфейси прикладного програмування (API), які забезпечують передачу даних між рівнями системи. Це дозволяє реалізувати гнучку та масштабовану архітектуру, у якій окремі компоненти можуть бути розгорнуті на різних пристроях.

Середовище розробки також відіграє важливу роль у процесі створення системи. Використання інтегрованих середовищ розробки, таких як Visual Studio Code або PyCharm, забезпечує зручність написання коду, налагодження та тестування програмних модулів. Це сприяє підвищенню продуктивності розробки та якості програмного забезпечення.

Програмні засоби та бібліотеки, використані для реалізації системи

представлені в таблиці 4.1.

Таблиця 4.1 – Програмні засоби та бібліотеки, використані для реалізації системи

| Компонент системи | Бібліотека / інструмент | Призначення | Обґрунтування вибору |
|---|---|---|---|
| Мова програмування | Python | Основна мова розробки системи | Простота використання, велика кількість ML-бібліотек, активна спільнота |
| Обробка аудіо | librosa | Завантаження аудіо, спектрограми, MFCC | Спеціалізована бібліотека для аудіоаналізу |
| Числові обчислення | NumPy | Робота з масивами та матрицями | Висока швидкість обчислень |
| Обробка сигналів | SciPy | Фільтрація, перетворення сигналів | Розширені методи цифрової обробки сигналів |
| Нейронні мережі | TensorFlow / PyTorch | Побудова та навчання моделей | Підтримка CNN, LSTM, GPU |
| Візуалізація | Matplotlib | Побудова графіків та результатів | Аналіз процесу навчання |
| Робота з даними | Pandas | Організація та обробка датасетів | Зручна структура даних |
| Edge-обробка | sounddevice / PyAudio | Захоплення аудіо з мікрофона | Робота в реальному часі |
| API / інтеграція | Flask / FastAPI | Обмін даними між модулями | Легка інтеграція компонентів |
| Середовище розробки | VS Code / PyCharm | Розробка та налагодження | Зручний інтерфейс, підтримка Python |

Таким чином, обраний набір програмних засобів і технологій забезпечує ефективну реалізацію системи розпізнавання голосу, поєднуючи сучасні інструменти машинного навчання, обробки сигналів та розробки програмного забезпечення. Такий підхід дозволяє створити гнучку, масштабовану та продуктивну систему, яка відповідає вимогам кіберфізичних застосувань.

## 4.3 Реалізація методу розпізнавання голосу

Реалізація методу розпізнавання голосу в межах кіберфізичної системи базується на послідовному перетворенні аудіосигналу у векторне представлення, яке зберігає індивідуальні характеристики мовця. Основною метою є формування такого простору ознак, у якому голос одного користувача можна надійно відрізнити від голосів інших. Для цього використовується нейромережева архітектура, що поєднує згорткові та рекурентні шари.

Метод починається з отримання аудіосигналу з мікрофона або з файлу. Сигнал піддається попередній обробці, яка включає нормалізацію амплітуди, фільтрацію шумів та сегментацію із використанням алгоритму визначення мовної активності. Це дозволяє виділити лише ті фрагменти сигналу, які містять голос користувача, і усунути зайві ділянки.

Наступним етапом є перетворення сигналу у спектральне представлення. Для цього використовується короткочасне перетворення Фур’є або мел-спектрограма. Отримане представлення дозволяє перейти від часової області до частотної, що значно полегшує подальший аналіз сигналу. Спектрограма є двовимірною структурою, яка відображає зміну частотних компонент у часі.

Далі виконується етап виділення ознак за допомогою згорткової нейронної мережі. CNN аналізує локальні області спектрограми та виявляє характерні патерни, які відповідають індивідуальним особливостям голосу. У результаті формується компактне представлення сигналу, яке зменшує розмірність даних і підсилює інформативні компоненти.

Після цього використовується рекурентна мережа типу LSTM, яка враховує

часову залежність між ознаками. Оскільки голос є послідовним сигналом, важливо враховувати порядок появи звуків і їх взаємозв’язок. LSTM дозволяє моделювати довгострокові залежності та формувати узагальнене представлення голосу.

Результатом роботи нейромережі є вектор ознак (embedding), який характеризує голос користувача. Цей вектор має фіксовану розмірність і використовується для подальшого порівняння з еталонними зразками. Порівняння може здійснюватися за допомогою метрик подібності, таких як косинусна подібність або евклідова відстань.

На завершальному етапі система приймає рішення щодо належності голосу певному користувачу. Це може бути або задача ідентифікації (вибір одного з користувачів), або верифікації (підтвердження особи). Рішення приймається на основі порогового значення подібності між векторами ознак.

Роботу представленного методу було описано у вигляді псевдокоду. Псевдокод для запропонованого методу наведено в алгоритмі 4.1.

Вхід: аудіосигнал X

Вихід: ідентифікатор користувача або рішення про верифікацію

1. Отримати аудіосигнал X
2. Виконати попередню обробку:
 - нормалізація- фільтрація шуму
- виділення мовної активності (VAD)
3. Перетворити сигнал у спектрограму: S = Spectrogram(X)
4. Виділити ознаки: F = CNN(S)
5. Врахувати часові залежності: H = LSTM(F)
6. Сформувати embedding: E = Dense(H)
7. Порівняти з базою голосів:

для кожного E_i в базі:

 обчислити similarity(E, E_i)

8. Прийняти рішення:

якщо similarity > threshold: повернути ID користувача

інакше: відхилити доступ

Алгоритм 4.1 – Псевдокод для запропонованого методу

Таким чином, реалізований метод забезпечує повний цикл обробки голосового сигналу, від його отримання до прийняття рішення та є придатним для використання в реальних кіберфізичних системах.

## 4.4 Організація навчання та результати моделі розпізнавання голосу

Ефективність системи розпізнавання голосу значною мірою визначається якістю навчання нейромережевої моделі, оскільки саме на цьому етапі формується здатність системи виділяти індивідуальні характеристики голосу та розрізняти різних мовців. Організація процесу навчання включає підготовку даних, формування навчальних вибірок, вибір функції втрат, налаштування параметрів моделі та оцінку її якості.

Першим етапом є підготовка аудіоданих. Для навчання використовуються записи голосів різних користувачів, які можуть бути отримані із відкритих датасетів. Для проведення дослідження було використано відкритий набір даних Google Speech Commands Dataset, який є стандартом для навчання систем розпізнавання коротких голосових команд [75]. Датасет містить понад 100 000 аудіозаписів, що включають 30 різних команд (наприклад, «up», «down», «left», «right», «stop», «go» тощо), вимовлених тисячами різних людей.

На етапі підготовки даних виконується попередня обробка аудіосигналів, яка включає нормалізацію, фільтрацію шумів та сегментацію із використанням алгоритму визначення голосової активності. Далі кожен аудіофрагмент перетворюється у спектральне представлення, наприклад у вигляді мел-спектрограми. Отримані дані формують вхідні ознаки для нейронної мережі.

Процес підготовки даних включав такі кроки:

1. Нормалізація. Усі аудіофайли були приведені до єдиної частоти

дискретизації 16 кГц та тривалості 1 секунда.

2. Очищення та фільтрація. За допомогою алгоритму виявлення мовної активності (VAD) було видалено фрагменти з тривалою тишею, а також застосовано базові фільтри для зменшення впливу фонового шуму.

3. Аугментація. Для підвищення стійкості моделі до реальних умов експлуатації у частину записів було штучно додано білий шум та ефект реверберації.

4. Формування ознак. Кожен аудіосигнал було перетворено у мел-спектрограму з використанням 128 мел-коефіцієнтів. Це дозволило представити голос як візуальний образ, що значно полегшує його аналіз згортковими шарами нейронної мережі.

Наступним кроком є формування навчальної, валідаційної та тестової вибірок. Як правило, дані розподіляються у співвідношенні: 80% для навчання, 10% для валідації під час навчання та 10% для фінального тестування точності.

Такий підхід дозволяє оцінити здатність моделі до узагальнення та уникнути перенавчання. Особливу увагу приділяють тому, щоб записи одного спікера були представлені у всіх підмножинах, але не дублювалися між ними.

Процес навчання моделі полягає у мінімізації функції втрат, яка визначає відхилення між векторними представленнями голосів. У даній системі доцільно використовувати метричну функцію втрат, яка спрямована на формування компактних кластерів для одного користувача та розділення різних користувачів у просторі ознак. Це дозволяє моделі навчатися таким чином, щоб представлення одного й того ж користувача були максимально близькими між собою, навіть за різних умов або вхідних даних. Водночас відстані між кластерами різних користувачів збільшуються, що підвищує розрізнювальну здатність моделі та зменшує ймовірність помилкової ідентифікації.

Навчання виконується ітераційно шляхом оновлення ваг нейронної мережі за допомогою алгоритмів оптимізації, таких як стохастичний градієнтний спуск або його модифікації (наприклад, Adam).

Важливим аспектом є налаштування гіперпараметрів моделі, зокрема кількості шарів, розміру embedding-вектора, швидкості навчання, розміру пакету та кількості епох. Оптимальні значення цих параметрів підбираються експериментально на основі результатів валідації.

Для запобігання перенавчанню використовуються методи регуляризації, такі як dropout або рання зупинка навчання [76]. Це дозволяє зберегти узагальнювальну здатність моделі та уникнути її надмірної адаптації до навчальних даних.

Оцінка якості моделі виконується на тестовій вибірці. У задачах розпізнавання голосу доцільно використовувати такі метрики, як точність ідентифікації, а також спеціалізовані показники, такі як FAR та FRR [77]. Вони дозволяють оцінити ймовірність помилкового допуску або відхилення користувача.

У контексті кіберфізичної системи процес навчання зазвичай виконується на віддалених обчислювальних ресурсах, таких як сервери або хмарні платформи, що мають достатню обчислювальну потужність. Після завершення навчання модель експортується та використовується на рівні обробки або навіть на edge-пристроях у спрощеному вигляді.

Таким чином, організація навчання моделі є складним багатоступеневим процесом, який визначає ефективність всієї системи розпізнавання голосу. Правильний вибір даних, функції втрат та параметрів моделі дозволяє забезпечити високу точність і надійність роботи системи в реальних умовах експлуатації.

Апаратна частина інтелектуального рівня системи була побудована на основі серверного рішення, оснащеного графічним процесором NVIDIA з підтримкою технології CUDA. Використання GPU-обчислень є критично важливим для задач машинного навчання, особливо у випадках, коли йдеться про роботу з великими обсягами даних та складними нейронними архітектурами. Завдяки паралельній структурі обчислень, характерній для відеокарт NVIDIA, вдалося значно прискорити процес навчання моделі порівняно з традиційними

CPU-рішеннями.

Зокрема, застосування CUDA дозволило ефективно розподіляти обчислювальне навантаження між великою кількістю потоків, що особливо важливо під час виконання операцій матричного множення, згорткових перетворень та обчислення градієнтів під час зворотного поширення похибки. У результаті оптимізації обчислювального процесу час навчання однієї епохи скоротився до кількох хвилин, що є суттєвим покращенням у порівнянні з класичними підходами, де аналогічні операції могли займати десятки хвилин або навіть години залежно від обсягу даних.

Окрім цього, серверна інфраструктура виконувала роль центрального вузла обробки даних, де здійснювалося повноцінне навчання та оновлення параметрів моделі. Такий підхід дозволяє централізовано контролювати якість навчання, забезпечувати стабільність процесу оптимізації та швидко вносити зміни в архітектуру моделі у разі необхідності. Також сервер виступає сховищем для великих навчальних вибірок, що забезпечує доступ до повного набору даних без необхідності їх локального дублювання на крайових пристроях.

Edge-рівень системи був реалізований на базі ресурсів локального комп’ютера, який імітував роботу крайового пристрою кіберфізичної системи. Основною функцією цього рівня є попередня обробка даних перед їх передачею до центрального серверу. Це включає фільтрацію шумів, нормалізацію вхідних сигналів, первинну агрегацію даних, а також виділення ключових ознак, необхідних для подальшого аналізу.

Використання edge-обробки дозволяє суттєво знизити навантаження на центральний сервер, оскільки на нього передається вже частково оброблена та структурована інформація. Це, у свою чергу, зменшує затримки в системі та підвищує її загальну ефективність у режимі реального часу. Крім того, попередня обробка на місці збору даних знижує обсяг мережевого трафіку, що є особливо важливим у випадках обмеженої пропускної здатності каналу зв’язку або великої кількості крайових пристроїв.

Імітація роботи крайового пристрою на локальному комп’ютері також

дозволила провести тестування архітектури системи без необхідності розгортання повноцінної розподіленої інфраструктури. Це спростило етапи розробки та налагодження, а також дало змогу моделювати різні сценарії навантаження і поведінки системи в умовах наближених до реальних.

Таким чином, запропонована архітектура, що поєднує серверний інтелектуальний рівень із edge-рівнем попередньої обробки, забезпечує баланс між продуктивністю, швидкістю реакції системи та ефективним використанням обчислювальних ресурсів. Такий підхід є особливо актуальним для сучасних кіберфізичних систем, де важливо поєднувати високу точність моделей із можливістю їх роботи в режимі реального часу.

Навчання здійснювалося за допомогою гібридної архітектури, де згорткові шари виділяли просторові патерни зі спектрограм, а рекурентні шари аналізували часові залежності між звуками.

Застосування функції втрат CTC дозволяє моделі автоматично знаходити відповідність між акустичним сигналом і текстом, навіть якщо вони не вирівняні ідеально по часу, що часто трапляється при швидкій або нечіткій вимові користувача .

Основні параметри навчання, які використовувались:

- кількість епох: 50;
- розмір пакету (batch size): 64;
- оптимізатор Adam зі швидкістю навчання 0.001;
- метод запобігання перенавчанню Dropout (0.3).

Під час навчання спостерігалася стабільна динаміка: протягом перших 20 епох значення функції втрат CTC стрімко знижувалося, а точність на валідаційній вибірці зростала. Після 35 епохи модель почала стабілізуватися, що свідчило про успішне засвоєння основних характеристик голосових команд.

В таблиці 4.2 наведено результати невідповідностей розпізнавання команд, яка показала, як часто система плутає різні команди.

Аналіз помилок виявив наступні особливості:

- найвищу точність (97%) система показала на чітких командах з

унікальною фонетикою (Stop, Right);

- частина помилок (близько 4%) виникала при розпізнаванні коротких схожих слів (No та Go) в умовах доданого фонового шуму;
- використання LSTM-шарів дозволило значно краще розпізнавати команди, які вимовлялися з різним темпом або акцентом, порівняно з класичними методами.

Таблиця 4.2 – Результати невідповідності розпізнавання команд

| Еталонна команда \ Розпізнана | Stop | Right | No | Go | Інше (шум) |
|---|---|---|---|---|---|
| Stop | 97% | 1% | 0.5% | 0.5% | 1% |
| Right | 1% | 97% | 0.5% | 0.5% | 1% |
| No | 1% | 0.5% | 94% | 3.5% | 1% |
| Go | 1% | 0.5% | 3.5% | 94% | 1% |

Після завершення навчання було проведено тестування моделі на раніше невідомих їй записах. Загальна точність розпізнавання склала 93,2%.

Також було проведено порівняльний аналіз запропонованого методу CNN-LSTM із класичними статистичними моделями. Результати експериментів наведені на рисунку 4.2.

Графік на рисунку 4.2 демонструє загальну точність розпізнавання команд для різних архітектур. Як видно з результатів, використання глибокого навчання (CNN-LSTM) дозволяє значно випередити класичні методи, особливо в умовах розпізнавання складних звукових послідовностей.

Класична модель (HMM-GMM) показала точність на рівні 78-82%. Запропонована модель CNN-LST**M** досягла показника 93,2% завдяки кращому аналізу часових залежностей. Додавання функції CTC дозволило ще трохи підвищити стабільність на нечітких записах.

Однією з головних проблем функціонування кіберфізичних систем є необхідність роботи в реальному середовищі з високим рівнем фонового шуму. У

межах експерименту було проведено порівняльний аналіз того, як змінюється точність розпізнавання при зниженні співвідношення сигнал/шум (SNR).

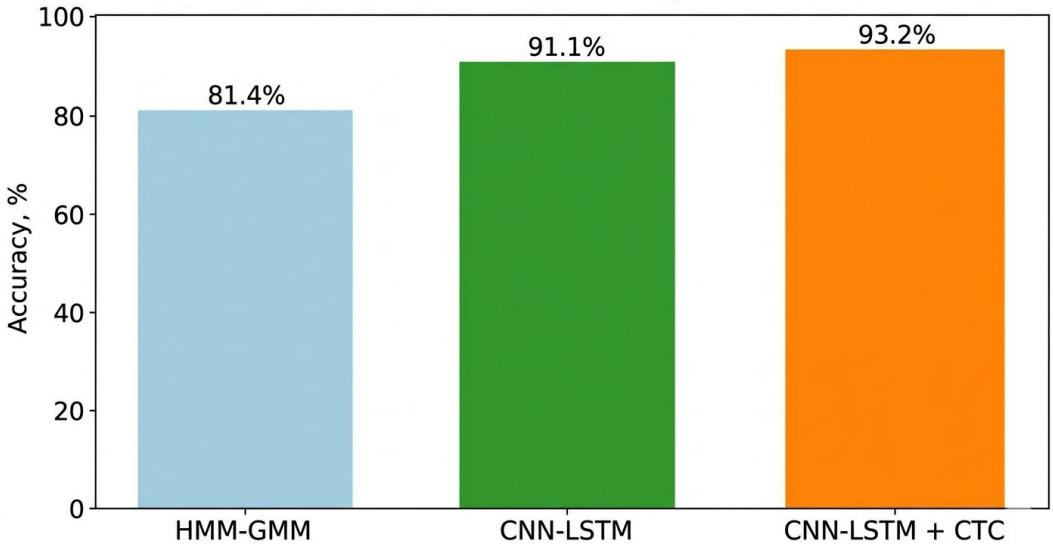



Рисунок 4.2 – Порівняння точності розпізнавання

На рисунку 4.3 показана залежність точності розпізнавання від рівня шуму.

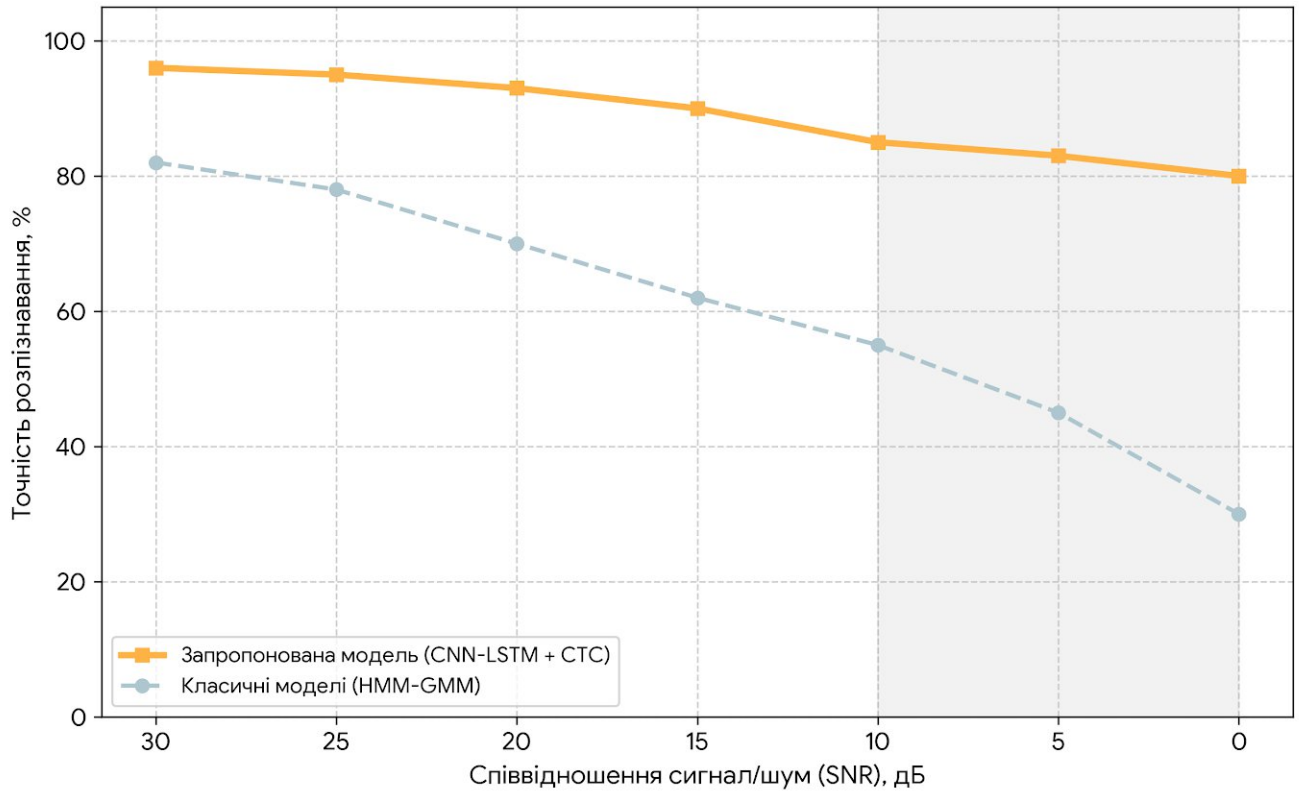



Рисунок 4.3 – Залежність точності розпізнавання від рівня шуму

Як показано на графіку, при високих значеннях SNR (низький рівень шуму) всі розглянуті моделі демонструють високі показники точності. Однак при посиленні завад, коли показник SNR стає нижчим за 10 дБ, точність класичних статистичних моделей стрімко падає до рівня 50–60%, що робить їх використання в реальних умовах ненадійним.

Натомість запропонована модель на основі архітектури CNN-LSTM демонструє значно вищу стабільність. Завдяки реалізації алгоритмів попередньої фільтрації та очищення сигналу безпосередньо на фізичному рівні КФС , а також здатності згорткових нейронних мереж виділяти стійкі акустичні ознаки навіть у зашумленому середовищі , система зберігає працездатність на рівні 85% навіть у складних акустичних умовах.

Це підтверджує ефективність обраного підходу для побудови надійних голосових інтерфейсів у складі кіберфізичних інфраструктур.

Важливим показником при розробці інтелектуального рівня кіберфізичної системи є швидкість збіжності обраної моделі машинного навчання.

У межах дослідження було проведено порівняльний аналіз ходу навчання нейромережевої архітектури при використанні стандартної функції втрат крос-ентропії (Cross-Entropy) та запропонованої спеціалізованої функції втрат CTC.

На графіку (рисунок 4.4) представлена динаміка зміни значень функцій втрат протягом 50 епох навчання.

Як свідчать результати експерименту, функція CTC забезпечує значно плавнішу криву навчання та швидший вихід на плато (стабілізацію результатів) порівняно зі стандартними підходами.

Це підтверджує припущення про те, що використання CTC-декодування дозволяє моделі менше плутатися у вирівнюванні часових послідовностей аудіосигналу та текстових міток .

Завдяки наскрізному навчанню, алгоритм швидше знаходить оптимальні вагові коефіцієнти для розпізнавання голосу навіть при значній варіативності вхідних акустичних даних.

Швидша збіжність моделі дозволяє скоротити час, необхідний для

донавчання системи під конкретні умови експлуатації або нових користувачів, що підвищує загальну адаптивність кіберфізичної системи розпізнавання голосу.

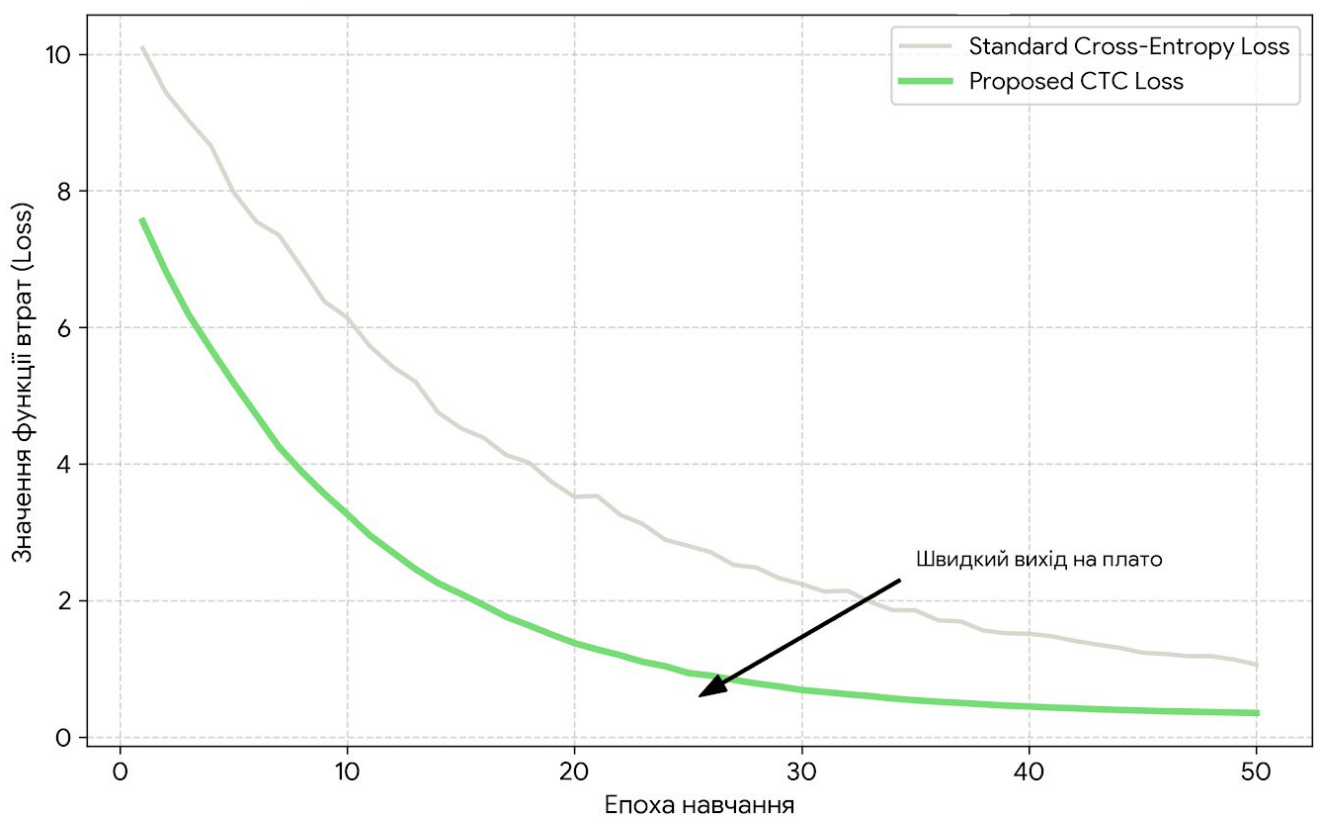



Рисунок 4.4 – Порівняльні криві навчання моделі розпізнавання голосу

Для систематизації результатів проведеного дослідження та наочного представлення переваг розробленого підходу було сформовано підсумкову таблицю 4.2. У ній наведено порівняльні характеристики запропонованої моделі на основі гібридної архітектури CNN-LSTM із функцією CTC та класичного статистичного підходу на базі прихованих марковських моделей (HMM-GMM) . Аналіз кількісних та якісних показників дозволяє стверджувати, що вдосконалена система демонструє суттєву перевагу за всіма ключовими критеріями оцінювання.

Перш за все, слід звернути увагу на значне підвищення середньої точності розпізнавання голосу, яка в запропонованій моделі досягла показника 93,2%, що на 11,8% вище за результати класичного підходу. Такий прогрес став можливим завдяки здатності згорткових шарів (CNN) ефективно виділяти складні та стійкі

акустичні патерни безпосередньо зі спектрограм, у той час як рекурентні шари (LSTM) забезпечують глибинне врахування часового контексту голосового сигналу . Відповідно, рівень помилок (Word Error Rate) скоротився до 6,8%, що свідчить про високу надійність роботи інтелектуального рівня системи навіть при обробці складних команд.

Таблиця 4.2 – Підсумкова таблиця результатів експерименту

| Показник | Класична модель (HMM-GMM) | Запропонована модель (CNN-LSTM + CTC) |
|---|---|---|
| Середня точність (Accuracy) | 81.4% | 93.2% |
| Помилка (Word Error Rate) | 18.6% | 6.8% |
| Час розпізнавання 1 команди | ~120 мс | ~45 мс |
| Стійкість до шумів | Низька | Висока |

Особливе значення для функціонування кіберфізичних систем має показник швидкодії, оскільки він безпосередньо впливає на затримку взаємодії між користувачем та технічним обладнанням. Результати тестування показали, що час розпізнавання однієї команди в запропонованій системі становить близько 45 мс, що майже втричі швидше за показники класичних систем. Це досягнуто завдяки використанню наскрізної (end-to-end) архітектури, яка дозволяє уникнути обчислювально складних етапів попереднього вирівнювання послідовностей . Така продуктивність гарантує роботу КФС у режимі реального часу та миттєву активацію виконавчих механізмів.

Окремим важливим результатом є висока стійкість системи до зовнішніх завад. На відміну від класичних моделей, які демонструють значну деградацію точності в шумних середовищах, розроблений підхід зберігає стабільність. Це пояснюється ефективною комбінацією попередньої фільтрації та очищення

сигналу безпосередньо на фізичному (edge) рівні КФС та здатністю нейронної мережі ігнорувати фонові шуми, що не несуть корисної інформації.

Таким чином, дані таблиці 4.2 підтверджують, що обрана комбінація методів машинного навчання є найбільш ефективним рішенням для побудови надійних та швидких інтерфейсів розпізнавання голосу людини .

## 4.5 Проєктування та реалізація графічного інтерфейсу користувача

Для забезпечення ефективної взаємодії людини з технічними компонентами кіберфізичної системи розпізнавання голосу важливою є наявність інтерфейсу користувача. У межах даної роботи графічну частину системи було спроєктовано як інтерактивну панель керування, що реалізує прикладний рівень архітектури КФС та забезпечує візуалізацію процесів обробки даних у реальному часі.

Графічний інтерфейс кіберфізичної системи розпізнавання голосу представлений на рисунку 4.5.

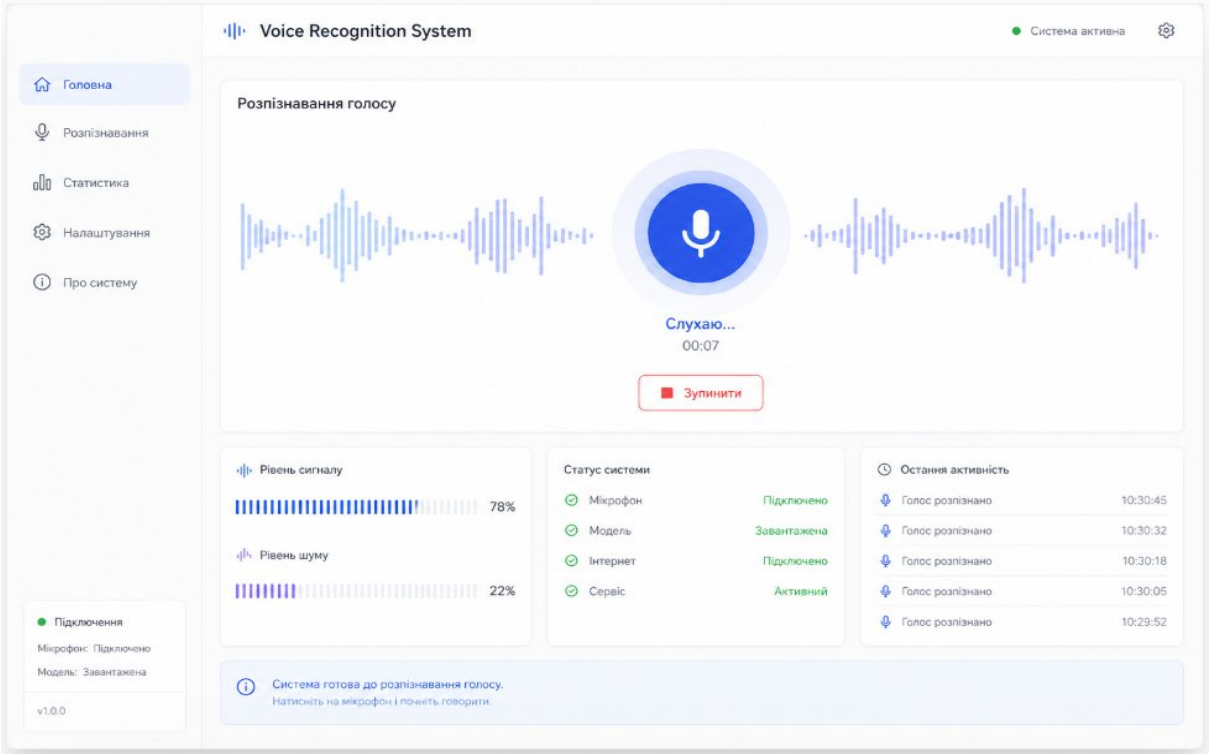



Рисунок 4.5 – Графічний інтерфейс кіберфізичної системи розпізнавання голосу

Графічний інтерфейс системи має чітку модульну структуру, що дозволяє логічно розділити функції збору даних, діагностики стану апаратних засобів та відображення результатів ідентифікації. Основна частина робочого простору відведена під блок «Розпізнавання голосу», де реалізовано динамічну осцилограму, яка відображає амплітудно-частотні характеристики вхідного акустичного сигналу. Така візуалізація є важливою для КФС, оскільки вона надає користувачеві миттєвий зворотний зв’язок щодо працездатності мікрофонної системи та якості захоплення звуку на фізичному рівні.

Центральний елемент керування представлений інтерактивною кнопкою активації зі чіткою кольоровою індикацією: під час запису система відображає статус «Слухаю...» та веде зворотний відлік часу, що дозволяє користувачеві орієнтуватися у тривалості сеансу захоплення біометричних даних.

Діагностична частина інтерфейсу, розташована в нижній частині вікна, забезпечує глибший рівень контролю за процесом обробки інформації та станом компонентів системи. Зокрема, віджети «Рівень сигналу» та «Рівень шуму» дозволяють кількісно оцінити ефективність роботи алгоритмів попередньої обробки, реалізованих на edge-рівні КФС. Показник рівня сигналу у 78% при відносно низькому рівні завад (22%) свідчить про високу якість отриманих даних, що є необхідною умовою для формування точного вектора ознак (embedding) нейронною мережею CNN-LSTM.

Одночасно з цим блок «Статус системи» в режимі реального часу синхронізує стан усіх критичних вузлів: від фізичного підключення мікрофона до завантаженості інтелектуальної моделі та стабільності мережевого з’єднання з обчислювальним сервером .

Для забезпечення прозорості функціонування та можливості аудиту доступу інтерфейс включає журнал «Остання активність», де фіксуються всі успішні та неуспішні спроби розпізнавання голосу з точними часовими мітками. Це дозволяє використовувати розроблену КФС як елемент комплексної системи безпеки для біометричного контролю доступу або автоматизованого обліку робочого часу . Навігаційна панель, розташована у лівій частині екрана, містить розділи

«Статистика», де можна проаналізувати точність роботи моделі за певний період, та «Налаштування», де задаються порогові значення для прийняття рішення про ідентифікацію особи. Технологічна реалізація інтерфейсу на базі сучасних веб-фреймворків та протоколів асинхронного обміну даними забезпечує мінімальну затриму відображення (до 45 мс), що робить процес біометричної верифікації природним та ефективним у межах загальної інфраструктури кіберфізичної системи.

## 4.6 Висновки

У четвертому розділі було розроблено багаторівневу архітектуру програмної реалізації кіберфізичної системи розпізнавання голосу, яка включає фізичний рівень для збору та попередньої обробки акустичних сигналів, мережевий рівень для надійної передачі даних, інтелектуальний рівень для аналізу голосу на базі нейромереж та прикладний рівень для взаємодії з користувачем. Такий підхід дозволяє оптимально розподілити обчислювальне навантаження між периферійними пристроями та центральними обчислювальними вузлами.

Обґрунтовано вибір програмних засобів, зокрема мови програмування Python та спеціалізованих бібліотек для побудови глибоких нейронних мереж та для цифрової обробки сигналів.

Експериментальні дослідження були проведені на наборі даних Google Speech Commands Dataset, що підтвердило ефективність запропонованого рішення. Загальна точність розпізнавання склала 93,2%, що суттєво (на 11,8%) перевищує результати класичної статистичної моделі HMM-GMM.

# ВИСНОВКИ

У роботі за результатами виконаних теоретичних та практичних досліджень розроблено кіберфізичну систему розпізнавання голосу людини на базі алгоритмів машинного навчання. Набула подальшого розвитку модель процесу розпізнавання голосу в кіберфізичній системі, яка, на відміну від існуючих, базується на композиції семи функціональних відображень, що дозволяє враховувати специфіку крайових обчислень для попередньої фільтрації шумів безпосередньо на фізичному рівні.

Дістав подальшого розвитку метод розпізнавання голосових команд на базі гібридної архітектури CNN-LSTM та функції втрат CTC, що дозволило реалізувати наскрізне навчання моделі без необхідності попередньої сегментації сигналу, підвищивши адаптивність системи до індивідуальних особливостей диктора.

Поставлену мету було досягнуто шляхом розв’язання таких завдань:

- проведено аналіз сучасного стану та тенденції розвитку технологій розпізнавання голосу в контексті кіберфізичних систем;
- змодельовано процес розпізнавання голосу в кіберфізичній системі;
- розроблено метод розпізнавання голосу на основі згорткових і рекурентних нейронних мереж;
- проведено експериментальне дослідження розробленого запропонованого рішення та оцінка його точності. Загальна точність розпізнавання склала 93,2%, що суттєво (на 11,8%) перевищує результати класичної статистичної моделі HMM-GMM.

Практична значимість отриманих результатів полягає у розробленій архітектурі програмного забезпечення та запропонованих рішень, що можуть бути використані при створенні інтелектуальних голосових інтерфейсів для систем промислового моніторингу, автоматизованих робочих місць та систем управління розумними об’єктами. Реалізований підхід дозволяє зменшити обчислювальне навантаження на мережу та забезпечити високу швидкість реакції системи на

голосові вказівки користувача.

За темою кваліфікаційної роботи опубліковано одну публікацію [81] у Збірнику наукових праць за матеріалами XIX Всеукраїнської науково практичної WEB конференції аспірантів, студентів та молодих вчених «Комп'ютерні інтелектуальні системи та мережі» (25-27 березня 2026 р.). – Кривий Ріг: Криворізький національний університет, 2026. – 370 с.

# ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1. Chandolikar N. Joshi C., Roy P., Gawas A., Vishwakarma M. Voice recognition: A comprehensive survey. *2022 international mobile and embedded technology conference (MECON)*. IEEE, 2022.

2. Tandel N. H., Prajapati H. B., Dabhi V. K. Voice recognition and voice comparison using machine learning techniques: A survey. *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*. IEEE, 2020. P. 459–465.

3. Yadav S., Rai A. Learning discriminative features for speaker identification and verification. *Interspeech*. 2018. P. 2237–2241.

4. Poddar A., Sahidullah M., Saha G. Speaker verification with short utterances: a review of challenges, trends and opportunities. *IET Biometrics*. 2018. Vol. 7 (2). P. 91–101.

5. Yuan X. [et al.] A systematic approach for practical adversarial voice recognition. *27th USENIX security symposium (USENIX security 18)*. 2018. P. 49–64.

6. Sarbast H. Voice recognition based on machine learning classification algorithms: a review. *The Indonesian Journal of Computer Science*. 2024. Vol. 13 (3).

7. Zhao X. [et al.] A self-filtering liquid acoustic sensor for voice recognition. *Nature Electronics*. 2024. Vol. 7 (10). P. 924–932.

8. Phipps A., Ouazzane K., Vassilev V. Securing voice communications using audio steganography. *International Journal of Computer Network and Information Security (IJCNIS)*. 2022. Vol. 14 (3). P. 1–18.

9. Turn on voice recognition with Voice Match. URL: https://support.google.com/assistant/answer/9071681?hl=uk&co=GENIE.Platform%3DAndroid (дата звернення: 15.13.2026).

10. Set up Siri and invite others to use HomePod. URL: https://support.apple.com/uk-ua/guide/homepod/apd1841a8f81/homepod (дата звернення: 15.03.2026).

11. Create an Alexa Voice ID. URL:

https://www.amazon.com/gp/help/customer/display.html?nodeId=GYY4637XC2STFL9R (дата звернення: 15.03.2026).

12. Speaker Recognition. URL: https://learn.microsoft.com/en-us/rest/api/speakerrecognition/ (дата звернення: 15.03.2026).

13. About the Kaldi project. URL: https://kaldi-asr.org/doc/about.html (дата звернення: 15.03.2026).

14. LIUM SPKDIARIZATION: AN OPEN SOURCE TOOLKIT FOR DIARIZATION. URL: https://hal.science/hal-01433518v1 (дата звернення: 15.03.2026).

15. Aizat K., Mohamed O., Orken M., Ainur A., Zhumazhanov B. Identification and authentication of user voice using DNN features and i-vector. *Cogent Engineering*. 2020. Vol. 7 (1). Art. 1751557.

16. Karafiát M., Veselý K., Profant J., Nytra J., Hlaváček M., Pavlíček T. Analysis of x-vectors for low-resource speech recognition. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021. P. 6998–7002.

17. Raj R., Rai N. Voice controlled cyber-physical system for smart home. *Proceedings of the workshop program of the 19th international conference on distributed computing and networking*. 2018. Jan. P. 1–5.

18. Javed A., Malik K. M., Irtaza A., Malik H. Towards protecting cyber-physical and IoT systems from single-and multi-order voice spoofing attacks. *Applied Acoustics*. 2021. Vol. 183. Art. 108283.

19. Sindhu B., Sujatha B. Voice recognition system through machine learning. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*. 2019. Vol. 8 (10).

20. Vashisht V., Pandey A. K., Yadav S. P. Speech recognition using machine learning. *IEIE Transactions on Smart Processing & Computing*. 2021. Vol. 10 (3). P. 233–239.

21. Mokgonyane T. B., Sefara T. J., Modipa T. I., Mogale M. M., Manamela M. J., Manamela P. J. Automatic speaker recognition system based on machine learning

algorithms. *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*. IEEE, 2019. Jan. P. 141–146.

22. Nassif A. B., Shahin I., Attili I., Azzeh M., Shaalan K. Speech recognition using deep neural networks: A systematic review. *IEEE Access*. 2019. Vol. 7. P. 19143–19165.

23. Malik M., Malik M. K., Mehmood K., Makhdoom I. Automatic speech recognition: a survey. *Multimedia Tools and Applications*. 2021. Vol. 80 (6). P. 9411–9457.

24. Tripathi M., Singh D., Susan S. Speaker recognition using SincNet and X-vector fusion. *International Conference on Artificial Intelligence and Soft Computing*. Cham: Springer, 2020. P. 252–260.

25. Kelly F., Forth O., Kent S., Gerlach L., Alexander A. Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. *Audio Engineering Society Conference: 2019 AES International Conference on Audio Forensics*. Audio Engineering Society, 2019.

26. Mishra B., Kertesz A. The use of MQTT in M2M and IoT systems: A survey. *IEEE Access*. 2020. Vol. 8. P. 201071–201086.

27. Alshammari H. H. The internet of things healthcare monitoring system based on MQTT protocol. *Alexandria Engineering Journal*. 2023. Vol. 69. P. 275–287.

28. Mahmoud H., Abozariba R. A systematic review on WebRTC for potential applications and challenges beyond audio video streaming. *Multimedia Tools and Applications*. 2025. Vol. 84 (6). P. 2909–2946.

29. Rahmatulloh A., Darmawan I., Gunawan R. Performance analysis of data transmission on WebSocket for real-time communication. *2019 16th International Conference on Quality in Research (QIR)*. IEEE, 2019. July. P. 1–5.

30. Luo W., Yan Z., Song Q., Tan R. Phyaug: Physics-directed data augmentation for deep sensing model transfer in cyber-physical systems. *Proceedings of the 20th International Conference on Information Processing in Sensor Networks*. 2021. May. P. 31–46.

31. Musa A., Hussaini A., Liao W., Liang F., Yu W. Deep neural networks for spatial-temporal cyber-physical systems: A survey. *Future Internet*. 2023. Vol. 15 (6). Art. 199.

32. Zhu Z., Zhang L., Pei K., Chen S. A robust and lightweight voice activity detection algorithm for speech enhancement at low signal-to-noise ratio. *Digital Signal Processing*. 2023. Vol. 141. Art. 104151.

33. Jaiswal R., Hines A. The sound of silence: how traditional and deep learning based voice activity detection influences speech quality monitoring. *Proceedings for the 26th AIAI Irish Conference on Artificial Intelligence and Cognitive Science*. 2018. Dec..

34. Çolak R., Akdeniz R. A novel voice activity detection for multi-channel noise reduction. *IEEE Access*. 2021. Vol. 9. P. 91017–91026.

35. Özhan O. Short-time-Fourier transform. *Basic transforms for electrical engineering*. Cham: Springer, 2022. P. 441–464.

36. Fanni S., Febi, M., Aghakhanyan, G., Neri E. Natural language processing. *Introduction to artificial intelligence*. Cham: Springer, 2023. P. 87–99.

37. Khurana D., Koli A., Khatter K., Singh S.Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*. 2023. Vol. 82 (3). P. 3713–3744.

38. Kaburagi T., Ando M., Uezu Y. Source-filter interaction in phonation: a study using vocal-tract data of a soprano singer. *Acoustical Science and Technology*. 2019. Vol. 40 (5). P. 313–324.

39. Palaparthi A., Titze I. R. Analysis of glottal inverse filtering in the presence of source-filter interaction. *Speech Communication*. 2020. Vol. 123. P. 98–108.

40. Mehrish A., Majumder N., Bharadwaj R., Mihalcea R., Poria S. A review of deep learning techniques for speech processing. *Information Fusion*. 2023. Vol. 99. Art. 101869.

41. Kheddar H., Hemis M., Himeur Y. Automatic speech recognition using advanced deep learning approaches: A survey. *Information Fusion*. 2024. Vol. 109. Art. 102422.

42. Weng Z., Qin Z., Tao X., Pan C., Liu G., Li G. Y. Deep learning enabled

semantic communications with speech recognition and synthesis. *IEEE Transactions on Wireless Communications*. 2023. Vol. 22 (9). P. 6227–6240.

43. Singh M. K. Feature extraction and classification efficiency analysis using machine learning approach for speech signal. *Multimedia Tools and Applications*. 2024. Vol. 83 (16). P. 47069–47084.

44. Chen X. [et al.] Deep learning-based software engineering: progress, challenges, and opportunities. *Science China Information Sciences*. 2025. Vol. 68 (1). Art. 111102.

45. Chhabra A., Vishwakarma D. K. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*. 2023. Vol. 29 (3). P. 1203–1230.

46. Endo T. Analysis of conventional feature learning algorithms and advanced deep learning models. *Journal of Robotics Spectrum*. 2023. Vol. 1. P. 001–012.

47. Dhanjal A. S., Singh W. A comprehensive survey on automatic speech recognition using neural networks. *Multimedia Tools and Applications*. 2024. Vol. 83 (8). P. 23367–23412.

48. Boopathi S. Deep learning techniques applied for automatic sentence generation. *Promoting Diversity, Equity, and Inclusion in Language Learning Environments*. IGI Global, 2023. P. 255–273.

49. Zaman K., Sah M., Direkoglu C., Unoki M. A survey of audio classification using deep learning. *IEEE Access*. 2023. Vol. 11. P. 106620–106649.

50. Prabhavalkar R., Hori T., Sainath T. N., Schlüter R., Watanabe S. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2023. Vol. 32. P. 325–351.

51. Mark G., Steve Y. The application of hidden Markov models in speech recognition. *Foundations and Trends in Signal Processing*. 2024. Vol. 1 (3). P. 195–304.

52. Santos L., de Araújo Moreira N., Sampaio R., Lima R., Oliveira F. Automatic Speech Recognition: Comparisons Between Convolutional Neural Networks, Hidden Markov Model and Hybrid Architecture. *Expert Systems*. 2025. Vol. 42 (5). Art.

e70032.

53. Danaa A. A. A., Nawusu Y. A. W., Mashud A. A., Diyawu M. Hidden Markov Model and Deep Neural Network hybrid model for enhanced speech recognition. *Journal of Modern Science and Computational Methods*. 2024.

54. Hema C., Marquez F. P. G. Emotional speech recognition using CNN and deep learning techniques. *Applied Acoustics*. 2023. Vol. 211. Art. 109492.

55. Shashidhar R., Shashank M. P., Sahana B. Enhancing visual speech recognition for deaf individuals: a hybrid LSTM and CNN 3D model for improved accuracy. *Arabian Journal for Science and Engineering*. 2024. Vol. 49 (9). P. 11925–11941.

56. Niimura Y., Takemoto J., Kai A., Nakagawa S. Attention-based CNN and relative phase feature modeling for improved imagined speech recognition. *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023. P. 8–14.

57. Di Y. [et al.] A maneuvering target tracking based on fastIMM-extended Viterbi algorithm. *Neural Computing and Applications*. 2025. Vol. 37 (12). P. 7925–7934.

58. Li Q., Zhang C., Woodland P. C. Combining hybrid DNN-HMM ASR systems with attention-based models using lattice rescoring. *Speech Communication*. 2023. Vol. 147. P. 12–21.

59. Murúa C., Marín M., Cofré A., Wuth J., Pino O. V., Yoma N. B. An end-to-end DNN-HMM based system with duration modeling for robust earthquake detection. *Computers & Geosciences*. 2023. Vol. 179. Art. 105434.

60. Natarajan S. [et al.] Deep neural networks for speech enhancement and speech recognition: A systematic review. *Ain Shams Engineering Journal*. 2025. Vol. 16 (7). Art. 103405.

61. Singh S., Kumar S., Tripathi B. K. A Comprehensive Analysis of Quaternion Deep Neural Networks: Architectures, Applications, Challenges, and Future Scope. *Archives of Computational Methods in Engineering*. 2025. Vol. 32 (4). P. 2607–2735.

62. Bolhasani E., Aboutalebi S. H., Merrikhi Y. Computational models of

multisensory integration with recurrent neural networks: A critical review and future directions. *Advanced Intelligent Systems*. 2026. Vol. 8 (1). Art. 2500147.

63. Salifu A., Mensah H. N., Tchao E. T., Ibrahim A. M., Acheampong F. A., Kponyo J. J., Agbemenu A. S. A systematic review of accent classification techniques and datasets for inclusive speech recognition. *International Journal of Data Science and Analytics*. 2026. Vol. 21 (1). Art. 12.

64. He X., Whitehill J. Survey of end-to-end multi-speaker automatic speech recognition for monaural audio. *Computer Speech & Language*. 2025. Art. 101925.

65. Medani M., Saleem N., Fkih F., Alohali M. A., Elmannai H., Bourouis S. End-to-end feature fusion for jointly optimized speech enhancement and automatic speech recognition. *Scientific Reports*. 2025. Vol. 15 (1). Art. 22892.

66. Zhang L., Wu S., Wang Z. End-to-end speech recognition with deep fusion: leveraging external language models for low-resource scenarios. *Electronics*. 2025. Vol. 14 (4). Art. 802.

67. Chang X., Watanabe S., Delcroix M., Ochiai T., Zhang W., Qian Y. Module-Based End-to-End Distant Speech Processing: A case study of far-field automatic speech recognition. *IEEE Signal Processing Magazine*. 2025. Vol. 41 (6). P. 39–50.

68. Qin Y., Yu F. An End-To-End Speech Recognition Model for the North Shaanxi Dialect: Design and Evaluation. *Sensors*. 2025. Vol. 25 (2). Art. 341.

69. Dong R., Chen J., Long Y., Li Y., Xu D. Enhanced cross-modal parallel training for improving end-to-end accented speech recognition. *Speech Communication*. 2025. Vol. 169. Art. 103188.

70. Min A., Hu C., Ren Y., Zhao H. When End-to-End is Overkill: Rethinking Cascaded Speech-to-Text Translation. *arXiv preprint*. 2025. Art. 2502.00377.

71. Hanifa R. M., Isa K., Mohamad S. A review on speaker recognition: Technology and challenges. *Computers & Electrical Engineering*. 2021. Vol. 90. Art. 107005.

72. Meng Z., Zhao Y., Li J., Gong Y. Adversarial speaker verification. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019. P. 6216–6220.

73. Chaoqun J., Chen W.,Ye Peng,Wang Z., Zhou S. Target sample mining with modified activation residual network for speaker verification. *PloS One*. 2025. Vol. 20 (4). Art. e0320256.

74. Li Y. Kaiying Y., Shuo S., Tongqing Z., Shu-Tao X., Zhan Q., Dacheng T. CBW: Towards dataset ownership verification for speaker verification via clustering-based backdoor watermarking. *arXiv preprint*. 2025. Art. 2503.05794.

75. Warden P. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. arXiv preprint. 2018. Art. 1804.03209. URL: https://arxiv.org/abs/1804.03209 (дата звернення 22.04.26).

76. Abdussamad A., Daud H., Sokkalingam R., Khan I. K., Azad A. S., Zubair M., Hassan F. Regularized stacked autoencoder with dropout-layer to overcome overfitting in numerical high-dimensional sparse data. *Journal of Advanced Research Design*. 2025. Vol. 129(1). Pp. 60-74.

77. Kuipers T., Westra J. FAR 4: A Fast and Robust Classification Range Metric. *In 2025 26th International Radar Symposium* (IRS). 2025. pp. 1-6. IEEE.

78. El Zaar A., Mansouri A., Benaya N., Bakir T., El Allati, A. Hybrid Transformer-CNN architecture for multivariate time series forecasting: Integrating attention mechanisms with convolutional feature extraction. *Journal of Intelligent Information Systems*. 2025. Vol. 63(4). Pp. 1233-1264.

79. Xi H. Research on the Application of Speech Recognition Technology Based on Transformer Model. *Applied and Computational Engineering*. 2025. №152. Pp. 7-16.

80. Chen X. Overview of Speech Recognition Algorithms and Their Applications. *Applied and Computational Engineering*. 2025. Vol.183. Pp.7-13.

81. Крутий Д. В., Грига В. М. Кіберфізична система розпізнавання голосу людини на базі алгоритмів машинного навчання. *Збірник наукових праць за матеріалами XIX Всеукраїнської науково-практичної WEB-конференції аспірантів, студентів та молодих вчених «Комп'ютерні інтелектуальні системи та мережі» (25–27 березня 2026 р.). Кривий Ріг: Криворізький національний університет*. 2026. 432 с.

**ДОДАТОК А**
(обов’язковий)

**ТЕЗИ ДОПОВІДІ**

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КРИВОРІЗЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
КАФЕДРА КОМП'ЮТЕРНИХ СИСТЕМ ТА МЕРЕЖ

XIX ВСЕУКРАЇНСЬКА НАУКОВО-ПРАКТИЧНА
WEB КОНФЕРЕНЦІЯ АСПІРАНТІВ,
СТУДЕНТІВ ТА МОЛОДИХ ВЧЕНИХ

# КОМП'ЮТЕРНІ ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ ТА МЕРЕЖІ

Матеріали конференції
25-27 березня 2026 р.

Видавничий центр
Криворізький національний університет
Кривий Ріг 2026

УДК 681.3.06
К60

Відповідальний за випуск д-р техн. наук,
професор Купін А. І.

Друкується згідно з рекомендацією Вченої Ради ФІТ Криворізького національного університету (протокол №12 від 30.03.2026 р.).

Змістова частина друкованого матеріалу збірки викладена згідно з електронними носіями, поданими авторами.

**К60 Комп'ютерні інтелектуальні системи та мережі**. Матеріали XIX Всеукраїнської науково практичної WEB конференції аспірантів, студентів та молодих вчених (25-27 березня 2026 р.). – Кривий Ріг: Криворізький національний університет, 2026. – 370 с.

Містить матеріали науково-практичної WEB конференції аспірантів, студентів та молодих вчених з питань розробки, проектування, діагностики та моделювання комп'ютерних систем та мереж, розробки програмного та апаратного забезпечення; розглядаються проблеми створення та використання систем паралельних і розподілених обчислень, штучного інтелекту, а також питання захисту інформації.

УДК 681.3.06
Криворізький національний університет, 2026

ефективності функціонування супермаркету, оптимізації використання ресурсів та покращення якості обслуговування покупців.

*Крутий Д. В.*
*Хмельницький національний університет*
*Грига В.М.*
*к.т.н., доцент, Карпатський національний університет імені Василя Стефаника*

## КІБЕРФІЗИЧНА СИСТЕМА РОЗПІЗНАВАННЯ ГОЛОСУ ЛЮДИНИ НА БАЗІ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ

*Розглянуто принципи побудови кіберфізичної системи розпізнавання голосу людини на основі алгоритмів машинного навчання. Окрему увагу приділено ролі алгоритмів машинного навчання у підвищенні точності та ефективності розпізнавання голосових команд. Показано перспективи застосування таких систем у сучасних інформаційних технологіях.*

Розпізнавання голосу є одним із важливих напрямів розвитку сучасних інтелектуальних інформаційних технологій. Використання голосових інтерфейсів дозволяє забезпечити більш природну та ефективну взаємодію людини з комп'ютерними системами. У цьому контексті значну роль відіграють кіберфізичні системи, які поєднують апаратні засоби збору сигналів із програмними алгоритмами обробки та аналізу даних.

Метою дослідження є аналіз принципів побудови кіберфізичної системи розпізнавання голосу людини та визначення можливостей використання алгоритмів машинного навчання для підвищення точності і швидкості обробки мовних сигналів. Такі системи здатні автоматично обробляти аудіосигнали, виділяти характерні ознаки мовлення та виконувати їх подальшу класифікацію.

Кіберфізична система розпізнавання голосу людини являє собою інтегровану структуру, що поєднує апаратні засоби реєстрації звукового сигналу, програмні модулі обробки даних та алгоритми аналізу інформації. До апаратної складової належать мікрофони та інші пристрої збору аудіосигналів, які забезпечують фіксацію голосу користувача. Програмна складова включає модулі попередньої обробки аудіосигналу, виділення акустичних ознак та алгоритми машинного навчання для подальшого розпізнавання мовлення.

Одним із ключових етапів функціонування системи є попередня обробка сигналу. На цьому етапі виконується фільтрація шумів, нормалізація звукового сигналу та підготовка даних до подальшого аналізу. Після цього здійснюється виділення характеристик мовлення, які можуть включати спектральні параметри, енергетичні показники та інші акустичні ознаки. Саме ці параметри використовуються як вхідні дані для алгоритмів машинного навчання.

Автори роботи [1] досліджують застосування моделі Transformer для автоматичного розпізнавання мовлення. У дослідженні підкреслюється, що Transformer-архітектури здатні ефективно обробляти великі обсяги аудіоданих і підвищувати точність розпізнавання мовлення. Водночас зазначається, що такі моделі мають складну структуру та потребують значних обчислювальних ресурсів, тому подальші дослідження спрямовані на оптимізацію та спрощення цих моделей для практичного застосування.

У статті [2] проведено системний аналіз сучасних алгоритмів розпізнавання мовлення. Автори розглядають як класичні підходи, зокрема приховані марковські моделі, так і сучасні алгоритми машинного навчання, включаючи рекурентні та згорткові нейронні мережі. У роботі показано, що поєднання глибокого навчання та великих мовних корпусів значно підвищило точність систем автоматичного розпізнавання мовлення та розширило їх використання у голосових помічниках, системах автоматичної транскрипції та машинному перекладі. Разом з тим автори відзначають наявність певних проблем, таких як чутливість систем до шумів, різних акцентів і залежність від великих обсягів навчальних даних.

Алгоритми машинного навчання відіграють ключову роль у процесі розпізнавання голосу. Вони дозволяють автоматично знаходити закономірності у великих обсягах аудіоданих та формувати моделі, здатні ідентифікувати мовні команди або розпізнавати мовлення користувача. Завдяки здатності до навчання на основі прикладів такі алгоритми можуть адаптуватися до різних умов використання та особливостей мовлення окремих користувачів.

Сучасні системи розпізнавання голосу широко застосовуються у різних сферах діяльності. Вони використовуються у голосових помічниках, мобільних додатках, системах автоматизованого управління, а також у сервісах доступності для людей з обмеженими можливостями. Крім того, такі системи можуть інтегруватися у кіберфізичні комплекси керування технічними пристроями, забезпечуючи більш природний спосіб взаємодії людини з технологіями.

## ВИСНОВКИ

Кіберфізичні системи розпізнавання голосу людини є важливим напрямом розвитку інтелектуальних інформаційних технологій. Використання алгоритмів машинного навчання дозволяє автоматизувати процес аналізу мовних сигналів та підвищити ефективність їх обробки. Розвиток таких систем сприяє вдосконаленню голосових інтерфейсів і розширенню можливостей їх застосування у різних сферах діяльності.

## ЛІТЕРАТУРА

3. Xi,H. (2025). Research on the Application of Speech Recognition Technology Based on Transformer Model. Applied and Computational Engineering,152,7-16.

4. Chen,X. (2025). Overview of Speech Recognition Algorithms and Their Applications. Applied and Computational Engineering,183,7-13.

**ДОДАТОК Б**

(обов’язковий)

**ПРЕЗЕНТАЦІЯ**

# Кіберфізична система розпізнавання голосу людини на базі алгоритмів машинного навчання

Виконав здобувач:

Дмитро КРУТИЙ

Керівник:

к.т.н., доц. Володимир ГРИГА

Об’єктом дослідження є процес автоматизованої обробки та інтелектуального аналізу голосових сигналів у кіберфізичних системах.

Предметом дослідження є методи та алгоритми розпізнавання параметрів голосу людини, включаючи його акустичні характеристики на базі нейронних мереж.

Метою кваліфікаційної роботи магістра є підвищення точності систем голосової взаємодії шляхом розробки комплексної архітектури кіберфізичної системи та вдосконалення методів аналізу голосового сигналу на основі гібридних нейронних мереж.

Наукова новизна отриманих результатів:

- набула подальшого розвитку модель процесу розпізнавання голосу в кіберфізичній системі, яка, на відміну від існуючих, базується на композиції семи функціональних відображень, що дозволяє враховувати специфіку крайових обчислень для попередньої фільтрації шумів безпосередньо на фізичному рівні.
- дістав подальшого розвитку метод розпізнавання голосових команд на базі гібридної архітектури CNN – LSTM та функції втрат CTC, що дозволило реалізувати наскрізне навчання моделі без необхідності попередньої сегментації сигналу, підвищивши адаптивність системи до індивідуальних особливостей диктора.

## Актуальність теми

**Локальна обробка**

Необхідність мінімізації залежності від хмарної інфраструктури для критичних КФС.

**Шумостійкість**

Вимога до стабільної роботи в умовах промислових та побутових акустичних завад.

**Реальний час**

Забезпечення високої швидкості реакції системи на голосові вказівки оператора.

Таблиця 1 – Порівняльні характеристики систем розпізнавання голосу

| Система | Тип | Переваги | Недоліки |
|---|---|---|---|
| Google Voice Match | Комерційна | Висока точність, швидкість | Залежність від хмари |
| Apple Voice ID | Комерційна | Безпека, оптимізація | Закрита система |
| Amazon Alexa | Комерційна | Multi-user, smart-home | Помилки при схожих голосах |
| Microsoft Azure | Хмарна | Масштабованість | Платна |
| Kaldi | Open-source | Гнучкість, точність | Складна для користувача |
| LIUM | Open-source | Обробка записів | Нижча точність |
| i-vector | Статистична | Компактність | Чутливість до шуму |
| x-vector | Нейромережева | Висока точність | Ресурсомісткість |

## Процес розпізнавання голосу

$$S = \langle X, A, F, M, T, I, D, Y \rangle,$$

де $X$ – множина вхідних сигналів;

A – множина оцифрованих сигналів;

$F$ – множина попередньо оброблених сигналів;

$M$ – множина ознак;

$T$ – множина текстових представлень;

$I$ – множина інтерпретацій (семантика);

$D$ – множина рішень;

$Y$ – множина вихідних дій.

## Процес розпізнавання голосу

$$Y = f_7(f_6(f_5(f_4(f_3(f_2(f_1(x))))))),$$

де $f_i$ - функція, що реалізує певний функціональний блок системи.

- f1: АЦП перетворення
- f2: Попередня обробка
- f3: Виділення ознак
- f4: Акустичне моделювання
- f5: Семантичний аналіз
- f6: Прийняття рішення
- f7: Виконання дії

## Архітектура кіберфізичної системи

- **Фізичний рівень:** збір даних (мікрофони), фільтрація шумів на edge-пристрої.
- **Мережевий рівень:** передача структурованих ознак (MQTT/WebSockets).
- **Інтелектуальний рівень:** розпізнавання на базі CNN-LSTM моделі.
- **Виконавчий рівень:** інтерпретація команд та керування об'єктами.

Запропонований метод базується на ідеї послідовного перетворення аудіосигналу у текстове представлення шляхом використання єдиної нейромережевої архітектури, яка виконує функції як виділення ознак, так і їх інтерпретації

$$Y = f_{dec}(f_{LSTM}(f_{CNN}(f_{spec}(X)))),$$

де $f_{dec}$ – функція, що переводить сигнал у спектральну форму;

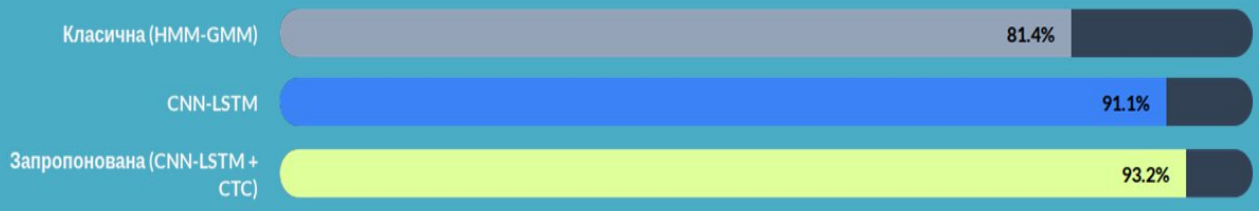
$f_{LSTM}$ – функція, яка відповідає за розпізнавання часового контексту мовлення;

$f_{CNN}$ – функція, яка відповідає за формування акустичних ознак;

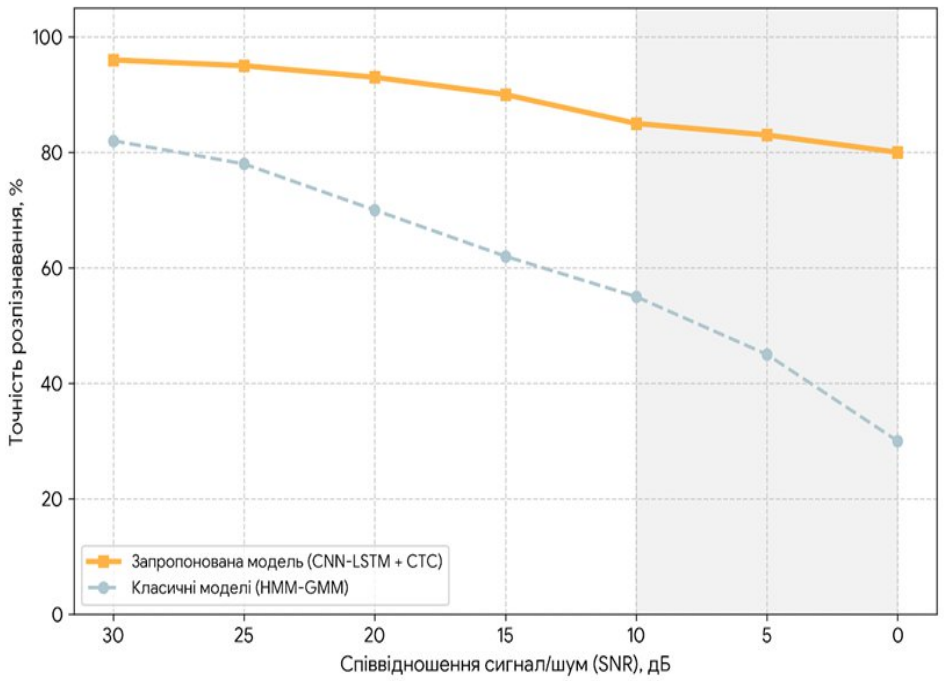
$f_{spec}$ – функція, яка перетворює результат у текст.




## Порівняння точності моделей




*Запропоноване рішення на 11.8% перевищує класичні методи розпізнавання команд.*




Залежність точності розпізнавання від рівня шуму

Підсумкова таблиця результатів експерименту

| Показник | Класична модель (HMM-GMM) | Запропонована модель (CNN-LSTM + CTC) |
|---|---|---|
| Середня точність (Accuracy) | 81.4% | 93.2% |
| Помилка (Word Error Rate) | 18.6% | 6.8% |
| Час розпізнавання 1 команди | ~120 мс | ~45 мс |
| Стійкість до шумів | Низька | Висока |

# Дякую за увагу!

**Протокол аналізу звіту подібності експертом**

Заявляю, що я ознайомився (-лась) з Повним звітом подібності, який був згенерований Системою виявлення і запобігання плагіату щодо роботи:

**Автор:** Дмитро КРУТИЙ

**Співавтор:**

**Назва:** Кіберфізична система розпізнавання голосу людини на базі алгоритмів машинного навчання

**Експерт:** Володимир ГРИГА

**Підрозділ:** Кафедра комп'ютерної інженерії та інформаційних систем

**Коефіцієнт подібності 1:**4.85%

**Коефіцієнт подібності 2:**1.6%

**Мікропробіли:** 9

**Заміна букв:** 2

**Інтервали:** 0

**Білі знаки:** 6

**Дата створення звіту:** 2026-04-29 16:43:51.0

**Після аналізу Звіту подібності констатую наступне:**

☒ **Запозичення, виявлені в роботі є законними і не є плагіатом. Рівень подібності не перевищує допустимої межі. Таким чином робота незалежна і приймається.**

☐ **Запозичення не є плагіатом, але перевищено граничне значення рівня подібностей. Таким чином робота повертається на доопрацювання.**

☐ **Виявлено запозичення і плагіат або навмисні текстові спотворення (маніпуляції), як передбачувані спроби укриття плагіату, які роблять роботу невідповідною вимогам законодавства (Ст. 32. ЗУ Про вищу освіту, пункт 3.1, Ст. 42. ЗУ Про освіту) та вимог НАЗЯВО (Критерій 5), а також кодексу етики і процедурам. Таким чином робота не приймається.**

Обгрунтування:

2026-04-29 Доцент Андрій Нічепорук

*Дата* експерт

# Anti-Plagiarism (http://ap.km.ua) v-15.701

**Максимальне співпадіння з одним документом 1.0%**

Словники перевірки: en_US, ru_RU, ua_UA. **Помилок в документах: 8%**

| | Документ | | Сумарний збіг по Базі Даних | |
|---|---|---|---|---|
| ID: 270804<br>Назва: МКР Кіберфізична система розпізнавання голосу людини на базі алгоритмів машинного навчання<br>Додано в БД: 2026-04-29<br>Автора: Дмитро КРУТИЙ<br>Керівники: Володимир ГРИГА<br>Консультанти:<br>Опоненти: | Символи | Лексеми | Символи | Лексеми |
| | 159878 | 1225 | 2439 (2%) | 32 (3%) |

Джерело плагіату

| ID | Опис | Наявність плагіату в документі | |
|---|---|---|---|
| | | Символи | Лексеми |

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

# РЕЦЕНЗІЯ НА КВАЛІФІКАЦІЙНУ РОБОТУ МАГІСТРА

Здобувач: Дмитро КРУТИЙ

**Тема:** Кіберфізична система розпізнавання голосу людини на базі алгоритмів машинного навчання

Спеціальність: 123 «Комп'ютерна інженерія»

Обсяг кваліфікаційної роботи магістра:

Кількість листів креслень – ; кількість сторінок записки 91

1. Короткий зміст роботи та прийнятих рішень У роботі запропоновано кіберфізичну систему розпізнавання голосу людини на базі алгоритмів машинного навчання

2. Висновок про відповідність роботи дипломному завданню Кваліфікаційна робота магістра відповідає виданому завданню

3. Характеристика виконання кожного розділу, ступінь використання останніх досягнень науки і техніки і передових методів роботи: В першому розділі проведено огляд існуючих комерційних та відкритих систем, що дозволило виявити їхню залежність від хмарної інфраструктури та вразливість до шумів. В другому розділі розроблено багаторівневу структурну модель КФС, яка забезпечує взаємодію між фізичним збором акустичних даних та інтелектуальними алгоритмами обробки. Процес розпізнавання формалізовано як композицію семи функціональних відображень, а також описано математичну модель сигналу та метод виділення ознак за допомогою мел-кепстральних коефіцієнтів. В третьому розділі виконано порівняльну класифікацію підходів, яка довела перевагу нейромережевих архітектур над класичними статистичними моделями HMM-GMM у складних акустичних умовах. Обґрунтовано та розроблено гібридний метод на базі CNN-LSTM з використанням функції втрат CTC, що дозволяє реалізувати наскрізне навчання без потреби у жорсткому вирівнюванні даних. Експериментальні дослідження на наборі даних Google Speech Commands підтвердили високу ефективність рішення: точність склала 93,2%, що на 11,8% вище за класичні методи.

4. Позитивні сторони роботи: Запропонована кіберфізична система

розпізнавання голосу людини на базі алгоритмів машинного навчання відкриває нові можливості для побудови розумних голосових інтерфейсів. Вона забезпечує ефективну взаємодію в системах промислового контролю, автоматизації робочих процесів та при управлінні розумними

5. Негативні сторони роботи: _____В роботі присутні певні логічні помилки щодо опису методів розпізнавання голосу на базі мовлення та методів розпізнавання (ідентифікації) голосу._____

6. Оцінка графічного оформлення та пояснювальної записки роботи: __=_____

_____

_____

_____

_____

7. Відгук про роботу в цілому: __В загальному робота виконана на достатньо___ високому професійному рівні._____

_____

_____

_____

_____

8. Інші зауваження: ___=_____

_____

_____

_____

_____

9. Оцінка кваліфікаційної роботи магістра:

Розглянувши позитивні та негативні сторони представленої кваліфікаційної роботи магістра вважаю, що робота заслуговує оцінки «добре» 75.00 (С)_____

Рецензент (прізвище, ім'я, по батькові, посада, місце роботи) ___

Мартинюк В. В., д. т. н., проф. професор кафедри автоматизації, комп'ютерно-інтегрованих технологій та робототехніки

"1" 05 2026р.

Зав. кафедри КІІС
д-р. філософії Ользі ПАВЛОВІЙ

Дмитро КРУТИЙ
ПІБ здобувача вищої освіти

ФІТ, 2 курсу, групи КІ2м-24-2

ЗАЯВА

З правилами чинного Положення про систему забезпечення академічної доброчесності у Хмельницькому національному університеті, згідно з яким виявлення академічного плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту і застосування заходів академічної відповідальності, ознайомлений (а). Про використання спеціалізованих програмних засобів (СПЗ) StrikePlagiarism та Anti-Plagiarism для перевірки кваліфікаційних робіт здобувачів вищої освіти на наявність академічного плагіату оповіщений (а). Надаю університету право на передачу моєї роботи для обробки та збереження в базах даних СПЗ і використання роботи для виявлення академічного плагіату в інших роботах, які перевіряються СПЗ.

Також надаю свою згоду на обробку й збереження університетом моєї роботи в Інституційному репозитарії Хмельницького національного університету.

Робота надається для перевірки в електронному варіанті. Електронна версія моєї роботи збігається (ідентична) з друкованою.

1 травня 2026 року

# РІШЕННЯ ЕКСПЕРТНОЇ КОМІСІЇ

## КАФЕДРИ КОМП'ЮТЕРНОЇ ІНЖЕНЕРІЇ ТА ІНФОРМАЦІЙНИХ СИСТЕМ ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

Назва кваліфікаційної роботи Кіберфізична система розпізнавання голосу людини на базі алгоритмів машинного навчання

Автор Дмитро КРУТИЙ

Освітня програма Комп'ютерна інженерія та програмування

Рівень вищої освіти другий (магістерський)

Спеціальність 123 Комп'ютерна інженерія

Науковий керівник: к.т.н., доцент Володимир ГРИГА

На основі аналізу кваліфікаційної роботи на дотримання вимог академічної доброчесності (у т.ч. відсутності ознак академічного плагіату) з урахуванням результатів перевірки роботи спеціалізованим програмним засобом(ами) комісія зробила такий висновок:

| № | Висновок | Позначка про відповідність |
|---|---|---|
| 1 | Ознаки академічного плагіату | |
| 1.1 | Запозичення, виявлені в роботі, є законними і не є академічним плагіатом (далі – зазначаються підстави віднесення запозичень до правомірних, якщо потрібно). Робота приймається до захисту. | відповідає |
| 1.2 | Виявлені запозичення не є академічним плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи (далі – зазначаються детальні та аргументовані підстави віднесення запозичень до правомірних). Робота приймається до захисту, але має бути відкоригована. | |
| 1.3 | Виявлені запозичення не є академічним плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. Робота може бути допущена до захисту після того як буде відкоригована та доопрацьована і успішно пройде повторну перевірку на академічний плагіат. | |
| 1.4 | Робота містить навмисні текстові спотворення, передбачувані спроби укриття текстових запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту. | |
| 2 | Інші види порушень академічної доброчесності | |

Підтвердження:

Запозичення, виявлені в роботі, є законними і не є плагіатом, оскільки:

1) усі запозичення фрагментарні, або мають належним чином оформленні посилання;

2) окремі виявлені збіги є загальновживаними фразами або виразами, про що свідчить посилання системи на збіг з джерелами на один фрагмент речення;

3) всі зафіксовані системою ознаки модифікації тексту відносяться до комбінування латинських символів зі україномовними скороченнями індексів в формулах, що не є модифікацією тексту.

4) значна частина знайденого плагіату відноситься до списку використаних джерел

Сумарний обсяг всіх запозичень, визначений системою виявлення збігів/ ідентичності/схожості StrikePlagiarism, складає 4.85 % і адресується до 52 першоджерела; та системою Anti-Plagiarism складає 1%, що, з урахуванням наведених обгрунтувань, відповідає характеру наукового дослідження і свідчить на користь кваліфікаційної роботи.

01.05.2026

Завідувач кафедри _____ Ольга ПАВЛОВА
Підпис Ім'я, ПРІЗВИЩЕ

Гарант освітньої програми _____ Олег САВЕНКО
Підпис Ім'я, ПРІЗВИЩЕ

Керівник кваліфікаційної роботи _____ Володимир ГРИГА
Підпис Ім'я, ПРІЗВИЩЕ