Міністерство освіти і науки України
Університет митної справи та фінансів

Факультет інноваційних технологій
Кафедра комп'ютерних наук та інженерії програмного забезпечення

Кваліфікаційна робота магістра

на тему: «Аналіз сучасних методів розпізнавання мовлення»

Виконав: студент групи К23-2м

Спеціальність 122 Комп'ютерні науки

Федосенко В.С.
(прізвище та ініціали)

Керівник д.е.н., проф. Корнєєв М.В.
(науковий ступінь, вчене звання, прізвище та ініціали)

Рецензент Дніпровський державний технічний університет
(місце роботи)

в.о. завідувача кафедри програмного забезпечення систем
(посада)

к.т.н., доц. Жульковський О.О.
(науковий ступінь, вчене звання, прізвище та ініціали)

Дніпро – 2025

АНОТАЦІЯ

Федосенко В.С. Аналіз сучасних методів розпізнавання мовлення.

Дипломна робота на здобуття освітнього ступеня магістр за спеціальністю 122 «Комп'ютерні науки» – Університет митної справи та фінансів, Дніпро, 2025.

Магістерська робота присвячена аналізу сучасних методів розпізнавання мовлення, яке є однією з ключових технологій у галузі штучного інтелекту. Дослідження охоплює широкий спектр підходів, починаючи від традиційних статистичних моделей (приховані марковські моделі – HMM, моделі сумішей Гауса – GMM) до сучасних методів на основі глибокого навчання та трансформаторів.

У роботі детально проаналізовано еволюцію систем розпізнавання мовлення. Експериментальний аналіз показав, що впровадження наскрізних моделей, таких як CTC і трансформатори, дозволяє значно покращити роботу систем у реальних умовах. Вивчено вплив шуму, акцентів, фонових звуків та інших факторів на ефективність алгоритмів. Проведено порівняння різних моделей та визначено їхні сильні і слабкі сторони, що стало основою для розробки рекомендацій з покращення систем.

Робота також акцентує увагу на етичних аспектах, таких як забезпечення конфіденційності, уникнення упередженості, створення доступних рішень для мов із низьким рівнем ресурсів.

Магістерська робота є внеском у розвиток технологій розпізнавання мовлення, спрямованих на підвищення точності, адаптивності та інклюзивності сучасних інформаційних систем.

Ключові слова: розпізнавання мовлення, штучний інтелект, глибоке навчання, трансформатори, шумостійкість, багатомовність.

ABSTRACT

Fedosenko V.S. Analysis of modern methods of speech recognition.

Diploma thesis for the degree of Master's Degree in specialty 122 «Computer Science» – University of Customs and Finance, Dnipro, 2025.

The master's thesis is devoted to the analysis of modern methods of speech recognition, which is one of the key technologies in the field of artificial intelligence. The research covers a wide range of approaches, from traditional statistical models (Hidden Markov Models - HMM, Gaussian Mixture Models – GMM) to modern methods based on deep learning and transformers.

The paper analyzes in detail the evolution of speech recognition systems. Experimental analysis has shown that the introduction of end-to-end models, such as CTC and transformers, can significantly improve the performance of systems in real-world conditions. The influence of noise, accents, background sounds, and other factors on the efficiency of the algorithms is studied. Different models were compared and their strengths and weaknesses were identified, which became the basis for developing recommendations for improving the systems.

The work also focuses on ethical aspects, such as ensuring privacy, avoiding bias, and creating affordable solutions for low-resource languages.

The master's thesis is a contribution to the development of speech recognition technologies aimed at improving the accuracy, adaptability, and inclusiveness of modern information systems.

Keywords: speech recognition, artificial intelligence, deep learning, transformers, noise immunity, multilingualism.

ЗМІСТ

ВСТУП 5
РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ 8
1.1 Основні підходи до розпізнавання мовлення 8
1.2 Традиційні методи розпізнавання мовлення 11
1.3 Сучасні методи розпізнавання мовлення 15
1.4 Проблеми та обмеження сучасних систем розпізнавання мовлення 19
1.5 Аналіз сучасної літератури 24
1.6 Висновки до першого розділу 34
РОЗДІЛ 2. ДОСЛІДЖЕННЯ СУЧАСНИХ МЕТОДІВ РОЗПІЗНАВАННЯ МОВЛЕННЯ 36
2.1 Постановка експерименту 36
2.2 Вибір методів для аналізу ефективності розпізнавання 40
2.3 Дослідження впливу шуму та інших факторів на точність 44
2.4 Порівняльний аналіз традиційних і сучасних алгоритмів 49
2.5 Використання мови програмування Python 53
2.6 Висновки до другого розділу 56
РОЗДІЛ 3. ПРОГРАМНА РЕАЛІЗАЦІЯ 58
3.1 Мета та вимоги до розробки 58
3.2 Вибір інструментів і технологій 59
3.3 Архітектура програмного забезпечення 62
3.4 Загальний опис роботи системи розпізнавання мовлення 66
3.5 Тестування системи розпізнавання мовлення 68
3.6 Висновки до третього розділу 71
ВИСНОВКИ 72
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 74
ДОДАТКИ 77

ВСТУП

Сучасний світ характеризується стрімким розвитком інформаційних технологій, що кардинально змінюють взаємодію людей з комп'ютерами та іншими пристроями. Однією з найбільш перспективних областей є технології розпізнавання мовлення, які знаходять застосування у найрізноманітніших сферах – від інтерфейсів користувача до медичних і освітніх рішень [1]. Завдяки їм користувачі отримують можливість більш природної взаємодії з машинами, що значно полегшує повсякденне життя. Однак, попри значні досягнення в цій галузі, існує низка проблем, які потребують вирішення: обмеження точності розпізнавання у складних акустичних умовах, обробка багатомовних даних, адаптація до індивідуальних особливостей голосу користувачів тощо.

Актуальність даного дослідження обумовлена зростаючою потребою в ефективних і надійних методах розпізнавання мовлення, які б відповідали сучасним вимогам до точності, швидкодії та універсальності. Технології на базі розпізнавання мовлення використовуються в таких критично важливих системах, як автоматичні перекладачі, голосові помічники, системи управління в автомобілях, медичні діагностичні системи та засоби для людей з обмеженими можливостями. Вдосконалення цих методів відкриває нові можливості для соціального й економічного розвитку, роблячи сучасні технології доступними для ширшого кола користувачів.

Метою кваліфікаційної роботи є проведення аналізу сучасних методів розпізнавання мовлення, визначення їхніх переваг та недоліків, а також формування рекомендацій для їх подальшого вдосконалення.

Для досягнення поставленої мети необхідно вирішити такі завдання:

– провести огляд сучасних підходів до розпізнавання мовлення, включаючи традиційні методи та алгоритми, засновані на штучному інтелекті;

– вивчити основні проблеми, з якими стикаються сучасні системи розпізнавання мовлення;

– проаналізувати результати застосування різних методів у практичних сценаріях;

– розробити рекомендації щодо вдосконалення існуючих систем розпізнавання мовлення;

– оцінити перспективи використання новітніх технологій у цій галузі.

Об’єктом дослідження є сучасні технології та алгоритми розпізнавання мовлення.

Предметом дослідження є методи та підходи до обробки та аналізу мовленнєвих даних.

У ході виконання роботи використовувались наступні методи дослідження:

– аналіз літературних джерел та наукових статей з метою вивчення сучасного стану технологій розпізнавання мовлення;

– порівняння ефективності різних алгоритмів на основі результатів емпіричних досліджень;

– експериментальний аналіз роботи існуючих систем розпізнавання мовлення;

– методи статистичного аналізу для оцінки точності та швидкодії алгоритмів.

Наукова новизна роботи полягає у систематизації сучасних методів розпізнавання мовлення, виявленні основних напрямів їх вдосконалення, а також у формуванні рекомендацій для підвищення ефективності цих технологій в умовах реального використання. Зокрема, запропоновано нові підходи до адаптації алгоритмів розпізнавання для роботи в умовах високого рівня шуму та багатомовного середовища.

Практична значимість роботи полягає у можливості використання результатів дослідження для створення більш ефективних систем

розпізнавання мовлення, які знайдуть застосування у різних галузях, таких як освіта, медицина, транспорт та побутові технології. Рекомендації, розроблені в рамках цієї роботи, можуть бути використані для вдосконалення існуючих програмних та апаратних рішень.

Кваліфікаційна робота сприяє вирішенню актуальних проблем у галузі технологій розпізнавання мовлення та робить внесок у розвиток сучасних інформаційних систем.

Структура магістерської роботи: робота складається з трьох розділів, об’єм роботи – 85 сторінок, робота містить 17 рисунків, перелік використаних джерел має 17 посилань.

РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

1.1 Основні підходи до розпізнавання мовлення

Розпізнавання мовлення – це інтегрована сфера штучного інтелекту та обробки сигналів, яка зазнала величезної еволюції за останні кілька десятиліть. Цей процес передбачає переклад усної мови в текст [1]. Завдання має характерну складність, оскільки потрібно розуміти як людську мову, так і акустичні властивості мови.

У розробці системи для розпізнавання мовлення застосовуються різні підходи. Більшість із них базуються на принципово інших принципах і технологіях для подолання викликів, властивих цьому завданню.

Розпізнавання мовлення спочатку почалося з примітивних систем, які залежали від інтенсивного використання простих методів зіставлення шаблонів. Перші спроби розпізнавання мовлення залежали від низьких обчислювальних ресурсів і елементарних алгоритмічних структур. Ці системи могли ідентифікувати лише обмежений набір слів і вимагали, щоб слова вимовлялися дуже точно. Зі збільшенням обчислювальної потужності та глибини теоретичного розуміння галузь поступово просувалась у бік просунутих імовірнісних і статистичних моделей, зокрема HMM [1, 2].

Здатність ефективно моделювати послідовні дані зробила HMM головною серед систем розпізнавання мовлення наприкінці XX століття. Представляючи мовлення як послідовність фонем і використовуючи ймовірність переходу між цими одиницями, HMM забезпечили надійний спосіб обробки мінливості, яка характеризує розмовну мову. Однак їхня залежність від марковських припущень і GMM для акустичного моделювання представляє кілька обмежень для фіксації складних часових залежностей і нелінійностей.

Прорив нейронних мереж, зокрема DMM, привніс можливість ієрархічного представлення даних в акустичне моделювання. Ці моделі чудово вловлюють складні зв’язки в мовних сигналах, перевершуючи традиційні GMM як з точки зору точності, так і надійності. Інтеграція їх у фреймворки на основі HMM стала важливим етапом, що поєднала сильні сторони ймовірнісного моделювання послідовності з експресивною силою нейронних мереж. Цей гібридний підхід був де-факто стандартизований на початку 2010-х років і дозволив точніше розпізнавати мовлення в різноманітних середовищах і серед мовців.

Наступним прогресом у розпізнаванні мовлення стали наскрізні нейронні архітектури. Найбільш репрезентативною з цих архітектур є так звана структура часової класифікації Connectionist Temporal Classification framework, яка безпосередньо відображає функції вхідного мовлення на вихідні текстові послідовності. На практиці це не вимагає попереднього сегментованого введення та добре узгоджується з природною варіативністю мовлення. Але найчастіше моделі, засновані на CTC, не впоралися з довгостроковою залежністю або моделюванням мови. У моделях кодеру-декору використовується кодер, який обробляє акустичні характеристики в представлення високого рівня, і декодер для виведення тексту, керований механізмом уваги, який динамічно фокусується на відповідних частинах введення.

Паралельно з розвитком послідовного навчання були розроблені моделі на основі трансформаторів, які підняли планку розпізнавання мови ще вище. Це природним чином позиціонує великі моделі трансформаторів як особливо придатні для глобальної контекстуалізації, і вони досить добре показали як акустичне, так і мовне моделювання [2]. Варіанти включають Transformer Transducer, а також BERT та його похідні, які були адаптовані до мови, досягаючи найсучаснішого статусу за різними тестами. Ці моделі були

попередньо навчені на великій кількості даних і, таким чином, добре узагальнювалися в різних акцентах, діалектах і шумних умовах.

Окрім розробок нейронних мереж, системи розпізнавання мовлення все частіше використовують деякі допоміжні методи для підвищення їх продуктивності. Нормалізація динаміків і вбудовування динаміків є одними з методів адаптації динаміків, які намагаються впоратися зі змінністю між динаміками, налаштовуючи моделі на конкретні динаміки. Шумостійкі функції та стратегії збільшення даних покращують стійкість до спотворень навколишнього середовища. Крім того, напівсупервізовані та неконтрольовані методики навчання привертають більше уваги зі зростаючим мотивом використання рясного некоментованого мовлення. Зазвичай вони передбачають самоконтрольовану попередню підготовку, що дозволяє моделям вловлювати загальні акустичні та лінгвістичні характеристики, які згодом можуть бути використані для конкретних завдань [1-3].

Інтеграція лінгвістичних знань також залишається важливою частиною розпізнавання мовлення. Явні, наприклад n-грами, чи неявні, такі як нейронні мовні моделі, мовні моделі є дуже важливим елементом у передбаченні правдоподібних послідовностей слів і вирішенні неоднозначностей. У наскрізних фреймворках включення зовнішніх мовних моделей або використання методів дрібного злиття може принести значні переваги. Ці підходи дуже добре узгоджуються з ієрархічною природою мови, дозволяючи системам розпізнавання збалансувати акустичне свідчення з лінгвістичною правдоподібністю.

Окрім основного завдання розпізнавання, системи розпізнавання мовлення все частіше вбудовуються в інтерактивні та мультимодальні програми. Наприклад, голосові помічники покладаються на модулі перетворення мови в текст як базові компоненти більших систем, які об’єднують розуміння природної мови, керування діалогом і синтез тексту в мову. У службах транскрипції розпізнавання мовлення в реальному часі

доповнюється алгоритмами постобробки для пунктуації, великих літер і форматування. Ці програми вказують на розробку систем, які є не тільки точними, але й ефективними та адаптованими до обмежень реального світу.

Незважаючи на перелічені успіхи, ще попереду стоять деякі проблеми з розпізнаванням мовлення, серед яких мови з низьким ресурсом. Вони завжди страждають від нестачі великих позначених наборів даних для навчання поточних моделей [3, 4]. Нещодавно було досягнуто багатообіцяючих рішень, дозволивши перенести навчання та міжмовні техніки для передачі знань з мов із високим рівнем ресурсу до мов із низьким ресурсом. Іншими проблемами є надійність у несприятливих акустичних умовах, таких як перекриття мови або середовища з високою реверберацією. Останні дослідження були зосереджені на використанні багатомікрофонних масивів і методів формування променя для покращення якості сигналу перед розпізнаванням.

Таким чином, особливої уваги потребують також етичні питання розпізнавання мовлення. Ключовими проблемами, які необхідно вирішити на шляху до відповідального впровадження цих технологій, є конфіденційність, упередженість і доступність. Механізми диференційованої конфіденційності, стратегії пом’якшення упередженості та принципи інклюзивного дизайну все частіше стають частиною процесу розробки як засоби узгодження систем розпізнавання мовлення з суспільними цінностями.

1.2 Традиційні методи розпізнавання мовлення

Традиційні методи розпізнавання мовлення були основою, на якій базуються сучасні розробки в розпізнаванні мовлення. Серед основних методів, які є найбільш значущими завдяки своїй математичній надійності та практичній ефективності, є приховані марковські моделі та моделі суміші Гауса. Ці ймовірнісні та статистично обґрунтовані методології забезпечують міцну структуру з точки зору моделювання послідовної та змінної природи

мовних сигналів, щоб дозволити системам фактично усвідомити, що розмовна мова може бути перекладена в текст із прийнятним ступенем точності [4, 5].

Загалом класифікацію методів розпізнавання мовлення наведено на рисунку 1.1.

Рисунок 1.1 – Класифікація методів розпізнавання мовлення

Сам ланцюг процесу розпізнавання мовлення наведено на рисунку 1.2.

Рисунок 1.2 – Ланцюг процесу розпізнавання мовлення

Приховані моделі Маркова – це один клас статистичних моделей, особливо придатних для даних часових рядів, таких як мова. Найважливішою інтуїцією, що лежить в основі HMM, є моделювання процесів із прихованими станами, які генерують спостережувані дані. Для розпізнавання мовлення прихованими станами є фонеми – найменші одиниці звуку, а спостережувані дані – це набір акустичних характеристик, витягнутих із мовних сигналів. Кожна фонема моделюється як послідовність станів, а переходи між станами контролюються імовірнісними правилами. Ця структура дозволяє HMM

фіксувати часові залежності в мовленні, враховуючи варіації в темпах мовлення та стилях артикуляції.

Потужність HMM випливає з їх марковського припущення, яке спрощує моделювання послідовностей, припускаючи, що майбутній стан залежить лише від поточного стану, а не від історії попередніх станів. Це припущення дозволяє проводити ефективні обчислення та, загалом, робить такі алгоритми, як алгоритм Вітербі, практичними для декодування послідовностей. Алгоритм Вітербі знаходить найбільш вірогідну послідовність станів, враховуючи послідовність спостереження, таким чином уможливлюючи узгодження акустичних сигналів із фонетичними представленнями [5]. Доповнюючи це імовірнісне декодування, алгоритм вперед-назад обчислює ймовірність спостережуваних послідовностей за даною моделлю. Разом ці методи забезпечують потужний інструментарій для моделювання послідовності.

Моделі суміші Гауса є природним доповненням до HMM, моделюючи акустичні характеристики, пов’язані з кожним прихованим станом. Мовні сигнали зазвичай представляють як послідовності векторів ознак, які фіксують інформацію про спектральні властивості, висоту та інші характеристики звуку. GMM моделюють розподіл ймовірностей цих векторів ознак як суміш розподілів Гауса, причому кожен компонент Гауса фіксує один режим мінливості даних. Гнучкість GMM у моделюванні складних розподілів дозволяє їм фіксувати різні акустичні моделі для різних фонем, акцентів і стилів мовлення.

Поєднання HMM і GMM дає дуже потужну гібридну структуру для розпізнавання мовлення. У цій парадигмі HMM вводять часову структуру, моделюючи часову еволюцію фонем, тоді як GMM моделюють акустичну мінливість кожної фонеми [6]. Це також забезпечується завдяки інтеграції, яка розпізнає безперервне мовлення, де спостережувані акустичні характеристики узгоджуються з найбільш імовірною послідовністю фонем. Такі системи традиційно навчені оцінювати параметри HMM і GMM, щоб максимізувати

вірогідність спостережуваних даних. Щоб впоратися з цим, використовуються такі методи, як алгоритм очікування-максимізації, який ітеративно оновлює параметри моделі, щоб найкраще відповідати навчальним даним.

Першим і головним серед існуючих обмежень є марковське припущення в HMM, що обмежує моделювання довгострокових залежностей і контекстної інформації в мовленні. Це обмеження стає помітним для таких складних завдань, як розпізнавання спонтанного мовлення або роботи з великим словниковим запасом. Подібним чином, лінійні комбінації компонентів Гауса в GMM погано підходять для моделювання дуже нелінійних і складних моделей, присутніх у мовних даних. Ці недоліки підкреслюють потребу в більш виразних моделях, здатних охопити повне багатство мови.

Дійсно, системи HMM-GMM давно користуються великою репутацією завдяки обчислювальній ефективності, що дозволяє розпізнавати мову в реальному часі навіть на скромному обладнанні. Іноді це відбувається за рахунок зниження гнучкості та точності порівняно з іншими, більш сучасними підходами [6, 7]. Наприклад, залежність від ручної роботи та апріорно визначених фонетичних одиниць може ускладнити адаптацію цих систем до нових мов чи діалектів. Крім того, їх продуктивність значно погіршується в шумних умовах або з колонками, акценти яких сильно відрізняються від тренувальних даних. Ці проблеми підкреслюють компроміси, з якими стикаються традиційні методи.

Теоретичні основи HMM і GMM також вплинули на загальну спільноту машинного навчання. Імовірнісні принципи, що лежать в основі цих моделей, такі як байєсівський висновок і оцінка максимальної правдоподібності, є основою багатьох методів сьогодні. Крім того, модульність систем HMM-GMM, де акустичне моделювання, моделювання послідовності та компоненти декодування настільки акуратні, надихнула на розробку сучасних архітектур. Саме ця модульність робить можливою інтеграцію лінгвістичних і

фонетичних знань, таким чином дозволяючи традиційним системам ефективно використовувати досвід предметної області [7].

Таким чином, прихована модель Маркова та модель суміші Гауса представляють базову основу розпізнавання мовлення. Вони заклали чітко визначену структуру для вирішення фундаментальних проблем мінливості та послідовної структури мовних сигналів. Їхні обмеження призвели до розгляду більш просунутих методів, але принципи та методи, розроблені для HMM та GMM, залишаються дуже необхідними в цій галузі сьогодні. Їхньою спадщиною є не лише створені ними системи, а й надані ними теоретичні ідеї, які формують траєкторію дослідження розпізнавання мовлення та його застосування.

1.3 Сучасні методи розпізнавання мовлення

Сучасні методи розпізнавання мовлення, з іншого боку, представляють радикальні зміни в останніх досягненнях щодо нейронних мереж, глибокого навчання та архітектури трансформатора. Ці системи значно перевершують деякі класичні підходи, такі як приховані моделі Маркова та моделі суміші Гауса, завдяки надзвичайно потужним структурам обчислень у поєднанні з методологіями, керованими даними [5-7]. Дійсно, ця продуктивність повністю заснована на використанні нейронних мереж під час моделювання складних нелінійних мовних зв’язків. Глибоке навчання запровадило значне розширення цих нейронних мереж, додавши набагато більше можливих архітектур, які можуть навчатися з ієрархічних представлень, тоді як трансформатори забезпечили парадигматичну зміну обробки послідовності завдяки механізмам самоуважності.

Нейронні мережі – це обчислювальна модель, натхненна структурою та діями людського мозку. Зазвичай вони складаються з кількох взаємопов’язаних шарів, що містять вузли або нейрони, з’єднані ваговими

зв’язками. Кожен нейрон приймає цей вхідний сигнал і застосовує до нього якусь функцію активації. Це стає його виходом, надаючи наступний шар. Нейронні мережі дуже добре навчаються за допомогою акустичних характеристик, які представляють спектральні та часові характеристики мовних сигналів. Ранні застосування нейронних мереж у розпізнаванні мовлення в основному стосувалися неглибоких архітектур, які, хоча й були дуже ефективними для конкретних завдань, насправді не мали можливості моделювати всю складність мовних даних [6].

Глибоке навчання ознаменувало зміну парадигми у застосуванні нейронних мереж для розпізнавання мови. DNN – це нейронна мережа, що складається з кількох шарів нейронів і має здатність вивчати ієрархічні представлення ознак. У той час як нижні рівні вловлюють низькорівневі характеристики, наприклад, грані або текстури в акустичному сигналі, вищі рівні абстрагують їх у складніші моделі, наприклад, фонетичні або лінгвістичні структури. Ця здатність вивчати представлення безпосередньо з необроблених або мінімально оброблених даних є однією з головних причин успіху глибокого навчання в розпізнаванні мовлення. Інші методи, які допомогли підвищити продуктивність і стабільність мереж DNN, включають активації ReLU, регулярізацію відключення та нормалізацію пакетів.

Одним із головних проривів у глибокому навчанні для розпізнавання мовлення стала розробка згорткових нейронних мереж. Хоча вони спочатку були розроблені для обробки зображень, CNN виявилися досить ефективними в мовленнєвих завданнях завдяки своїй здатності вловлювати локальні шаблони та ієрархічні відносини [7, 8]. Застосовуючи згорткові фільтри до спектрограм, які представляють візуальне зображення мовних сигналів, CNN можуть виділяти інваріантні характеристики на фоні невеликих коливань висоти чи швидкості. Ця стійкість до невеликих коливань висоти та швидкості зробила CNN популярним вибором для ідентифікації мовців і розпізнавання фонем.

Повторювані нейронні мережі також не залишилися позаду на шляху розпізнавання мовлення. На відміну від мереж прямого зв'язку, RNN містять з'єднання зворотного зв'язку, які надають RNN можливість обробки даних на основі послідовності. Це робить їх придатними для завдань часового ряду, таких як мова. Впровадження спеціалізованих варіантів, таких як мережі довготривалої короткочасної пам'яті (LSTM) і Gated Recurrent Units (GRUs), приборкало стандартну модель RNN, зафіксувавши довгострокові залежності, які важко моделювати. І LSTM, і GRU зрозуміли це, запропонувавши деякі структури воріт, які можуть контролювати потік інформації, що дозволяє їм запам'ятовувати минулі вхідні дані протягом тривалого часу та забувати нерелевантну інформацію. Це було важливо в таких програмах, як безперервне розпізнавання мовлення, де контекст відіграє важливу роль у розпізнаванні фонем і слів.

Традиційні RNN та їх варіанти, незважаючи на їхній успіх, мають певні проблеми, пов'язані з обчислювальною неефективністю та труднощами з розпаралелюванням навчання [8]. Трансформатори відмовляються від повторюваних зв'язків RNN і натомість покладаються на механізми самоконтролю, які обчислюють зв'язки між усіма елементами послідовності за один раз. Це дозволяє трансформаторам більш ефективно фіксувати довгострокові залежності, оскільки ваги уваги динамічно змінюють свій фокус на відповідні частини вхідної послідовності.

Механізм самоуважності в основі трансформаторних архітектур працює шляхом обчислення балів уваги, які вказують на важливість кожного елемента в послідовності по відношенню до інших елементів. Потім ці бали використовуються для виконання зважених комбінацій представлень вхідних даних і дозволяють моделі висвітлити основні характеристики. Трансформери використовують кілька концентраційних головок, кожна з яких вивчає різні особливості входу, і поєднують їх шарами, щоб створити глибокі та виразні моделі. Іншим важливим компонентом є позиційне кодування, яке дозволяє

трансформаторам осягати порядок елементів у послідовності та виконувати ще одну критичну вимогу до обробки мовних даних.

Ця галузь розвивалася завдяки розробці попередньо навчених моделей трансформаторів, у тому числі представлень двонаправленого кодера від Transformers та їх мовних адаптацій. Ці моделі вивчають уявлення загального призначення за допомогою широкомасштабного попереднього навчання на різноманітних наборах даних, які потім можна налаштувати для конкретних завдань [8, 9]. Попередньо підготовлені трансформатори вже адаптовано для розпізнавання мовлення для покращення продуктивності в мовах з низьким ресурсом, у шумному середовищі та розмовному мовленні. Такі архітектури, як Wav2Vec та його варіанти, успішно поєднують трансформатори з необробленим аудіовходом, позбавляючи потреби в ручних функціях і даючи найсучасніші результати.

Глибоке навчання та методи, засновані на трансформаторах, також виграли від розробок уздовж осі методів оптимізації та збільшення обчислювальних ресурсів. Використання варіантів стохастичного градієнтного спуску, таких як Adam і RMSProp, прискорило конвергенцію та покращило конвергенцію. Лише з розповсюдженням GPU та TPU навчання глибоких моделей на великомасштабних наборах даних забезпечило неперевершену продуктивність в історії. Крім того, кілька фреймворків демократизували такі інструменти для розробки та розгортання моделей нейронних мереж: TensorFlow і PyTorch – два з них.

Таким чином, різні програми залишили свій відбиток у розпізнаванні мови за допомогою сучасних методів. Віртуальні помічники, автоматизовані служби транскрипції та системи перекладу в реальному часі використовують глибоке навчання та трансформатори для забезпечення точної та надійної роботи [8-10]. Це дозволило зробити величезний крок у багатомовному розпізнаванні мовлення, коли моделі можуть обробляти та розпізнавати мовлення кількома мовами одночасно. Інтеграція з мовними моделями також

дозволила системам розпізнавання мовлення бути набагато більш контекстними та зв’язними у виведенні тексту.

Таким чином, система розпізнавання мовлення повинна впоратися з мінливістю, що виникає через акценти, діалекти та стилі мовлення, не кажучи про фоновий шум і перекриту мову. Ці проблеми викликали дослідження різних методів, таких як збільшення даних, адаптація домену та самоконтрольоване навчання. Щоб технології розпізнавання мовлення використовувалися рівноправно та відповідально, окрім етичних питань, необхідно розглянути питання конфіденційності та упередженості. Іншими словами, це означає, що перехід від традиційних методологій до сучасних, які включають нейронні мережі, глибоке навчання та трансформатори, означає зміну парадигми розпізнавання мовлення. Нові шляхи, які вони відкривають, включають навчання безпосередньо з даних, захоплення складних шаблонів вищого порядку та навіть обробку послідовностей з неперевершеним рівнем ефективності. З кожним кроком еволюції в поточних дослідженнях поєднання цих методів із майбутніми технологіями може вивести технології розпізнавання мовлення далеко вперед у продуктивності та їх застосуванні.

1.4 Проблеми та обмеження сучасних систем розпізнавання мовлення

Незважаючи на те, що сучасні системи розпізнавання мовлення досягли надзвичайного розвитку, проблем і обмежень все ще багато, що заважає їм досягти універсальної застосовності та надійності. Це питання, пов’язані з технічними, лінгвістичними, екологічними та етичними аспектами, що вказує на складність перекладу людської мови на точні та значущі цифрові інтерпретації. Оскільки ці системи пронизують різні сфери, розуміння їх обмежень є важливим для сприяння подальшим інноваціям щодо їхніх недоліків [11].

Схема, що наводить проблеми розпізнавання мовлення, наведена на рисунках 1.3-1.4.

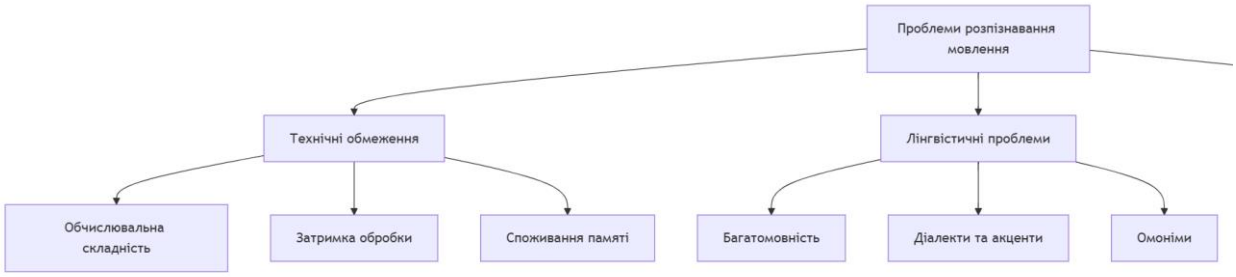

Рисунок 1.3 – Технічні обмеження та лінгвістичні проблеми розпізнавання мовлення

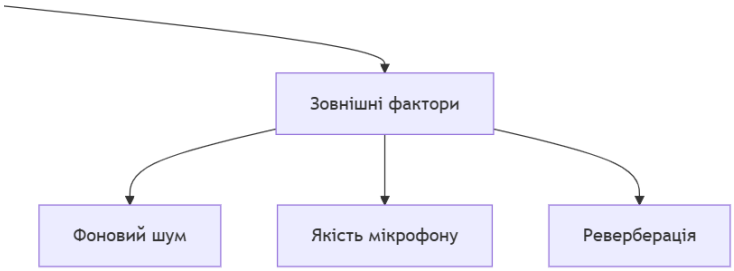

Рисунок 1.4 – Зовнішні фактори проблеми розпізнавання мовлення

Найголовнішою проблемою передового розпізнавання мовлення є його вразливість до факторів навколишнього середовища. Ефективність значно погіршується через фоновий шум, накладання мови та зміни акустичних умов. Моделі глибокого навчання досягли найсучаснішої шумостійкості завдяки доповненню даних і шумоадаптивним алгоритмам, але такі методи часто не можуть ефективно узагальнити в дуже мінливих середовищах реального світу. Наприклад, нездатність сигналів відокремити мову від навколишнього шуму зберігається в людному середовищі або коли людина бере участь у кількох розмовах одночасно. Тому для цього потрібні ще більш складні методи розділення сигналу та покращення, які працюють у динамічних і

непередбачуваних акустичних сценаріях [11, 12]. Ще одна важлива проблема – варіативність акценту, діалекту та стилю мовлення. Системи розпізнавання мовлення зазвичай мають упередження щодо даних, на яких вони навчалися, і тому працюють по-різному в різних мовних і культурних контекстах. Основні проблеми часто виникають через неносіїв мови, регіональні акценти та соціолекти, оскільки всі ці варіації відрізняються від стандартних мовних шаблонів, закодованих у навчальних наборах даних. Це упередження не тільки підриває інклюзивність технологій розпізнавання мовлення, але й обмежує їх застосування в багатомовних і мультикультурних умовах. Здебільшого це вирішується шляхом включення більшої та різноманітнішої колекції навчальних наборів даних, розробки мовно-агностичних систем та інших подібних функцій. Однак розробка та впровадження цих рішень пов’язані з дуже високими витратами на обчислення та використання ресурсів.

Контекстуальне розуміння є ще одним моментом, де системи розпізнавання дають збій у сучасні дні. У той час як нейронні мережі та трансформаторні архітектури розширили можливості моделювання послідовних залежностей і захоплення далекого контексту, ці системи зазвичай виходять з ладу через неоднозначність і нюанси мови. Омофони, ідіоматичні вирази та контекстно-залежні значення часто призводять до неправильного тлумачення [12]. Такі неоднозначності вирішуються вдосконаленими мовними моделями, які семантично набагато глибші та інтегровані із зовнішніми джерелами знань. Тим не менш, ще один рівень ускладнень у системах розпізнавання мовлення може бути доданий самими мовцями. Відмінності в характеристиках голосу, таких як висота, тон або швидкість мовлення, можуть вплинути на точність розпізнавання. Крім того, такі фізіологічні фактори, як вік, стать і стан здоров’я, вносять додаткову мінливість. У спробах подолати ці ефекти було розроблено низку методів адаптації та нормалізації мовця. Вони часто вимагали тонкого налаштування або додаткових обчислювальних витрат ціною масштабованості для програм

реального часу. Незалежні від динаміків моделі, хоч і більш універсальні, часто поступаються точністю, коли стикаються з дуже різними групами користувачів. Використання великомасштабних анотованих наборів даних для навчання є значним вузьким місцем у розробці систем розпізнавання мовлення. Оскільки анотація даних займає багато часу та багато коштує, це особливо збільшує розрив у продуктивності між менш представленими мовами та діалектами. Таким чином, нові підходи до самоконтролю та неконтрольованого навчання стали найпривабливішими альтернативами, які дозволяють моделям навчатися з необроблених, немаркованих аудіоданих. Однак ці методи все ще перебувають у початковому стані та стикаються з проблемами досягнення паритету з контрольованими підходами з точки зору точності чи надійності [13].

Вимоги до обробки в реальному часі ще більше ускладнюють розгортання систем розпізнавання мовлення. Такі програми, як віртуальні помічники, автоматичні служби транскрипції та живий переклад, потребують низької затримки та високої пропускної здатності. Збалансування цих вимог з обчислювальною складністю моделей глибокого навчання представляє величезну проблему. Хоча апаратні прискорювачі, такі як GPU і TPU, пом’якшили деякі з цих обмежень, енергоефективність і масштабованість залишаються критичними проблемами, особливо в середовищах з обмеженими ресурсами, таких як мобільні пристрої та периферійні обчислювальні платформи.

Ще один важливий вимір обмежень, які оточують сучасні системи розпізнавання мовлення, пов’язаний із міркуваннями конфіденційності та безпеки. Збір і обробка аудіофайлів викликає низку проблем, пов’язаних із згодою користувача, правом власності на дані та потенційним зловживанням. Такі ризики посилюються в архітектурах централізованої обробки, де аудіодані надсилаються на хмарні сервери, створюючи можливості для несанкціонованого доступу та порушень [14]. Такі методи, як інтегроване

навчання та обробка на пристрої, спрямовані на вирішення цих проблем шляхом мінімізації передачі даних і підвищення конфіденційності користувачів, однак ці підходи часто обходяться більшими обчислювальними вимогами та зниженою продуктивністю моделі. Крім того, існують етичні наслідки систем розпізнавання мовлення, оскільки ці системи самі є упередженими.

Іншою важливою проблемою є можливість інтерпретації та пояснення моделей глибокого навчання. Важко зрозуміти процеси прийняття рішень, що лежать в основі систем розпізнавання мовлення, через природу нейронних мереж і трансформаторів як «чорний ящик». Відсутність прозорості заважає довірі та підзвітності, особливо в програмах із високими ставками, де наслідки помилок є серйозними. Зусилля щодо покращення інтерпретації, наприклад візуалізації уваги та дистиляції моделі, тривають, але залишаються недостатніми для забезпечення повного розуміння поведінки моделі.

Це призводить до остаточної різниці – економічного та інфраструктурного доступу до технологій розпізнавання мовлення – підкреслюючи обмеженість цих технологій у всьому світі. У той час як розвинені регіони користуються передовими системами в поєднанні з підключенням, регіони з недостатнім ресурсом не мають інфраструктури та обчислювальних ресурсів, необхідних для розгортання цих технологій [14]. Такі зусилля мають бути узгоджені шляхом розробки полегшених моделей, які є економічно ефективними та ефективно працюють в умовах обмежених ресурсів.

Таким чином, проблеми та обмеження сучасних систем розпізнавання мовлення дають натяк на складність та багатогранність завдання. Виклики величезні, починаючи від екологічних факторів і закінчуючи мовним розмаїттям, контекстуальною неоднозначністю та мінливістю мовців, які вимагають інноваційних рішень. Питання залежності від даних, обробки в режимі реального часу, конфіденційності, етики та адаптивності до конкретної

області також визначають напрямок продовження досліджень і розробок. Подолання цих викликів вимагатиме скоординованих міждисциплінарних зусиль із використанням останніх досліджень у галузі машинного навчання, лінгвістики, обробки сигналів та етики для розробки більш надійних, справедливих і надійних систем розпізнавання мовлення.

1.5 Аналіз сучасної літератури

Стаття [1] зосереджена на розробці автоматичної системи розпізнавання мовлення для англійських лекцій, яка також займається підсумовуванням змісту та надає японські субтитри. Повне субтитрування аудіо англійської лекції може ускладнити розуміння і знизити зручність сприймання, тому необхідна система підсумовування. Використовуючи систему розпізнавання мовлення на основі DNN-HMM, дослідники досягли точності розпізнавання на рівні 88% для лекцій TED. Результати перекладу мовлення показали менший BLEU-скор, близько 14%, порівняно з перекладом тексту. Водночас система підсумовування виявила свою стійкість до помилок розпізнавання мовлення, оскільки витягнуті важливі речення були майже ідентичні тим, що були б отримані в процесі підсумовування тексту.

Стаття [2] зосереджена на вирішенні проблеми розпізнавання мовлення у пацієнтів з дисартрією за допомогою технологій автоматичного розпізнавання мовлення. Для цього були зібрані мовні дані чотирьох пацієнтів з дисартрією та чотирьох здорових спікерів для цифрових команд. Було проведено кілька експериментів з розпізнавання мовлення за допомогою двох попередньо навчених моделей, Aishell-2 та Wenetspeech, на основі відкритої архітектури WeNet, а також чотирьох основних комерційних систем розпізнавання API. У результаті було запропоновано метод оптимізації даних для дисартричних пацієнтів за допомогою інструменту so-vits-svc. Експерименти показали, що хоча сучасні моделі розпізнавання мовлення

добре працюють для мовлення загальної популяції, є ще простір для покращення результатів для мовлення людей з дисартрією. Метод оптимізації, запропонований у статті, ефективно покращує якість мовних даних, і для двох згаданих відкритих моделей розпізнавання рівень точності збільшився на 60% і 67,64% порівняно з початковими даними.

Стаття [3] досліджує метод розпізнавання емоцій в мовленні, який використовує як акустичні, так і лінгвістичні ознаки. Багато існуючих методів емоційного розпізнавання застосовують лінгвістичні ознаки, однак часто для цього використовуються довідкові транскрипти, оскільки розпізнавання емоційного мовлення вважається складнішим, ніж неемоційного. Акустичні ознаки емоційного мовлення відрізняються від неемоційного, і ці ознаки значно варіюються залежно від типу та інтенсивності емоцій. У статті розглядається новий метод розпізнавання емоційного мовлення, який поєднує адаптацію акустичної моделі та мовної моделі, що дозволяє досягти високої ефективності в розпізнаванні емоційного мовлення. Автори намагаються видобути лінгвістичні ознаки за допомогою результатів розпізнавання мовлення, досягаючи точності розпізнавання слів на рівні 82,2%. Попри наявність помилок у розпізнаванні, лінгвістичні ознаки, отримані з результатів розпізнавання, виявляються корисними, і показано, що поєднання лінгвістичних та акустичних ознак є ефективним для розпізнавання емоцій.

У статті [4] пропонується новий метод для виявлення енергетичних модулювань з різних частотних смуг у мовленні, який використовує фреймові вектори ознак, що отримали назву модульовані вектори (M-вектори), для застосування в системах автоматичного розпізнавання мовлення (ASR). Показано, що в різних багатоканальних налаштуваннях, де паралельно використовуються M-вектори та популярні мел-частотні кепстральні коефіцієнти (MFCC), можна досягти підвищення точності розпізнавання слів в кінцевих системах на ≈ 5%, а для системи HMM-GMM з моногенеричними

та тригенераційними моделями точність зростає на ≈ 18% і ≈ 16% відповідно порівняно з використанням традиційних MFCC ознак.

Стаття [5] присвячена виявленню неналежних пауз у мовленні пацієнтів з дисартрією, що є поширеною проблемою серед пацієнтів після інсульту, і значно впливає на зрозумілість мови. Неналежні паузи є важливими індикаторами для оцінки тяжкості порушень та в реабілітаційній терапії. У статті пропонується розширити велику модель розпізнавання мовлення для виявлення таких пауз у мовленні дисартричних пацієнтів. Для цього було розроблено стратегію позначення пауз, концепцію завдання та модель розпізнавання мовлення з додатковим шаром для прогнозування неналежних пауз. Спочатку виявлення пауз розглядається як завдання розпізнавання мовлення, використовуючи модель автоматичного розпізнавання мовлення (ASR) для перетворення мовлення в текст з мітками пауз. Паузи позначаються на рівні тексту відповідно до їхньої доцільності, а критерії позначення були розроблені у співпраці з логопедами для забезпечення високоякісних анотованих даних. Нарешті, було розширено модель ASR додатковим шаром прогнозування неналежних пауз для кінцевого виявлення таких пауз. Для оцінки ефективності запропонованого методу була розроблена спеціальна метрика, яка дозволяє оцінювати виявлення пауз незалежно від загальної точності ASR. Експерименти показали, що запропонований метод дозволяє краще виявляти неналежні паузи в мовленні дисартричних пацієнтів порівняно з базовими методами, досягаючи рівня помилок неналежних пауз на рівні 14,47%.

Стаття [6] пропонує метод прогресивного навчання для адаптивної оцінки шуму та мовлення (PL-ANSE) в процесі попередньої обробки мовлення для розпізнавання мовлення в умовах шуму. Метод поєднує здатність відстежувати шум на рівні фреймів за допомогою покращеного мінімального рекурсивного середнього (IMCRA) та глибоке прогресивне навчання взаємодій між мовленням і шумом на рівні висловлювань. Спочатку для

навчання прогресивних масок відносного співвідношення сигнал/шум (PRM) використовують двонаправлену модель довготривалої пам’яті (LSTM) на кожному шарі мережі, при цьому співвідношення сигнал/шум поступово зростає. Потім, оцінені PRM на рівні висловлювань комбінуються в рамках традиційного алгоритму підвищення якості мовлення для покращення мовлення на рівні фреймів. У підсумку, покращене мовлення, що базується на багаторівневому злитті інформації, передається безпосередньо в систему розпізнавання мовлення для покращення її ефективності. Експерименти показали, що запропонований підхід дозволяє досягти зниження помилки в розпізнаванні слів (WER) на 22,1% порівняно з результатами, отриманими на не обробленому шумному мовленні (з 23,84% до 18,57%) на реальних даних CHiME-4 з одним каналом.

Стаття [7] зосереджена на використанні технології розпізнавання мовлення в логопедичній терапії для дітей з порушеннями вимови. В роботі розроблено просту веб-гру, яка сприяє покращенню вимови фонем через дитячі вірші. У грі гравець рухає свою фішку по ігровому полю після того, як правильно вимовить вірш. Вимова записується та розпізнається системою, яка визначає рівень відповідності з оригіналом, і цей бал перетворюється на кількість клітинок, на які можна рухатися. Для покращення розпізнавання дитячого мовлення була навчена модель на корпусі дитячої мови, оскільки традиційні системи оптимізовані для дорослих. Нова модель значно підвищила точність розпізнавання мовлення. Гра також збільшує мотивацію дітей до домашніх занять завдяки своїм ігровим елементам та зручному інтерфейсу.

Стаття [8] присвячена розробці автоматичної системи розпізнавання мовлення (ASR) для людей з дисартрією, порушенням, яке значно впливає на зрозумілість мовлення через параліч м’язів і органів, що беруть участь у процесі артикуляції. Оскільки дисартрія часто супроводжується фізичними обмеженнями, такі люди стикаються не лише з проблемами в комунікації, але

й з труднощами у взаємодії з цифровими пристроями. Для них технології ASR можуть стати важливим інструментом, що дозволяє їм спілкуватися з іншими людьми та комп'ютерами. Проте традиційні ASR системи погано справляються з розпізнаванням дисартричного мовлення, особливо при тяжких формах дисартрії, через кілька основних проблем: неточність дисартричних фонем, обмежену кількість даних для тренування та неточність позначення фонем. У статті представлено нову систему ASR для дисартрії під назвою Speech Vision (SV), яка вирішує ці проблеми за допомогою новаторського підходу, де мовні ознаки витягуються візуально, і система вчиться «бачити» форми слів, вимовлених людьми з дисартрією. Такий підхід дозволяє уникнути проблем, пов'язаних з фонемами. Щоб подолати проблему обмежених даних, система SV використовує техніки візуальної аугментації даних, генерує синтетичні акустичні візуальні дані для дисартрії та використовує трансферне навчання. У порівнянні з іншими сучасними системами ASR для дисартрії, SV показала кращі результати, покращивши точність розпізнавання на 67% серед користувачів UA-Speech, з найбільшими покращеннями для важких форм дисартрії.

Стаття [9] розглядає використання технологій автоматичного розпізнавання мовлення (ASR) для осіб з дисартрією – порушенням мовлення, що впливає на артикуляційні м'язи і призводить до незрозумілого мовлення. Хоча сучасні ASR системи досягли високих результатів для здорового мовлення, їх ефективність для дисартрії залишається низькою, особливо для пацієнтів з тяжкими формами дисартрії, які найбільше потребують цієї технології. У цій роботі автори заповнюють цей розрив, пропонуючи нову систему Dysarthric Speech Transformer, яка використовує спеціалізовану глибоку архітектуру трансформера. Для вирішення проблеми обмеженості даних вони розробили двофазний процес трансферного навчання, який використовує дані здорового мовлення, досліджували налаштування заморожування нейронів і застосовували аугментацію аудіо даних. Загалом у

дослідженні було навчено 45 адаптованих до спікера систем ASR для дисартрії. Результати показали ефективність використання трансферного навчання та аугментації даних, підкреслюючи важливість більш глибоких архітектур трансформерів. Запропонована система ASR перевершила сучасні методи, досягнувши кращої точності розпізнавання для 73% учасників дослідження, з покращеннями до 23%.

Стаття [10] зосереджена на використанні великих мовних моделей (LLM) для різних задач, пов’язаних з мовленням, таких як розпізнавання мовлення, перетворення тексту в мовлення та розуміння усної мови. Останнім часом звертається увага на використання дискретизованих мовних ознак як ефективної та сумісної альтернативи безперервним ознакам для LLM. Це пов’язано з меншими вимогами до зберігання та кращою узгодженістю цих ознак з простором вводу LLM. Проте типова практика заморожування енкодера мовлення під час навчання створює труднощі в подоланні розриву між модальностями мовлення та тексту. Для вирішення цієї проблеми запропоновано використання шару змішаного масштабування ретокенізації, що інтегрує декілька рівнів дискретизованих мовних ознак безпосередньо в модуль вводу LLM. Експериментальні результати показали, що запропонований метод ефективно підвищує ефективність системи автоматичного розпізнавання мовлення (ASR) у контексті безперервного навчання LLM, підкреслюючи важливість ретельно розробленого модуля вводу для інтеграції дискретизованих мовних ознак в LLM.

В останні роки спільне навчання фронтальної частини для покращення мовлення та задньої частини для автоматичного розпізнавання мовлення (ASR) стало широко використовуватися для підвищення надійності ASR систем. Традиційні методи спільного навчання використовують тільки покращене мовлення як вхід для бекенду. Однак, через різноманіття типів шуму з різними інтенсивностями, системам покращення мовлення важко безпосередньо відокремити мовлення від вхідних даних. Крім того, в

покращеному мовленні часто спостерігаються спотворення мовлення та залишковий шум, причому спотворення мовлення та шуму відрізняються. Більшість існуючих методів фокусується на об’єднанні покращених і шумових ознак для вирішення цієї проблеми. У статті [11] пропонується подвійний потік мережі для уточнення спектрограм, яка одночасно уточнює мовлення та шум і розділяє шум від шумного вхідного сигналу. Запропонований метод демонструє кращу ефективність з відносним зниженням коефіцієнта помилок розпізнавання слів (CER) на 8,6%.

У статті [12] пропонується новий багаторівневий показник спотворення (MLDM), який використовується як ціль для оптимізації систем покращення мовлення на основі глибоких нейронних мереж в аудіо та аудіо-візуальних сценаріях. Метою є одночасне поліпшення якості мовлення, зрозумілості та зменшення кількості помилок розпізнавання. Крім того, проведено всебічний кореляційний аналіз, який показав високу кореляцію між трьома оцінками: якістю мовлення, зрозумілістю та результатами розпізнавання, з трьома загальними цілями оптимізації – середньоквадратичною помилкою між ідеальним та оціненим масками, шкалою, інваріантною до сигнал-шуму, та мірою на основі крос-ентропії. Для покращення ефективності запропоновано ще один показник спотворення – корельований багаторівневий показник (C-MLDM), який є ваговою комбінацією MLDM та середнього кореляційного показника, заснованого на трьох Pearson-коефіцієнтах кореляції. Експериментальні результати на корпусі TCD-TIMIT, що містить додатковий шум, показали, що MLDM перевершує системи, оптимізовані з кожною з цих цілей, як в аудіо, так і в аудіо-візуальних сценаріях, покращуючи всі три показники. C-MLDM також стабільно перевершує MLDM у всіх тестах. Останнє підтверджує загальну здатність обох методів працювати на різноманітних наборах даних, архітектурах моделей SE та мовних умовах.

Стаття [13] присвячена вдосконаленню автоматичного розпізнавання мовлення (ASR) з використанням додаткової текстової інформації з відео-

слайдів, що синхронізовані з мовленням в онлайн-конференціях та курсах. Автори пропонують нову модель, що використовує довгий контекст інформації з слайдів, зокрема розробляють мережу для аудіо-візуального розпізнавання мовлення (AVSR), яка ефективно інтегрує цей контекст. Для цього вони використовують архітектуру з двома енкодерами для одночасного моделювання аудіо та довгого контексту. Також запропоновано модуль прогнозування упереджених фраз, який використовує бінарну крос-ентропію для виявлення упереджених фраз. Додатково введено динамічну симуляцію контекстних фраз, щоб покращити узагальненість і стійкість мережі LCB-net. Експерименти на великомасштабному аудіо-візуальному корпусі SlideSpeech показали, що запропонована модель перевершує загальну модель ASR на 9.4%/9.1%/10.9% у зниженні WER/U-WER/B-WER на тестовому наборі. Окрім того, модель була оцінена на корпусі LibriSpeech, де досягнуто покращень у зниженні WER/U-WER/B-WER на 23.8%/19.2%/35.4% відповідно порівняно з ASR.

У статті [14] запропоновано нову архітектуру нейронної мережі для розпізнавання мовлення в умовах кількох джерел звуку з багатоканальним входом і виходом – MIMO-Speech, яка розширює підхід послідовність-до-послідовності (seq2seq) для ефективної обробки багатоканальних сигналів і їх розпізнавання. Система повністю нейронна та оптимізується виключно за допомогою критерію автоматичного розпізнавання мовлення (ASR). Вона складається з монозвукової маскуючої мережі, нейронного багатоджерельного підсилювача та багатовихідної моделі для розпізнавання мовлення. Цей підхід дозволяє безпосередньо перетворювати накладене мовлення на текстові послідовності. Для покращення результатів була використана стратегія навчання за планом, що ефективно використовує тренувальний набір даних. Експерименти на корпусі wsj1-2mix показали, що модель досягає зменшення кількості помилок на 60% порівняно з системою для одного каналу, з високоякісними поліпшеними сигналами.

У статті [15] представлено WenetSpeech – багатодоменною китайською мовною корпорацією, яка містить понад 10 000 годин якісно маркованого мовлення, понад 2400 годин слабо маркованого мовлення та близько 10 000 годин неструктурованих даних, загалом понад 22 400 годин. Дані були зібрані з YouTube та подкастів і охоплюють різноманітні стилі мовлення, сценарії, теми та умови шуму. Для обробки даних YouTube використано метод оптичного розпізнавання символів (OCR) для генерації кандидатів аудіо/текстових сегментацій на основі відповідних субтитрів, а для подкастів застосовано високоякісну систему автоматичного розпізнавання мовлення (ASR) для створення пар аудіо/тексту. Далі була запропонована нова методика виявлення помилок маркування для подальшої перевірки і фільтрації кандидатів. Також до WenetSpeech включено три ручні тестові набори для оцінювання: Dev – для крос-валідації під час навчання, Test_Net – для тестування на даних з Інтернету, і Test_Meeting – для більш складного тестування на даних із реальних зустрічей. Для трьох популярних систем розпізнавання мовлення – Kaldi, ESPnet та WeNet – були надані базові системи для тренування на WenetSpeech та результати розпізнавання для цих тестових наборів. WenetSpeech є найбільшою відкритою китайською мовною корпорацією з транскрипціями і має велике значення для досліджень у сфері розпізнавання мовлення на виробничому рівні.

У статті [16] розглядається проблема розпізнавання мовлення для осіб з дизартрією – моторним розладом мовлення, що характеризується зниженою розбірливістю через порушення координації м’язів, що відповідають за вироблення мови. Системи автоматичного розпізнавання мовлення (ASR) можуть допомогти таким людям спілкуватися більш ефективно. Однак для створення надійної ASR системи, орієнтованої на дизартрію, необхідні великі обсяги навчальних даних, яких не вистачає. Останні досягнення в області синтезу мовлення на основі технології Text-To-Speech (TTS) дозволяють використовувати синтезоване мовлення для розширення навчальних наборів.

У цій роботі запропоновано поліпшити багатокористувацькі кінцеві TTS системи для синтезу дизартрійного мовлення, що допоможе поліпшити навчання спеціалізованої ASR для дизартрії. У синтезованому мовленні додаються параметри, такі як рівень тяжкості дизартрії та механізми вставки пауз, поряд з іншими контролюючими параметрами, такими як висота тону, енергія та тривалість. Результати показали, що модель DNN-HMM, тренована на додаткових синтезованих даних з дизартрією, досягає покращення WER на 12,2% порівняно з базовою моделлю, а додавання контролів рівня тяжкості та вставки пауз знижує WER на 6,5%, що демонструє ефективність цих параметрів.

У статті [17] розглядається використання музичного мовлення в реабілітації пацієнтів з мовними порушеннями, що полягає у виробленні мовлення за допомогою простих музичних (ритмічних чи мелодійних) патернів для полегшення обробки мовлення пацієнтами. Автори досліджували ефективність сучасних алгоритмів автоматичного розпізнавання мовлення (ASR) при розпізнаванні нормального та музичного мовлення. З початкового списку з 28 алгоритмів було відібрано лише 4, оскільки інші мали низьку точність, високу вартість чи проблеми з використанням. Вибраними були алгоритми від Amazon Web Services (AWS Transcribe), Google Speech Recognition, IBM Watson і Rev AI. Ці алгоритми тестували на 60 реченнях, записаних за чотирма умовами мовлення (мелодійне, ритмічне, нормальне повільне та нормальне). Всі алгоритми успішно впоралися з розпізнаванням нормального мовлення, а найкраще впоралися з музичним мовленням AWS Transcribe та IBM Watson, досягнувши точності понад 98%. При додаванні помірного рівня білого шуму та реверберації, AWS Transcribe зберігав прийнятні результати (>70%) або дуже добрі (>95%) показники розпізнавання. Ці результати можуть сприяти розробці програмного забезпечення, яке використовує ASR для проведення самостійних сесій музичної реабілітації

мовлення, таких як терапія мелодійною інтонацією після інсульту, що є особливо корисним у відсутність лікаря.

1.6 Висновки до першого розділу

У даному розділі було розглянуто еволюцію підходів до розпізнавання мовлення, що демонструють поступове вдосконалення методів обробки сигналів та використання штучного інтелекту для автоматичного перекладу усного мовлення у текст. Розпізнавання мовлення, як міждисциплінарна галузь, пройшло шлях від використання примітивних систем зіставлення шаблонів до сучасних моделей глибокого навчання та трансформаторів.

Методи, базовані на прихованих моделях Маркова (HMM) та моделях сумішей Гауса (GMM), створили основу для розпізнавання мовлення завдяки математичній строгості та здатності моделювати послідовність мовних сигналів. Попри їх обмеження, зокрема у моделюванні довгострокових залежностей та варіацій мовлення, ці методи залишаються важливими в історичному та теоретичному контексті, надихаючи розробку сучасних підходів.

Сучасні системи розпізнавання мовлення значно вдосконалили точність і надійність завдяки впровадженню глибоких нейронних мереж (DNN), згорткових нейронних мереж (CNN), повторюваних нейронних мереж (RNN) і архітектур трансформаторів. Їх перевага полягає у здатності моделювати складні нелінійні зв’язки, захоплювати глобальні контексти та забезпечувати ієрархічне представлення даних. Особливої уваги заслуговують попередньо навчені моделі, такі як Wav2Vec і Transformer, які успішно адаптовані до багатомовних завдань та шумових середовищ.

Попри досягнення, сучасні системи розпізнавання мовлення стикаються з низкою викликів:

– варіативність акцентів, діалектів та фонових шумів, необхідність значних обчислювальних ресурсів і затримка в реальному часі;

– обробка неоднозначностей, контекстної інформації, ідіоматичних виразів та омонімів;

– конфіденційність даних, уникнення упередженості та забезпечення доступності технологій для мов із низьким рівнем ресурсів.

Подальші дослідження мають бути спрямовані на:

– розробку більш стійких до шумів та контекстуально обізнаних моделей;

– інтеграцію новітніх технологій обробки сигналів для забезпечення реальної адаптивності до умов;

– використання методів самоконтрольованого навчання для зниження залежності від анотованих даних;

– розробку економічно ефективних рішень, здатних працювати на пристроях з обмеженими ресурсами.

Отже, дослідження та розробки у сфері розпізнавання мовлення не лише підвищують ефективність сучасних систем, але й сприяють інклюзивності, справедливості та відповідальному впровадженню технологій у суспільство.

РОЗДІЛ 2. ДОСЛІДЖЕННЯ СУЧАСНИХ МЕТОДІВ РОЗПІЗНАВАННЯ МОВЛЕННЯ

2.1 Постановка експерименту

Експериментальна установка в сучасній області розпізнавання мовлення має бути виконана з належною продуманістю та реалізацією, щоб переконатися, що результат повторюється, дійсний і надійний. Експериментальна установка створює базову основу для оцінки та порівняльного аналізу продуктивності розпізнавачів мовлення в контрольованих умовах, допускаючи притаманну складність мовних даних. Це відбувається в кілька етапів: визначення цілей, вибір наборів даних, побудова обчислювальної інфраструктури, впровадження моделей вибору та встановлення деяких протоколів оцінювання.

Першим кроком у постановці експерименту з розпізнавання мовлення є чітке формулювання цілей дослідження та гіпотез. Це включає в себе визначення того, на які аспекти розпізнавання мовлення націлено, наприклад, підвищення точності, стійкість до шуму або міжмовну адаптивність. Ці визначені цілі дають вказівки щодо подальших кроків щодо відбору даних, розробки моделі та вибору критеріїв оцінки [13-15]. Як правило, на цьому етапі проводиться розширений огляд літератури, щоб окреслити прогалини в наявних знаннях і встановити контрольний рівень для порівняння.

Відбір даних є одним із найважливіших завдань, пов’язаних із налаштуванням експерименту, оскільки він забезпечує якість і різноманітність набору даних для визначення продуктивності та можливості узагальнення системи розпізнавання мовлення. Вибраний набір даних має відповідати цілям дослідження та представляти передбачувану область застосування. Наприклад, експерименти, зосереджені на покращенні розпізнавання в шумі, можуть базуватися на наборах даних, таких як CHiME або Aurora, які

включають записи з широким діапазоном умов фонового шуму. Багатомовні експерименти потребуватимуть таких наборів даних, як Common Voice або VoxForge, які забезпечують записи багатьма мовами та діалектами.

Після вибору набору даних попередня обробка даних обов'язково стане ключовою передумовою узгодженості та адекватності для навчання та тестування. Він включає такі кроки, як нормалізація аудіо, сегментація та вилучення функцій. Нормалізація нормалізує рівні звуку, тоді як сегментація розбиває безперервні записи на менші керовані одиниці, такі як фонеми, слова або речення [15]. Методи виділення ознак, такі як кепстральні коефіцієнти Mel-частоти або генерація спектрограм, використовуються для перетворення необробленого аудіо в представлення, які більше піддаються алгоритмам машинного навчання. Іноді застосовується розширення даних, наприклад додавання синтетичного шуму, зміщення висоти або розтягнення часу, щоб зробити моделі більш стійкими до варіацій у вхідних даних.

Для проведення експериментів необхідно налаштувати обчислювальне середовище для навчання та оцінювання: обрати відповідні апаратні ресурси, такі як, наприклад, високопродуктивні графічні процесори або TPU, і налаштувати програмну структуру, на кшталт TensorFlow, PyTorch або Kaldi. Фреймворк буде вибрано відповідно до складності моделі, простоти впровадження та можливості підтримки певних функцій: налаштованих втрат або навчання кількох GPU. Крім того, інструменти налаштування, такі як Optuna або Ray Tune, інтегровані в робочий процес і автоматизують процес вибору найкращих параметрів: швидкості навчання, розміру партії та інших налаштувань, що залежать від архітектури.

Реалізація моделі включатиме проектування та налаштування архітектури розпізнавання мовлення. Загальна обробка в сучасних мовленнєвих експериментах часто включає модель глибокого навчання, включаючи CNN, рекурентну нейронну мережу або трансформатори. Ці архітектури відрізняються точними особливостями, які відповідають різним

сформульованим питанням, їх вхідним представленням, глибиною та вихідними конфігураціями. Більшість поточних зусиль спрямовано на вирішення деяких із цих проблем для кращої продуктивності. Наприклад, наскрізні моделі, такі як CTC або кодер-декодер на основі уваги, сьогодні все частіше застосовуються, оскільки вони можуть дозволити спільну оптимізацію як акустичної, так і мовної моделі без будь-якої форми явного узгодження.

Під час навчання моделі відбувається передача попередньо оброблених даних через мережу з одночасним ітеративним оновленням параметрів, які передбачають мінімізацію попередньо визначених втрат. Ваги та зміщення коригуються за допомогою таких методів, як стохастичний градієнтний спуск або оптимізатор Адама [16]. Техніки регулярізації, включаючи відсівання та зниження ваги, реалізуються, щоб уникнути надмірного навантаження, тоді як критерії ранньої зупинки запобігають додатковим циклам тренувань. Моніторинг виконується на основі таких показників, як конвергенція втрат і точність перевірки, які гарантують, що модель добре навчається.

Для будь-якої даної системи розпізнавання мовлення ретельно розробляються спеціальні протоколи оцінки, які перевіряють її продуктивність і узагальнення. Зазвичай це робиться шляхом поділу даних на підмножини навчання, перевірки та тестування. Тестовий набір, який зберігається осторонь процесу навчання, діє як еталон для оцінки моделі на предмет її точності, надійності та адаптивності. Ефективність розпізнавання кількісно визначається за допомогою таких показників, як частота помилок у словах, частота помилок фонем і частота помилок символів. Іноді якісний аналіз також може бути частиною експериментів, наприклад, можна перевірити візуалізацію карти уваги або результати транскрипції, щоб отримати уявлення про поведінку моделі.

Порівняльний аналіз для різних конфігурацій моделі або базових ліній часто є частиною експериментальної установки. Він охоплюватиме навчання

та порівняння кількох архітектур, наприклад, порівняння результатів між традиційними RNN та моделями на основі трансформаторів або вивчення ефекту різних гіперпараметрів. Щоб переконатися, що відмінності в продуктивності між різними конфігураціями є статистично значущими, проводяться парні t-тести або тести статистичної значущості початкового завантаження. Ці порівняння є цінними, оскільки демонструють ефективність запропонованих методів і поміщають їх у контекст більшого дослідження розпізнавання мовлення [15-17].

Іншим важливим аспектом експериментальної установки є перевірка надійності: здатність системи працювати в складних умовах випробувань, таких як високий рівень шуму, невидимі акценти або аудіо низької якості. Тести надійності дають підказку щодо узагальнення навчальних даних і визначають конкретні області для вдосконалення моделі. Іноді він може використовувати змагальні атаки, коли до звуку додаються непомітні збурення, щоб перевірити його стійкість до шкідливих вхідних даних. Це означало б наступне: ретельно задокументувати всі деталі експериментальної установки, починаючи від попередньої обробки даних до впровадження моделі та процедур навчання, а також критеріїв оцінки. Цей репозиторій є одним із прикладів, і в більшості випадків він контролюється версіями, щоб ділитися кодом і наборами даних, а іноді й попередньо підготовленими моделями, з іншими людьми в дослідницькій спільноті. Прозорість такого роду дозволяє незалежно перевіряти результати, ділитися досвідом і стимулювати інновації всередині спільноти.

Етичні міркування є невід’ємною частиною експериментальної установки, особливо щодо конфіденційних даних або даних, що дозволяють ідентифікувати особу. Необхідно дотримуватися вказівок, які забезпечують дотримання етики та дозволів на використання. Захист здійснюватиметься шляхом анонімізації, шифрування та зберігання в безпечному місці для збереження конфіденційності та прав учасників. Соціальні наслідки системи

розпізнавання мовлення критично досліджуються, щоб переконатися, що її розгортання відповідальне, враховуючи, наприклад, такі питання, як упередженість і справедливість.

Це передбачає підготовку експерименту в галузі сучасного розпізнавання мовлення, який є однією з таких багатовимірних робіт: комплексне та систематичне осмислення відбору та попередньої обробки даних для впровадження та оцінки моделі. Таким чином, приділяючи детальну увагу елементам із суворістю та точністю, допомагає отримати більш значущу інформацію, удосконалити сучасні технології та зробити внесок у створення надійніших і справедливіших систем розпізнавання мовлення.

2.2 Вибір методів для аналізу ефективності розпізнавання

Методи, за допомогою яких вищеописаний аналіз може бути проведений, утворюють втілення будь-яких значущих дій, які можуть бути виведені з нього. Ефективність розпізнавання в будь-якій системі розпізнавання мовлення може бути виконана за допомогою вичерпної кількості методів, як якісних, так і кількісних, щоб забезпечити найкращі характеристики для надійності та адаптивності на різних рівнях [14]. Ці методи вибираються з особливою ретельністю, щоб забезпечити всебічну оцінку досліджуваної системи щодо конкретних цілей, сценаріїв застосування та потенційних обмежень. Основною основою вибору аналітичних методів є визначення цілей оцінювання. Серед загальних цілей – визначення точності системи, стійкості до варіацій і здатності до узагальнення для різноманітних наборів даних і умов. Встановлення чітких цілей не тільки допомагає у виборі показників, але й впливає на розробку тестових сценаріїв та інтерпретацію результатів. Наприклад, система, орієнтована на транскрипцію в режимі реального часу в галасливому середовищі, потребує аналітичних методів, що

підкреслюють стійкість до шуму, тоді як багатомовна система вимагає підходів, які враховують лінгвістичне розмаїття та міжмовну передачу.

Розрахунок частоти помилок є одним із основних прийомів аналізу ефективності розпізнавання, який забезпечує кількісну оцінку точності системи. Серед практичних заходів метрики WER, PER і CER стали популярними для дослідження якості узгодження між виходом системи та істинністю землі. Зокрема, WER є стандартним тестом, який вимірює частку помилок у термінах замін, вставок і видалень відносно загальної кількості слів. Цей показник буде особливо корисним для вимірювання загальної ефективності та порівняння з встановленими базовими показниками. Однак показники частоти помилок не є хорошими представниками семантичної точності транскрипцій, і тому інші методи повинні використовуватися для доповнення цих заходів.

Набір даних має бути репрезентативним для передбачуваної області застосування, оскільки мають бути охоплені всі лінгвістичні, акустичні та контекстуальні варіації [6-9]. Необхідно, щоб система була піддана наборам даних, які мають різні акценти, стилі розмови та умови фонового шуму щодо її адаптивності та надійності. На додаток до цього, статистична надійність стосується розміру набору даних, у якому більша впевненість в узагальненні результату досягається, коли більший набір даних. Щоб уникнути будь-якої упередженості у виборі наборів даних, зазвичай дотримуються методів перехресної перевірки, розділяючи дані на набори для навчання, перевірки та тестування, щоб уникнути переобладнання та правильно оцінити продуктивність системи.

Окрім частоти помилок, ефективність розпізнавання часто аналізується методами, які оцінюють контекстне та семантичне розуміння системи. Інші методи включають використання вимірювань здивування та семантичної схожості, щоб визначити, наскільки система вловлює значення та структуру розмовної мови. Здивування, адаптоване з мовного моделювання, забезпечує

кількісну міру передбачуваної здатності системи щодо невизначеності в передбаченні послідовних слів у послідовності. Чим менше збентеження, тим кращі можливості моделювання мови – дуже важливий фактор у таких програмах, як автоматичне транскрибування та діалогові системи. У той час як вимірювання семантичної подібності, з іншого боку, обчислюють, наскільки вихідні дані системи концептуально збігаються з основною правдою і, отже, вказують на збереження значення навіть у випадках, коли мають місце поверхневі розбіжності.

Аналіз стійкості є важливою частиною оцінки ефективності розпізнавання, особливо у випадку систем, які піддаються впливу реальних умов. Тобто систему необхідно випробувати в складних умовах щодо фонового шуму, спотворень мови та мінливості акценту [1-3, 6]. Різноманітність методів модифікації мови, що застосовуються для синтетичного створення таких умов, включає використання додаткового шуму, реверберації та частотного маскування для перевірки системи на несприятливі акустичні умови. Аналіз стійкості дає розуміння стійкості системи і, таким чином, керує розробкою стійких до шуму моделей і методів попередньої обробки.

Тимчасова продуктивність вимірюється такими методами, як аналіз часу відгуку та вимірювання пропускної здатності. Ці методи використовуються для визначення здатності системи обробляти мовні дані з ефективністю та низькою затримкою. Аналіз часу відповіді розглядатиме час, потрібний системі для транскрипції певного вхідного сигналу, тоді як вимірювання пропускної здатності розглядатиме кількість зразків мовлення, оброблених протягом заданого проміжку часу. Ці показники дуже актуальні для додатків, які вимагають взаємодії в реальному часі, включаючи віртуальних помічників і системи живих субтитрів. Забезпечення низької затримки та високої пропускної здатності буде важливим для покращення взаємодії з

користувачем і забезпечення бездоганної інтеграції в робочі процеси, чутливі до часу.

Можливість інтерпретації та аналіз помилок йдуть поруч, підтримуючи розуміння обмежень і областей вдосконалення систем розпізнавання мовлення. Аналізуючи типи та шаблони вироблених помилок, система може знайти слабкі місця в обробці омофонів, рідкісних слів або термінології, що стосується предметної сфери. Як правило, аналіз помилок вимагає ручної перевірки транскрипцій та їх категоризації на основі мовних або акустичних характеристик. Наприклад, помилки можна згрупувати за фонетичною плутаниною, граматичною невідповідністю або семантичною невідповідністю [15]. Цей детальний аналіз дає корисну інформацію для вдосконалення моделі та вирішення конкретних проблем у розпізнаванні. Інший вимір аналізу ефективності розпізнавання передбачає оцінку адаптивності системи до невидимих даних і нових доменів. Трансферне навчання та методи адаптації домену зазвичай використовуються для розширення можливостей системи за межі початкового обсягу навчання. Наприклад, щоб оцінити цю адаптивність, дослідники розробляють експерименти, які перевіряють продуктивність системи на наборах даних, що вводять нові лінгвістичні чи акустичні особливості. Деякі заходи для кількісної оцінки передачі знань і здатності до адаптації домену включають підвищення точності після точного налаштування та помилки узагальнення. Ці експерименти демонструють, як системи розпізнавання мовлення можна використовувати в широкому діапазоні середовищ і контекстів.

Хоча багато уваги приділялося оцінюванню ефективності систем розпізнавання мовлення, на перший план вийшли етичні міркування та аналіз справедливості. Вони стосуються тестування відхилень у продуктивності системи для різних демографічних груп. Для кількісної оцінки розбіжностей можуть використовуватися різні показники справедливості, включаючи індекси розбіжностей і зрівняні шанси. Розуміння цих упереджень є

основоположним для досягнення справедливого доступу до кількох додатків, від охорони здоров’я до правової транскрипції, тим самим зменшуючи негативні наслідки.

Щоб переконатися, що аналіз ефективності розпізнавання є відтворюваним і прозорим, необхідно прийняти експериментальні протоколи та практику документування [16]. Це включає в себе всі параметри та конфігурації, які можуть бути використані під час оцінки, включаючи попередню обробку даних, архітектуру моделі та процедури навчання. Багато дослідників використовують інструменти та фреймворки з відкритим кодом, щоб стандартизувати свій аналіз і забезпечити незалежну перевірку своїх результатів. Крім того, зростає кількість практик, які застосовуються дослідниками для подальшого підвищення довіри та впливу висновків, таких як попередня реєстрація та публікація докладних звітів про оцінку. Таким чином, вибір методів для аналізу систем розпізнавання мовлення є багаторівневим процесом, який потребує детальної уваги щодо цілей, наборів даних і методів оцінювання. Використовуючи кількісні показники, тести надійності, аналізи інтерпретації та етичні оцінки разом, дослідники можуть отримати повне розуміння продуктивності системи та шляхів подальшого вдосконалення. Ці методології виходять за рамки розвитку сучасного розпізнавання мовлення до розробки систем, які є точними, надійними та інклюзивними.

2.3 Дослідження впливу шуму та інших факторів на точність

Аналіз ефективності систем розпізнавання мовлення належить до ключових аспектів у процесі її розробки, а вибір відповідних методів для проведення такого аналізу дозволяє зробити значущі та корисні висновки. Методології різноманітні, починаючи від різних кількісних і якісних методів, де кожна адаптована до різних аспектів, включаючи продуктивність системи,

надійність і адаптивність [15-17]. Для найбільш ефективного оцінювання ці методи мають бути обрані з урахуванням конкретних цілей, сценаріїв застосування та можливих обмежень досліджуваної системи.

На рисунку 2.1 наведено вплив шуму на точність розпізнавання мовлення.

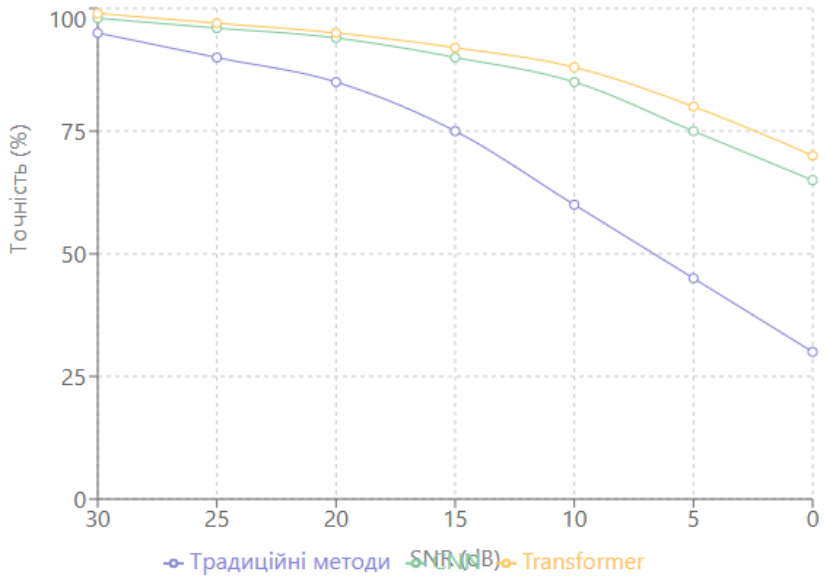

Рисунок 2.1 – Вплив шуму на точність розпізнавання мовлення

Визначення цілей оцінювання є першим важливим фактором у виборі аналітичних методів. У більшості випадків оцінка спрямована на встановлення точності, стійкості до варіацій і здатності системи до узагальнення для різних наборів даних і умов. Чіткі цілі також допомагають у виборі показників, але, що більш важливо, керують розробкою тестових сценаріїв та інтерпретацією результатів. Наприклад, система, яка буде використовуватися для транскрипції в галасливому середовищі, потребує методів, що підкреслюють шумостійкість, тоді як багатомовна система вимагає підходів, які враховують мовне розмаїття та міжмовну передачу [13]. Один із найбільш фундаментальних методів аналізу ефективності розпізнавання, обчислення частоти помилок, дозволяє кількісно визначити точність системи. Загальні

показники включають частоту помилок слів, частоту помилок фонем і частоту помилок символів. Серед них WER є однією зі стандартних опорних точок, які кількісно визначають помилки як відношення загальної кількості слів із заміною, вставкою та видаленням до загальної кількості. Зокрема, це зручно як метрика для порівняння глобальної продуктивності систем із встановленими базовими показниками. Однак показники частоти помилок мають власні обмеження щодо фіксації семантичної точності в результатах транскрипції, отже мотивація для включення додаткових методів для доповнення цих показників.

Набори даних мають бути репрезентативними для передбачуваної області застосування та охоплювати різні лінгвістичні, акустичні та контекстуальні варіації. До них належать набори даних, що відрізняються акцентом, стилем мовлення та умовами фонового шуму в реченні, що стає критичним фактором для оцінки адаптивності та надійності системи. Крім того, розмір набору даних є важливим фактором статистичної надійності: більші набори даних породжують більшу впевненість у можливості узагальнення результатів. Щоб уникнути упередженості у виборі наборів даних, дослідники зазвичай використовують перехресну перевірку: поділяють дані на навчальні [12], перевірочні та тестові набори, щоб уникнути переобладнання, щоб отримати реалістичне уявлення про продуктивність системи.

Окрім частоти помилок, аналіз ефективності розпізнавання зазвичай включає методи, які оцінюють контекстне та семантичне розуміння системи. Ефективність цього зазвичай вимірюється за допомогою таких методів, як вимірювання здивування та семантичної схожості, які показують, який зміст і структуру система вловить із усного мовлення. Здивування, запозичене з мовного моделювання, кількісно фіксує передбачуваність системи через невизначеність у передбаченні наступного слова в послідовності. Чим нижчі значення здивування, тим краща здатність моделювати мову, що є дуже

важливим у програмах, таких як автоматична транскрипція та діалогові системи [4, 5]. З іншого боку, вимірювання семантичної подібності оцінюють відповідність вихідних даних системи базовій істині на концептуальному рівні та дають зрозуміти, наскільки добре значення зберігається навіть у випадках, коли виникають поверхневі розбіжності.

Аналіз стійкості є важливою частиною оцінки ефективності розпізнавання, особливо для систем, що працюють у реальному середовищі. Це включає в себе випробування системи різними умовами, такими як підвищений рівень фонового шуму, спотворення мови та різноманітність акцентів. На практиці це включає генерацію синтетичного шуму та методи модифікації мовлення для контрольованої імітації таких умов. Додатковий шум, реверберація та частотне маскування є найпоширенішими методами, які використовуються для оцінки надійності системи в несприятливих акустичних сценаріях. Аналіз надійності надає низку підказок щодо стійкості системи та керує розробкою стійких до шуму моделей і методів попередньої обробки.

Тимчасова продуктивність вимірюється такими методами, як аналіз часу відгуку та вимірювання пропускної здатності [7]. Ці методи оцінюють ефективність системи в обробці мовних вхідних даних і надання вихідних даних із мінімальною затримкою. Аналіз часу відповіді вимірює час, потрібний системі для транскрипції даного вхідного сигналу, тоді як вимірювання пропускної здатності передбачає кількість зразків мовлення, оброблених протягом заданого періоду часу. Вищезазначені показники є найбільш актуальними для додатків, які вимагають взаємодії в реальному часі, таких як віртуальні помічники та живі субтитри. Низька затримка та висока пропускна здатність забезпечують покращену взаємодію з користувачем, що робить його безперебійним для будь-яких користувачів, які працюють у робочих процесах, чутливих до часу. Можливість інтерпретації та аналіз помилок є невід’ємною частиною розуміння обмежень і обсягу систем розпізнавання мовлення. Різні типи допущених помилок разом із їх

індивідуальними моделями розподілу дозволяють вказати на конкретні слабкі сторони системи, такі як омофони, складність рідкісних слів і проблеми термінів, пов’язаних із доменом. У деяких випадках перевірка транскрипції вручну визначила класи помилок або як частину мовних чи акустичних атрибутів, або як помилку фонетичної, граматичної чи семантичної природи. Це дає детальний аналіз, необхідний для ефективних ідей для вдосконалення моделі та вирішення конкретних проблем у розпізнаванні. Іншим аспектом аналізу ефективності розпізнавання є визначення здатності системи адаптуватися до невидимих даних і нових доменів [9]. Трансферне навчання та методи адаптації домену широко використовуються для розширення можливостей системи за межі початкового обсягу навчання. Це дозволяє оцінити його адаптивність, коли плануються експерименти, перевіряючи продуктивність на наборах даних із новими лінгвістичними чи акустичними характеристиками. Заходи для кількісної оцінки цієї можливості включають покращення точності після тонкого налаштування та помилки узагальнення системи для передачі знань та адаптації домену. Ці експерименти, з іншого боку, вказують на можливість позиціонування систем розпізнавання мови в найрізноманітніших і динамічних контекстах.

Аналіз етики та справедливості сьогодні став поширеним явищем у системах розпізнавання мовлення, перевіряючи упередженість, що виникає в її продуктивності між демографічними ознаками статі, віку чи етнічної приналежності. Таким чином, різні показники справедливості – кожен з різними індексами невідповідності для оцінки, вирівнювання шансів – визначають, як продуктивність може відрізнятися між собою, і стають джерелом упередженості. Вкрай важливо мати ці міркування, інакше це зрештою може призвести до небажаних наслідків для будь-якого набору програм, починаючи від охорони здоров’я й закінчуючи легальним доступом до транскрипції [10].

Щоб зробити процес аналізу ефективності розпізнавання більш відтворюваним і прозорим, застосовуються суворі експериментальні протоколи та практика документування. Це включає в себе специфікацію всіх параметрів і конфігурацій, які використовуються під час оцінки, наприклад етапи попередньої обробки набору даних, архітектури моделі та процедури навчання. Інструменти та фреймворки з відкритим кодом часто використовуються для стандартизації процесу аналізу та забезпечення незалежної перевірки результатів. Крім того, зростає тенденція робити дослідження більш надійними та вагомими за допомогою таких практик, як попередня реєстрація та публікація розширених звітів про оцінку.

Отже, вибір методів для аналізу ефективності розпізнавання систем розпізнавання мовлення є складним процесом, який вимагає ретельного розгляду цілей, наборів даних і методів оцінювання. Поєднуючи кількісні показники з тестами надійності, аналізом інтерпретації та етичними оглядами, можна досягти загального розуміння продуктивності системи та шляхів її вдосконалення в майбутньому. Ці методології вдосконалюють сучасні технології розпізнавання мовлення, а також сприяють розробці точних, надійних і комплексних систем.

2.4 Порівняльний аналіз традиційних і сучасних алгоритмів

Традиційні системи розпізнавання мовлення базуються на класичних алгоритмах, а саме на прихованих моделях Маркова та моделях суміші Гауса. HMM представляють дуже хороший спосіб моделювання часових залежностей і послідовних структур, присутніх у мовних сигналах завдяки їх імовірнісній природі. Вони моделюють мовлення як послідовність станів, кожен з яких представляє одну фонему або акустичну одиницю, і включають ймовірності переходу, які визначають можливість переходу з одного стану в інший. У цьому формулюванні HMM здатні вловлювати варіації в тривалості

мовлення та часовому вирівнюванні [11]. Доповнюючи HMM, GMM використовуються для моделювання розподілу акустичних характеристик. Вони моделюються в GMM як зважені комбінації компонентів Гауса, що дозволяє системі вивчати мінливість мовного сигналу через відмінності в динаміках, акцентах і умовах навколишнього середовища.

На рисунку 2.2 наведено порівняння методів розпізнавання мовлення.

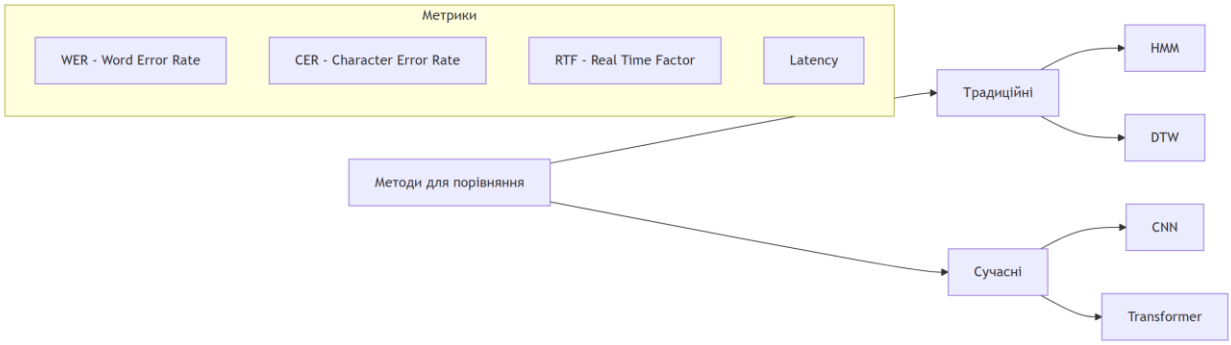

Рисунок 2.2 – Порівняння методів розпізнавання мовлення

Таким чином, інтеграція HMM і GMM формує надійну структуру для розпізнавання мовлення, де акустичні та мовні моделі співпрацюють у декодуванні мовних вводів у текст. Незважаючи на їх потужність, ці моделі обмежені рядом обмежень. По-перше, довіра до створених вручну функцій, таких як кепстральні коефіцієнти Mel-частоти, додає експертну залежність домену та зменшує здатність системи узагальнювати в абсолютно новій ситуації. Припущення про незалежність, які використовуються в GMM, обмежують їх здатність моделювати складні зв’язки між функціями, що призводить до низької продуктивності за несприятливих акустичних умов. Отже, незважаючи на те, що обмеження стали важливою складовою, вони переросли в інший напрямок досліджень і вивчення поза системами HMM-GMM.

Сучасні алгоритми, створені за допомогою нейронних мереж і методів глибокого навчання, змінили обличчя розпізнавання мовлення. Нейронні

мережі, в основному глибокі архітектури, вивчають ієрархічне представлення особливостей мови з великих обсягів даних і обчислень. Ці моделі працюють безпосередньо з необробленими сигналами або спектрограмами без необхідності самостійного виділення ознак. Отже, вони вловлюють складні закономірності та їх кореляцію в даних. Застосування CNN та RNN покращує систему обробки просторової та часової інформації відповідно. Таким чином, CNN є кращими щодо вилучення локального шаблону зі спектрограм у таких аспектах, як форманти та гармоніки, тоді як довгострокові залежності та послідовну динаміку, притаманні мовним сигналам, можуть бути захоплені RNN [13]. Яскравою тенденцією сучасного алгоритму є широке впровадження наскрізної архітектури, наприклад, таких моделей, як коннекціоністська часова класифікація та структури кодера-декодера на основі уваги. Ці методи обходять складнощі в конвеєрі розпізнавання мовлення шляхом прямого зіставлення вхідних функцій із вихідним текстом, минаючи явне моделювання фонем та їх вирівнювання. CTC використовує ймовірнісний механізм для вирівнювання, який дозволяє вводити змінну довжину, поступаючись місцем гнучкому навчанню наскрізних систем. У той час як механізми уваги дозволяють моделі динамічно приділяти більше уваги відповідним частинам введення під час декодування, підвищуючи здатність обробки довгого висловлювання. Ці дві інновації значно підвищили точність розпізнавання та масштабованість, зробивши сучасні алгоритми де-факто стандартом для найсучасніших систем.

Серед останніх розробок у розпізнаванні мовлення трансформатори, особливо ті, що включають такі моделі, як Transformer-XL і wav2vec, представляють зміну парадигми. Трансформатори покладаються на механізми самоконтролю, які моделюють глобальні залежності у вхідних даних, фіксуючи контекстну інформацію про цілі послідовності. Ось чому трансформатори перевершують моделі на основі RNN, особливо там, де задіяно важке моделювання контексту, таке як розпізнавання розмовного

мовлення та транскрипція довгого звуку. Подібним чином деякі попередньо підготовлені моделі трансформаторів, такі як BERT і GPT, були переведені на завдання розпізнавання мовлення шляхом тонкого налаштування, яке використовує контекстні вбудовування для покращення продуктивності. Незважаючи на ці переваги, сучасні алгоритми не позбавлені проблем. Використання великих наборів даних і обчислювальних ресурсів викликає занепокоєння щодо доступності та сталості [16]. Навчання глибоких нейронних мереж вимагає величезних витрат енергії, що знову ж таки сприяє впливу на навколишнє середовище. Крім того, керовані даними моделі потребують високоякісних маркованих наборів даних, які можуть бути дефіцитними або навіть недоступними для багатьох мов і діалектів. Ці обмеження вказують на потребу в інноваціях у способах ефективного навчання з даними, таких як самоконтрольовані та напівконтрольовані методи, які намагаються зменшити залежність від позначених даних без погіршення продуктивності.

Порівняння, проведене між традиційними та сучасними алгоритмами, є чітким у лінії еволюції: більшість обмежень, яких зазнавали традиційні моделі, вже були усунені за допомогою сучасних підходів. Однак синергію можна побачити саме в інтеграції методологій. Гібридні моделі, які поєднують інтерпретабельність і структуру HMM з репрезентативною потужністю глибоких нейронних мереж, показали велику перспективу в подоланні розриву між двома парадигмами. Ці моделі використовують сильні сторони обох підходів, підвищуючи надійність і адаптивність, зберігаючи можливість інтерпретації.

Розвиток від традиційних до сучасних алгоритмів розпізнавання мовлення відображає динамічну взаємодію між теоретичним прогресом і практичними інноваціями. У той час як традиційні моделі, такі як HMM і GMM, забезпечили фундаментальне уявлення про акустичне моделювання та послідовну обробку, сучасні алгоритми, що використовують нейронні мережі

та глибоке навчання, створили роботу для досягнення безпрецедентної точності та універсальності. Порівняння між цими підходами не лише підкреслює відповідний внесок, який вони зробили, але й вказує на постійні виклики та можливості. У своєму постійному розвитку системи розпізнавання мовлення відіграватимуть значну роль в інтеграції традиційних і сучасних методологій з метою збереження їх надійності, адаптованості та доступності для кожного користувача в технологічному світі, який постійно розвивається.

2.5 Використання мови програмування Python

Застосування мови програмування Python у розробці систем розпізнавання мовлення значно зростає завдяки її гнучкості, широкій бібліотечній підтримці та активній спільноті. Розпізнавання мовлення ініціюється отриманням звуку, який у цьому випадку зазвичай має форму записів мовлення, зібраних із різних джерел. Python, завдяки своїй простоті та просторості в екосистемі своїх бібліотек, є чудовим інструментом для маніпулювання аудіоданими [17]. Такі бібліотеки, як pydub і librosa, дозволяють легко змінювати аудіофайли, включаючи конвертацію між форматами, обрізання частин аудіо, повторну вибірку та сегментацію, серед іншого. Ці кроки є важливими для попередньої обробки, оскільки вони кондиціонують дані та передають аудіовхідні дані в узгоджений і придатний для використання формат.

Виділення ознак є досить важливим кроком у розпізнаванні мовлення, пов’язаним із виділенням значущих ознак із необроблених аудіосигналів. Завдяки потужній підтримці бібліотек Python можна легко реалізувати методи вилучення функцій. Наприклад, бібліотека librosa надає простий у використанні інтерфейс для аналізу аудіосигналу та виділення таких функцій, як MFCC, функції кольоровості та спектральний контраст. Перш за все, MFCC отримали привілейоване місце в системах розпізнавання мови, оскільки

спектральні властивості, які вони представляють, дуже схожі на мовні сигнали, які сприймаються людьми. Завдяки цим представленням функцій Python забезпечує представлення аудіоданих у виразній, але компактній формі, що найкраще підходить для подальших завдань моделювання.

Після виділення функцій Python має надати інструменти для розробки та навчання моделей, що відображають ці функції в тексті. Традиційні підходи, такі як HMM і GMM, можна легко реалізувати за допомогою таких бібліотек, як hmmlearn і scikit-learn. Ці бібліотеки містять утиліти для побудови імовірнісних моделей, навчання їх на основі позначених даних і оцінки продуктивності [13]. Однак перехід у бік глибокого навчання призвів до широкого впровадження сучасних фреймворків, таких як TensorFlow і PyTorch, які набагато краще підходять для створення систем розпізнавання мовлення. Глибинні нейронні мережі є основою сучасного розпізнавання мовлення. Крім того, інтеграція Python із фреймворками глибокого навчання під назвою TensorFlow і PyTorch надає можливість реалізувати складні архітектури, включаючи CNN, RNN і трансформатор. Для виділення просторових характеристик аудіо, представленого у вигляді спектрограми, багато хто використовує CNN, тоді як RNN добре вивчає часові залежності мови, особливо ті, що мають осередки LSTM або GRU.

Python також природно підтримує реалізацію чистого наскрізного розпізнавання мовлення, де функції необробленого аудіо відображаються безпосередньо в тексті без проміжного представлення. У цю парадигму входять SpeechBrain і ESPnet: кожен надає набір попередньо визначених будівельних блоків разом із деякими базовими утилітами для побудови та навчання наскрізної моделі. Це економить час користувачів, уникаючи необхідності звертатися до деталей низького рівня, але дозволяє користувачеві просто експериментувати з високою абстракцією з легкістю. Ефективне впровадження в ESPnet для навчання та декодування як CTC, так і архітектури кодера-декодера на основі уваги дозволяє знайти широке застосування. Іншим

ключовим моментом, пов’язаним із застосуваннями Python у розпізнаванні мовлення, є підтримка передачі навчання та попередньо навчених моделей. Зокрема, найсучасніші моделі, такі як Wav2Vec 2.0 і Whisper від Hugging Face’s Transformers і PyTorch Hub, відповідно, більшість із них попередньо навчені на великомасштабних наборах даних для точного налаштування доменно-специфічних даних [8]. Гнучкість Python означає, що всі ці попередньо навчені моделі можна легко інтегрувати в робочий процес, прискорюючи процес розробки та знижуючи бар’єри для початківців.

Варто зазначити, що оцінка завжди буде критичною частиною життєвого циклу розробки, і для цього в Python є численні інструменти. Типові показники, які використовуються для кількісної оцінки точності результату розпізнавання, включають частоту помилок слів, частоту помилок фонем і частоту помилок символів. Такі бібліотеки, як jiwer, спрощують обчислення таких показників, тому можна фактично інтерпретувати результати та знаходити шляхи покращення. Крім того, бібліотеки візуалізації Python, включаючи Matplotlib і Seaborn, можуть створювати глибокі діаграми, що відображають тенденції продуктивності та шаблони помилок, щоб надати цінний зворотний зв’язок для вдосконалення моделі.

На таких платформах, як GitHub і Kaggle, є багато відкритих проектів і наборів даних, на яких можна базувати роботу та сприяти подальшому розвитку галузі. Звичайно, використання Python для розпізнавання мовлення має недоліки. Для навчання моделі глибокого навчання потрібне високопродуктивне обладнання, як-от графічні процесори або процесори TPU. Крім того, Python залежить від динамічного введення та інтерпретованого виконання, що робить його неефективним у деяких сценаріях, особливо під час роботи з великомасштабними даними [10]. Через це не дивно, що для пом’якшення цих проблем зазвичай використовують взаємодію Python із такими мовами, як C++ і CUDA, а також його підтримку для фреймворків розподілених обчислень, таких як Apache Spark і Dask.

Підсумовуючи, можна зазначити, що Python – це потужний і універсальний інструмент для розробки систем розпізнавання мовлення, який надає розгалужену екосистему бібліотек і фреймворків для всіх аспектів конвеєра розробки. Python допомагає створювати надійні та точні системи, надаючи їм легкий робочий процес у попередній обробці даних, вилученні функцій, навчанні й оцінюванні. Це допомагає розширити межі розпізнавання мовлення.

2.6 Висновки до другого розділу

Висновками до поточного розділу є:

1) експериментальна установка в галузі розпізнавання мовлення повинна бути ретельно спланована, включаючи визначення цілей, вибір відповідних наборів даних, обчислювальної інфраструктури та протоколів оцінки. Такі підходи дозволяють досягти високої точності, стійкості до шуму та адаптивності до міжмовних варіацій. Особлива увага приділяється узгодженості наборів даних та адекватності передобробки, що забезпечує надійність результатів;

2) аналіз продуктивності систем розпізнавання мовлення передбачає використання якісних і кількісних методів, таких як частота помилок слів (WER), фонем (PER), символів (CER), а також семантична подібність та здивування. Дослідження також зосереджується на адаптивності системи до різних умов та вимірюванні її надійності за несприятливих умов;

3) одним із ключових напрямів є дослідження впливу шуму, акцентів і низької якості аудіо на точність системи. Використання синтетичних шумів і методів частотного маскування дозволяє перевірити стійкість моделей і визначити шляхи для підвищення їх ефективності;

4) традиційні алгоритми, такі як HMM та GMM, забезпечують базову функціональність, але обмежені у продуктивності в складних сценаріях.

Сучасні підходи на базі глибоких нейронних мереж (CNN, RNN, трансформери) значно підвищують точність і масштабованість. Гібридні моделі, які поєднують традиційні та сучасні методи, демонструють перспективу у вирішенні обмежень кожного підходу;

5) мова програмування Python забезпечує багатий інструментарій для розробки, включаючи бібліотеки для обробки аудіо (librosa, pydub), побудови моделей (TensorFlow, PyTorch), та оцінки результатів (jiwer). Гнучкість мови сприяє швидкому розвитку як початкових, так і вдосконалених систем розпізнавання мовлення. Разом з тим важливо враховувати необхідність високопродуктивного обладнання для ефективного навчання моделей.

Вказані результати підкреслюють важливість систематичного підходу до побудови, оцінювання та вдосконалення систем розпізнавання мовлення, а також роль інновацій у забезпеченні їхньої стійкості та ефективності.

РОЗДІЛ 3. ПРОГРАМНА РЕАЛІЗАЦІЯ

3.1 Мета та вимоги до розробки

Постановка задачі програмної реалізації базується на необхідності створення системи розпізнавання мовлення з розширеним функціоналом, яка здатна виконувати запис, обробку, аналіз аудіоданих та інтерпретацію тексту з аудіосигналів. Ця задача розглядається з використанням мови програмування Python, яка забезпечує широкий набір бібліотек та інструментів для роботи з аудіоданими, машинним навчанням і статистичним аналізом.

Метою є розробка модульної програми, що забезпечує:

1) запис аудіозразків. Тобто користувач може записати серію аудіофайлів із заданою тривалістю та кількістю, які зберігаються у визначеній директорії;

2) попередню обробку аудіо. Отримані аудіофайли проходять стадії вилучення основних акустичних характеристик, таких як тривалість, середній рівень енергії сигналу (RMS), частота нульових переходів та спектральні властивості;

3) розпізнавання тексту. Аудіодані аналізуються за допомогою механізмів перетворення мовлення в текст із використанням хмарних сервісів та локальних моделей;

4) аналіз та візуалізація результатів. Забезпечується створення статистичних звітів щодо успішності розпізнавання та побудова візуалізацій, таких як кореляційні матриці чи розподіл ознак;

5) розрахунок статистичних показників. визначаються метрики успішності роботи системи (наприклад, частота вдалого розпізнавання) та середні значення ключових ознак сигналу.

Функціональні вимоги до системи включають:

– підтримку багатомовності при розпізнаванні мовлення;

– можливість розширення функціоналу через інтеграцію нових методів машинного навчання;

– легке налаштування параметрів для збору даних та їх обробки.

Система призначена для практичного використання у дослідницьких цілях, для створення навчальних наборів даних або розробки комерційних додатків із розпізнавання мовлення.

3.2 Вибір інструментів і технологій

Вибір інструментів і технологій для розробки програмного забезпечення є ключовим етапом, який визначає ефективність реалізації системи, її функціональні можливості, масштабованість і зручність використання. У розробці розширеної системи розпізнавання мовлення були використані сучасні інструменти й технології, що забезпечують максимальну продуктивність і гнучкість. Основними аспектами вибору інструментів стали їх здатність працювати з аудіоданими, забезпечувати підтримку машинного навчання, реалізовувати складні обчислення та створювати зручний інтерфейс для користувача.

Основою реалізації стала мова програмування Python, яка є універсальним і широко вживаним засобом розробки програмного забезпечення. Вибір Python обумовлений його багатофункціональним екосистемним підходом, підтримкою численних бібліотек для обробки аудіоданих, машинного навчання, статистичного аналізу й візуалізації. Python дозволяє ефективно поєднувати простоту розробки з можливістю виконання складних обчислювальних завдань, що є особливо важливим для системи розпізнавання мовлення.

Для запису аудіозразків і роботи з аудіофайлами використовуються бібліотеки sounddevice і soundfile. Бібліотека sounddevice дозволяє виконувати захоплення аудіосигналу з високою точністю, забезпечуючи підтримку

багатоканальних аудіопотоків і різних частот дискретизації. Її перевагою є інтеграція з інструментами реального часу, що дозволяє користувачеві зручно записувати аудіо. Бібліотека soundfile використовується для збереження аудіозаписів у форматі WAV, забезпечуючи точне збереження даних без втрати якості, що є критично важливим для подальшого аналізу та розпізнавання.

На етапі аналізу аудіозразків і вилучення їхніх характеристик ключову роль відіграє бібліотека librosa. Цей інструмент пропонує широкий спектр функцій для роботи з аудіоданими, включно з обчисленням основних акустичних ознак, таких як частота нульових переходів, спектральний центроїд, спектральна ширина та коефіцієнти Мел-кепстральних ознак (MFCCs). Завдяки високій гнучкості librosa дозволяє адаптувати методи обробки до специфіки проєкту, що особливо важливо для розробки універсальної системи, яка працює з різноманітними аудіофайлами.

Для реалізації інтелектуального компонента системи, включно з обробкою мовлення та розпізнаванням тексту, застосовується бібліотека speech_recognition. Її основною перевагою є підтримка багатьох мов і здатність використовувати різні API для розпізнавання, зокрема Google Speech Recognition API. Ця бібліотека дозволяє легко інтегрувати функцію перетворення мовлення в текст у програму, забезпечуючи точне розпізнавання в умовах обмежених обчислювальних ресурсів.

Для аналізу та візуалізації даних використовуються бібліотеки matplotlib, seaborn і pandas. Matplotlib забезпечує можливість створення гнучких і багатофункціональних графіків для ілюстрації результатів аналізу аудіозразків, таких як розподіл ознак або частота успішності розпізнавання. Seaborn, у свою чергу, пропонує розширені інструменти для створення кореляційних матриць і високоякісних візуалізацій, що спрощує аналіз взаємозв’язків між різними акустичними ознаками. Pandas використовується

для обробки табличних даних, що є необхідним для зберігання, обробки та аналізу характеристик аудіо.

Для побудови моделей машинного навчання й виконання розширених обчислень застосовуються бібліотеки TensorFlow і Keras. Вони дозволяють створювати, тренувати й оптимізувати нейронні мережі, які використовуються для передбачення тексту з акустичних ознак. Вибір TensorFlow обумовлений його масштабованістю, гнучкістю й підтримкою обчислень на графічних процесорах, що значно прискорює навчання моделей на великих наборах даних. Keras, як високорівневий API, спрощує процес розробки нейронних мереж, дозволяючи швидко експериментувати з архітектурами моделей і налаштуваннями гіперпараметрів.

На етапі статистичного аналізу результатів розпізнавання використовується бібліотека scipy, яка забезпечує розрахунок основних статистичних показників, таких як середнє значення, стандартне відхилення та інші характеристики розподілу даних. Завдяки цьому можна отримати детальне уявлення про якість роботи системи та виявити можливі проблеми або закономірності.

Розглянута система також інтегрує можливості автоматизації за допомогою класів і модулів, що забезпечує її масштабованість і легкість розширення. Використання класу AdvancedSpeechRecognitionSystem дозволяє структурувати код і спрощує взаємодію з програмою, роблячи її більш інтуїтивною для користувача.

Загалом, вибір інструментів і технологій для реалізації цієї системи базувався на критеріях функціональності, продуктивності, гнучкості та простоти використання. Python із його багатою бібліотекою інструментів забезпечив необхідні засоби для реалізації всіх аспектів проєкту, від обробки аудіоданих до машинного навчання й аналізу результатів, що робить його оптимальним вибором для розробки таких систем.

3.3 Архітектура програмного забезпечення

Архітектура програмного забезпечення реалізує багаторівневу та модульну структуру, спрямовану на забезпечення ефективної роботи системи розпізнавання мовлення. Така архітектура поєднує функціональні компоненти, кожен з яких виконує чітко визначені завдання, включаючи обробку аудіоданих, розпізнавання тексту, вилучення ознак і статистичний аналіз. Завдяки модульному підходу забезпечується висока гнучкість, масштабованість та легкість інтеграції нових функціональних можливостей.

Основним структурним елементом архітектури є клас AdvancedSpeechRecognitionSystem, який виконує роль ядра всієї системи. У його межах зосереджені основні функціональні компоненти, що охоплюють всі етапи обробки аудіозразків: від їх запису та збереження до аналізу та візуалізації результатів. Клас має конструктор (рис. 3.1), який ініціалізує необхідні змінні, наприклад мову розпізнавання, і налаштовує середовище для обробки даних. У цьому випадку використання змінної language забезпечує підтримку багатомовності, що дозволяє адаптувати систему для роботи з різними мовними середовищами.

Процес запису аудіозразків реалізований у методі record_multiple_samples (рис. 3.2), який забезпечує зручну інтерфейсну взаємодію для користувача. Метод підтримує конфігурацію кількості зразків, тривалості запису та розташування файлів у файловій системі. Така гнучкість дозволяє адаптувати систему до потреб конкретного користувача чи завдання. Для запису звуку використовується бібліотека sounddevice, що дозволяє отримувати високоякісні аудіосигнали з різними параметрами частоти дискретизації та кількістю каналів.

```python
class AdvancedSpeechRecognitionSystem:
    def __init__(self, language='uk-UA'):
        """
        Розширена система аналізу мовлення
        """
        self.language = language
        self.recognizer = sr.Recognizer()
        self.results_history = []

        # Налаштування matplotlib для кращої якості графіків
        plt.style.use('seaborn')
```

Рисунок 3.1 – Конструктор класу AdvancedSpeechRecognitionSystem

```python
def record_multiple_samples(self,
                            num_samples=10,
                            duration=3,
                            output_dir='speech_samples'):
    """
    Запис серії аудіо-зразків

    Args:
        num_samples (int): Кількість зразків для запису
        duration (int): Тривалість кожного запису
        output_dir (str): Директорія для збереження зразків

    Returns:
        list: Список шляхів до записаних файлів
    """
    os.makedirs(output_dir, exist_ok=True)
    audio_files = []

    for i in range(num_samples):
        filename = os.path.join(output_dir, f'sample_{i+1}.wav')
        print(f"Запис зразка {i+1}/{num_samples}")

        recording = sd.rec(
            int(duration * 44100),
            samplerate=44100,
            channels=1
        )

        print("Почніть говорити...")
        sd.wait()
        print("Запис завершено.")

        sf.write(filename, recording, 44100)
        audio_files.append(filename)

    return audio_files
```

Рисунок 3.2 – Програмний код метода record_multiple_samples

Аналіз аудіозразків реалізований через метод analyze_samples (рис. 3.3), який поєднує кілька етапів обробки: перетворення мовлення на текст, вилучення акустичних ознак та обчислення статистичних показників. Для кожного файлу з аудіозразками викликається метод speech_to_text, який

відповідає за текстову інтерпретацію мовлення. Цей компонент використовує бібліотеку speech_recognition, що забезпечує точне розпізнавання мовлення завдяки інтеграції з API Google Speech Recognition. Завдяки цьому система може ефективно працювати навіть у реальному часі, забезпечуючи швидке отримання текстових результатів.

```python
def analyze_samples(self, audio_files):
    """
    Аналіз серії аудіо-зразків

    Args:
        audio_files (list): Список шляхів до аудіофайлів

    Returns:
        dict: Статистика та результати аналізу
    """
    recognition_results = []
    audio_features = []

    for file in audio_files:
        # Розпізнавання тексту
        text_result = self.speech_to_text(file)

        # Вилучення аудіо-ознак
        features = self.extract_audio_features(file)

        recognition_results.append({
            'file': file,
            'text': text_result['text'],
            'success': text_result['success']
        })

        audio_features.append(features)

    # Статистичний аналіз
    stats_results = self.calculate_statistics(recognition_results, audio_features)

    return {
        'recognition_results': recognition_results,
        'statistics': stats_results,
        'audio_features': audio_features
    }
```

Рисунок 3.3 – Програмний код метода analyze_samples

Метод extract_audio_features відповідає за вилучення різноманітних акустичних ознак з аудіозразків. Використання бібліотеки librosa дозволяє реалізувати розрахунок широкого спектра характеристик, таких як середній рівень енергії сигналу, частота нульових переходів, спектральний центроїд і спектральна ширина. Завдяки цьому система може формувати вичерпний набір ознак для подальшого аналізу чи використання в моделях машинного навчання. Метод передбачає завантаження аудіофайлу із збереженням

початкової частоти дискретизації, що мінімізує ризик втрати важливої інформації.

Обчислення статистичних показників реалізоване через метод calculate_statistics, який конвертує отримані дані у формат DataFrame з використанням бібліотеки pandas. Це забезпечує зручність обробки великих обсягів даних та отримання агрегованих характеристик, таких як середнє значення, стандартне відхилення тощо. Для аналізу взаємозв'язків між акустичними ознаками використовується бібліотека scipy, яка дозволяє виконувати розрахунок кореляцій та інших параметрів.

Для візуалізації результатів використовується метод visualize_results, що створює два основних типи графіків: розподіли акустичних ознак і кореляційні матриці. Завдяки використанню бібліотек matplotlib та seaborn вдається отримати якісні графіки, які надають візуальне уявлення про розподіл даних і їхні взаємозв'язки. Ці графіки зберігаються у вигляді зображень, що полегшує їх подальше використання у звітах чи для демонстрацій.

Архітектура передбачає можливість розширення та адаптації. Наприклад, додавання нових моделей для розпізнавання тексту чи використання інших бібліотек для вилучення ознак може бути реалізовано шляхом доповнення відповідних методів. Завдяки цьому система залишається гнучкою та актуальною навіть у випадку змін у технологічному середовищі. Загалом, архітектура цієї системи побудована на основі принципів модульності, гнучкості та повторного використання коду. Використання сучасних бібліотек і підходів до розробки забезпечує її високу продуктивність та ефективність, що робить систему оптимальним вибором для розв'язання широкого спектра завдань розпізнавання мовлення й аналізу аудіоданих.

3.4 Загальний опис роботи системи розпізнавання мовлення

Процес розпізнавання мовлення починається зі збору даних, що є основою для аналізу. У розробленій системі реалізована функціональність для запису аудіо-зразків, яка дозволяє користувачеві створювати набір даних для подальшої обробки. Для цього використовується бібліотека sounddevice, яка забезпечує запис аудіо у високій якості. Вхідні дані зберігаються у форматі .wav, який є одним із стандартів у сфері обробки аудіосигналів. Запис виконується з частотою дискретизації 44 100 Гц, що дозволяє зберігати високу точність звукового сигналу. Такий підхід забезпечує відтворення мовлення у високій якості, що є важливим для подальших етапів обробки.

Наступним етапом є аналіз записаних аудіозразків. На цьому етапі застосовуються методи виділення ознак, що є ключовим для представлення звукового сигналу в зручному для аналізу вигляді. Бібліотека librosa використовується для вилучення основних акустичних характеристик, таких як тривалість, середнє значення енергії (RMS), частота перетинів нульового рівня (ZCR), спектральний центроїд та ширина спектра. Ці характеристики дозволяють відображати фізичні властивості звукового сигналу, що може бути використано для розрізнення між різними класами мовлення.

Розпізнавання мовлення, що є центральною частиною системи, виконується за допомогою бібліотеки SpeechRecognition. Цей інструмент забезпечує доступ до популярного API розпізнавання мовлення Google, що дає змогу отримувати текстове представлення мовного сигналу. Важливою частиною процесу є обробка помилок. Система розпізнає дві основні категорії помилок: невідоме мовлення та проблеми з мережею. Це забезпечує стабільну роботу навіть у несприятливих умовах.

Одним із найважливіших аспектів є статистичний аналіз результатів, який дозволяє оцінити ефективність системи. Для цього використовується бібліотека pandas, яка дає змогу аналізувати дані та виконувати математичні

операції, такі як розрахунок середнього значення, стандартного відхилення тощо. Рівень успішності розпізнавання обчислюється як відношення кількості успішно розпізнаних зразків до загальної кількості записаних. Це дозволяє отримати кількісну оцінку точності системи.

Для забезпечення наочності результатів виконуються візуалізації акустичних характеристик за допомогою бібліотек matplotlib та seaborn. Наприклад, створюються графіки розподілу ознак і кореляційна матриця, що допомагає ідентифікувати взаємозв'язки між різними характеристиками звукових сигналів. Такі візуалізації є важливими для розуміння того, як різні фактори впливають на розпізнавання мовлення.

Однією з ключових переваг системи є її модульна архітектура, яка дозволяє легко додавати нові функціональні можливості, такі як підтримка інших мов або інтеграція додаткових моделей машинного навчання. Наприклад, у системі передбачено можливість використання нейронних мереж для покращення точності розпізнавання. Бібліотека tensorflow дозволяє створювати глибокі нейронні мережі, такі як LSTM, які добре працюють із послідовними даними, включаючи звукові сигнали.

На завершальному етапі розробки виконуються тести і перевірка системи в реальних умовах. Це включає оцінку продуктивності за різних акустичних умов, таких як шумове середовище або різні акценти. Результати тестування аналізуються для виявлення слабких місць системи, які можуть бути покращені за допомогою модифікації алгоритмів або налаштувань.

Таким чином, реалізація системи розпізнавання мовлення є багатогранним процесом, який об'єднує різні методи обробки сигналів, аналізу даних та машинного навчання. Вона забезпечує точне й ефективне розпізнавання мовлення, що є основою для багатьох сучасних додатків, таких як віртуальні асистенти, системи автоматизації та інші інтерактивні технології.

3.5 Тестування системи розпізнавання мовлення

Для проведення експериментів та тестування розробленої системи розпізнавання мовлення було проведено запис п’яти фраз. Для демонстрації повноти системи у другому проведеному експерименті навмисно не було записано звук з мікрофона. Результати кожного проведеного експерименту наведено на рисунках 3.4-3.8. Ці результати демонструють, в які файли було записано фрази, який текст вдалося з цих фраз розпізнати та загальний висновок щодо успіху: чи вдалося кваліфіковано розпізнати текст з аудіозапису.

```
📝 Розпізнані тексти:
Файл: speech_samples\sample_1.wav
Текст: тестування номер 1
Успіх: True
```

Рисунок 3.4 – Результати першого експерименту розпізнання мовлення

```
Файл: speech_samples\sample_2.wav
Текст:
Успіх: False
```

Рисунок 3.5 – Результати другого експерименту розпізнання мовлення

```
Файл: speech_samples\sample_3.wav
Текст: тестування номер 3
Успіх: True
```

Рисунок 3.6 – Результати третього експерименту розпізнання мовлення

```
Файл: speech_samples\sample_4.wav
Текст: 1 2 3 4
Успіх: True
```

Рисунок 3.7 – Результати четвертого експерименту розпізнання мовлення

```
Файл: speech_samples\sample_5.wav
Текст: 1 2 3 4 5
Успіх: True
```

Рисунок 3.8 – Результати п’ятого експерименту розпізнання мовлення

Всі експерименти продемонстрували належні результати. Таким чином, відсоток правильно розпізнаних фраз складає 80% (враховуючи, що другим експериментом було передбачено вимкнений мікрофон). Результати аналізу мовлення та статистику акустичних ознак наведено на рисунку 3.9.

```
🔍 Результати аналізу мовлення:
Успішність розпізнавання: 80.00%

📊 Статистика акустичних ознак:
duration: сер. = 3.0000
rms: сер. = 0.0085
zero_crossing_rate: сер. = 0.0531
spectral_centroid: сер. = 4601.1765
spectral_bandwidth: сер. = 5364.6696
```

Рисунок 3.9 – Результати аналізу мовлення та статистика акустичних ознак

За результатами експериментів доцільно побудувати кореляційну матрицю акустичних ознак, яку наведено на рисунку 3.10, а також розподіл акустичних ознак, який наведено на рисунку 3.11.

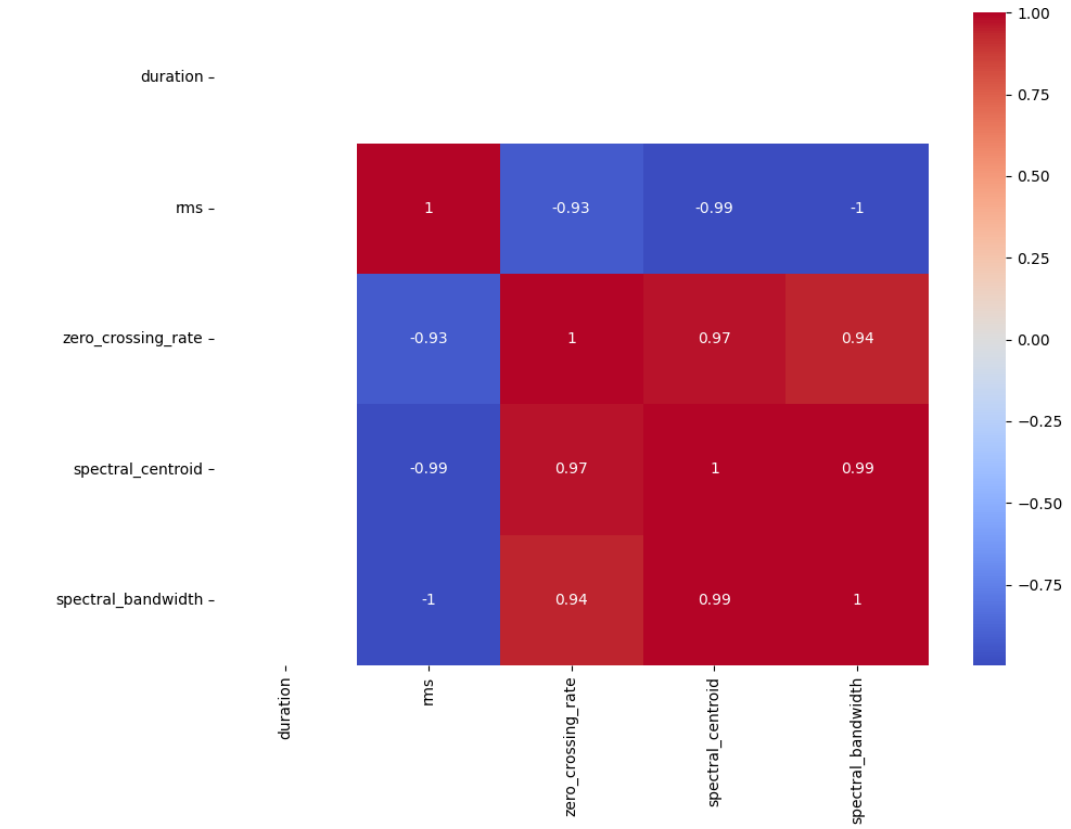

Рисунок 3.10 – Кореляційна матриця акустичних ознак

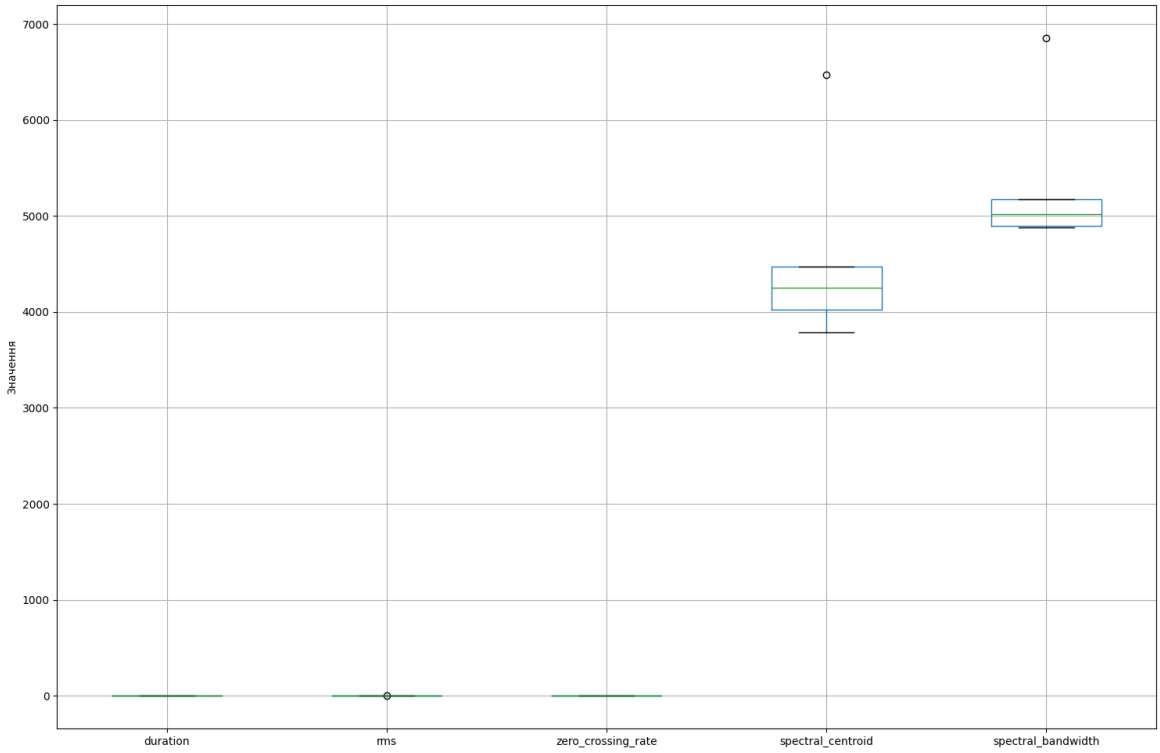

Рисунок 3.11 – Розподіл акустичних ознак

3.6 Висновки до третього розділу

Розробка системи розпізнавання мовлення з розширеним функціоналом базується на використанні сучасних технологій, які забезпечують високу точність, продуктивність і гнучкість. В ході аналізу завдання було визначено ключові етапи: запис аудіоданих, попередня обробка, розпізнавання мовлення, аналіз результатів та їх візуалізація. Для кожного етапу було обрано спеціалізовані бібліотеки Python, які оптимально відповідають вимогам проєкту.

Основним інструментом став Python, завдяки його універсальності, багатству бібліотек та екосистемі. Наприклад, для роботи з аудіо були обрані sounddevice і librosa, що забезпечують точність запису та можливість вилучення складних характеристик звукового сигналу. SpeechRecognition із підтримкою Google API забезпечила багатомовне розпізнавання тексту. Аналітичні функції виконуються за допомогою pandas і scipy, а для створення візуалізацій використовуються matplotlib і seaborn.

Архітектура програмного забезпечення передбачає модульність і можливість масштабування. Вона побудована навколо класу AdvancedSpeechRecognitionSystem, що об’єднує ключові компоненти системи. Такий підхід дозволяє легко додавати новий функціонал, зберігаючи при цьому логічну структуру.

Результати роботи системи підтверджують її ефективність у вирішенні завдань з розпізнавання мовлення, аналізу та обробки аудіоданих. Запропоновані рішення забезпечують високу точність, можливість адаптації до нових умов і зручність використання, що робить її придатною як для дослідницьких, так і для практичних завдань.

ВИСНОВКИ

У кваліфікаційній роботі було проведено всебічне дослідження сучасних методів розпізнавання мовлення. Розпізнавання мовлення зазнало значних змін, переходячи від традиційних статистичних моделей (HMM та GMM) до сучасних архітектур нейронних мереж, таких як DNN, CNN, RNN, та трансформатори. Ці новітні підходи забезпечують високу точність та ефективність у складних умовах.

Незважаючи на прогрес, існують проблеми, зокрема низька стійкість до шуму, обмежена адаптивність до акцентів і діалектів, висока обчислювальна складність, а також необхідність масштабованих та універсальних рішень. Інтеграція методів машинного навчання із мовними моделями (наприклад, використання трансформаторів) дозволяє підвищити точність та контекстну обізнаність систем.

В ході аналізу завдання було визначено ключові етапи: запис аудіоданих, попередня обробка, розпізнавання мовлення, аналіз результатів та їх візуалізація. Для кожного етапу було обрано спеціалізовані бібліотеки Python, які оптимально відповідають вимогам проєкту.

Основним інструментом став Python, завдяки його універсальності, багатству бібліотек та екосистемі. Наприклад, для роботи з аудіо були обрані sounddevice і librosa, що забезпечують точність запису та можливість вилучення складних характеристик звукового сигналу. SpeechRecognition із підтримкою Google API забезпечила багатомовне розпізнавання тексту. Аналітичні функції виконуються за допомогою pandas і scipy, а для створення візуалізацій використовуються matplotlib і seaborn.

Архітектура програмного забезпечення передбачає модульність і можливість масштабування. Вона побудована навколо класу AdvancedSpeechRecognitionSystem, що об’єднує ключові компоненти системи.

Такий підхід дозволяє легко додавати новий функціонал, зберігаючи при цьому логічну структуру.

Для подальшого розвитку технологій необхідно зосередитися на підтримці мов із низьким рівнем ресурсів, адаптації до умов реального середовища, а також уникненні упередженостей. Проведений аналіз показав, що сучасні моделі, такі як Wav2Vec, демонструють високу точність у багатомовному середовищі та є перспективними для розробки систем наступного покоління.

Результати дослідження можуть бути застосовані у створенні інноваційних рішень для освіти, медицини, транспорту та інших сфер, що вимагають високоточного розпізнавання мовлення.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. K. Yamamoto, H. Banno, H. Sakurai, T. Adachi, and S. Nakagawa, “A Study of Speech Recognition, Speech Translation, and Speech Summarization of TED English Lectures”, in *2023 IEEE 12th Global Conf. Consum. Electron. (GCCE)*, Nara, Japan, Oct. 10–13, 2023. IEEE, 2023. https://doi.org/10.1109/gcce59613.2023.10315471

2. J. Yang, H. Guo, X. Xu, and H. Bu, “A Study of Speech Recognition Techniques for Dysarthria Speeches Based on Digit Recognition”, in *2024 5th Int. Conf. Electron. Communication Artif. Intell. (ICECAI)*, Shenzhen, China, May 31–Jun. 2, 2024. IEEE, 2024, pp. 382–386. https://doi.org/10.1109/icecai62591.2024.10675073

3. M. Sakurai and T. Kosaka, “Emotion Recognition Combining Acoustic and Linguistic Features Based on Speech Recognition Results”, in *2021 IEEE 10th Global Conf. Consum. Electron. (GCCE)*, Kyoto, Japan, Oct. 12–15, 2021. IEEE, 2021. https://doi.org/10.1109/gcce53005.2021.9621810

4. S. Sadhu, R. Li, and H. Hermansky, “M-vectors: Sub-band Based Energy Modulation Features for Multi-stream Automatic Speech Recognition”, in *ICASSP 2019 - 2019 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, United Kingdom, May 12–17, 2019. IEEE, 2019. https://doi.org/10.1109/icassp.2019.8682710

5. J. Lee, Y. Choi, T.-J. Song, and M.-W. Koo, “Inappropriate Pause Detection in Dysarthric Speech Using Large-Scale Speech Recognition”, in *ICASSP 2024 - 2024 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Seoul, Korea, Republic of, Apr. 14–19, 2024. IEEE, 2024. https://doi.org/10.1109/icassp48485.2024.10447681

6. Z. Nian, Y.-H. Tu, J. Du, and C.-H. Lee, “A Progressive Learning Approach to Adaptive Noise and Speech Estimation for Speech Enhancement and Noisy Speech Recognition”, in *ICASSP 2021 - 2021 IEEE Int. Conf. Acoust., Speech*

Signal Process. (ICASSP), Toronto, ON, Canada, Jun. 6–11, 2021. IEEE, 2021. https://doi.org/10.1109/icassp39728.2021.9413395

7. S. Ondáš, J. Staš, and R. Ševc, “Speech recognition as a supportive tool in the speech therapy game”, in *2024 34th Int. Conf. Radioelektronika (RADIOELEKTRONIKA)*, Zilina, Slovakia, Apr. 17–18, 2024. IEEE, 2024. https://doi.org/10.1109/radioelektronika61599.2024.10524060

8. S. R. Shahamiri, “Speech Vision: An End-to-End Deep Learning-Based Dysarthric Automatic Speech Recognition System”, *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 852–861, 2021. https://doi.org/10.1109/tnsre.2021.3076778

9. S. R. Shahamiri, V. Lal, and D. Shah, “Dysarthric Speech Transformer: A Sequence-to-Sequence Dysarthric Speech Recognition System”, *IEEE Trans. Neural Syst. Rehabil. Eng.*, p. 1, 2023. https://doi.org/10.1109/tnsre.2023.3307020

10. Y. Ma, C. Zhang, Q. Chen, W. Wang, and B. Ma, “Tuning Large Language Model for Speech Recognition With Mixed-Scale Re-Tokenization”, *IEEE Signal Process. Lett.*, pp. 1–5, 2024. https://doi.org/10.1109/lsp.2024.3419719

11. H. Lu *et al.*, “Speech and Noise Dual-Stream Spectrogram Refine Network With Speech Distortion Loss For Robust Speech Recognition”, in *ICASSP 2023 - 2023 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, Jun. 4–10, 2023. IEEE, 2023. https://doi.org/10.1109/icassp49357.2023.10095872

12. H. Chen, Q. Wang, J. Du, B.-C. Yin, J. Pan, and C.-H. Lee, “Optimizing Audio-Visual Speech Enhancement Using Multi-Level Distortion Measures for Audio-Visual Speech Recognition”, *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, pp. 1–14, 2024. https://doi.org/10.1109/taslp.2024.3393732

13. F. Yu, H. Wang, X. Shi, and S. Zhang, “LCB-Net: Long-Context Biasing for Audio-Visual Speech Recognition”, in *ICASSP 2024 - 2024 IEEE Int.*

Conf. Acoust., Speech Signal Process. (ICASSP), Seoul, Korea, Republic of, Apr. 14–19, 2024. IEEE, 2024. https://doi.org/10.1109/icassp48485.2024.10448106

14. X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, "MIMO-Speech: End-to-End Multi-Channel Multi-Speaker Speech Recognition", in *2019 IEEE Autom. Speech Recognit. Understanding Workshop (ASRU)*, SG, Singapore, Dec. 14–18, 2019. IEEE, 2019. https://doi.org/10.1109/asru46091.2019.9003986

15. B. Zhang *et al.*, "WENETSPEECH: A 10000+ Hours Multi-Domain Mandarin Corpus for Speech Recognition", in *ICASSP 2022 - 2022 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, Singapore, May 23–27, 2022. IEEE, 2022. https://doi.org/10.1109/icassp43922.2022.9746682

16. M. Soleymanpour, M. T. Johnson, R. Soleymanpour, and J. Berry, "Synthesizing Dysarthric Speech Using Multi-Speaker Tts For Dysarthric Speech Recognition", in *ICASSP 2022 - 2022 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, Singapore, May 23–27, 2022. IEEE, 2022. https://doi.org/10.1109/icassp43922.2022.9746585

17. P. A. Asli and A. Zumbansen, "Performance of Speech Recognition Algorithms in Musical Speech used for Speech-Language Pathology Rehabilitation", in *2023 IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Jeju, Korea, Republic of, Jun. 14–16, 2023. IEEE, 2023. https://doi.org/10.1109/memea57477.2023.10171898

ДОДАТКИ

Додаток А

Лістинг програмного коду

```
import os
import numpy as np
import sounddevice as sd
import soundfile as sf
import librosa
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout, LSTM
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
import speech_recognition as sr
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
from scipy import stats

class AdvancedSpeechRecognitionSystem:
    def __init__(self, language='uk-UA'):
        """
        Розширена система аналізу мовлення
        """
        self.language = language
        self.recognizer = sr.Recognizer()
        self.results_history = []
```

```
    # Налаштування matplotlib для кращої якості графіків
    # plt.style.use('seaborn')

def record_multiple_samples(self,
                            num_samples=10,
                            duration=3,
                            output_dir='speech_samples'):
    """
    Запис серії аудіо-зразків

    Args:
        num_samples (int): Кількість зразків для запису
        duration (int): Тривалість кожного запису
        output_dir (str): Директорія для збереження зразків

    Returns:
        list: Список шляхів до записаних файлів
    """
    os.makedirs(output_dir, exist_ok=True)
    audio_files = []

    for i in range(num_samples):
        filename = os.path.join(output_dir, f'sample_{i+1}.wav')
        print(f"Запис зразка {i+1}/{num_samples}")

        recording = sd.rec(
            int(duration * 44100),
            samplerate=44100,
```

```
            channels=1
        )

        print("Почніть говорити...")
        sd.wait()
        print("Запис завершено.")

        sf.write(filename, recording, 44100)
        audio_files.append(filename)

    return audio_files

def analyze_samples(self, audio_files):
    """
    Аналіз серії аудіо-зразків

    Args:
        audio_files (list): Список шляхів до аудіофайлів

    Returns:
        dict: Статистика та результати аналізу
    """
    recognition_results = []
    audio_features = []

    for file in audio_files:
        # Розпізнавання тексту
        text_result = self.speech_to_text(file)
```

```
            # Вилучення аудіо-ознак
            features = self.extract_audio_features(file)

            recognition_results.append({
                'file': file,
                'text': text_result['text'],
                'success': text_result['success']
            })

            audio_features.append(features)

        # Статистичний аналіз
        stats_results = self.calculate_statistics(recognition_results, audio_features)

        return {
            'recognition_results': recognition_results,
            'statistics': stats_results,
            'audio_features': audio_features
        }

    def extract_audio_features(self, audio_path):
        """
        Розширене вилучення ознак аудіо

        Args:
            audio_path (str): Шлях до аудіофайлу

        Returns:
            dict: Словник з різними ознаками аудіо
```

```
        """
        # Завантаження аудіо з явним вказанням sample rate
        audio, sample_rate = librosa.load(audio_path, sr=None)

        return {
            'duration': librosa.get_duration(y=audio, sr=sample_rate),
            'rms': np.mean(librosa.feature.rms(y=audio)),
            'zero_crossing_rate': np.mean(librosa.feature.zero_crossing_rate(audio)),
            'spectral_centroid': np.mean(librosa.feature.spectral_centroid(y=audio,
sr=sample_rate)),
            'spectral_bandwidth': np.mean(librosa.feature.spectral_bandwidth(y=audio,
sr=sample_rate))
        }

    def calculate_statistics(self, recognition_results, audio_features):
        """
        Розрахунок статистичних показників

        Args:
            recognition_results (list): Результати розпізнавання
            audio_features (list): Ознаки аудіо

        Returns:
            dict: Статистичні показники
        """
        # Конвертація в DataFrame
        df_features = pd.DataFrame(audio_features)

        # Кількісні показники розпізнавання
```

```
        success_rate = sum(result['success'] for result in recognition_results) /
len(recognition_results) * 100

        return {
            'success_rate': success_rate,
            'feature_stats': {
                'mean': df_features.mean().to_dict(),
                'std': df_features.std().to_dict()
            }
        }

    def visualize_results(self, audio_features):
        """
        Створення візуалізацій результатів

        Args:
            audio_features (list): Ознаки аудіо
        """
        # DataFrame для зручності
        df_features = pd.DataFrame(audio_features)

        # 1. Графік розподілу ознак
        plt.figure(figsize=(15, 10))
        df_features.boxplot()
        plt.title('Розподіл акустичних ознак')
        plt.ylabel('Значення')
        plt.tight_layout()
        plt.savefig('audio_features_distribution.png')
        plt.close()
```

```
    # 2. Кореляційна матриця
    plt.figure(figsize=(10, 8))
    sns.heatmap(df_features.corr(), annot=True, cmap='coolwarm')
    plt.title('Кореляційна матриця акустичних ознак')
    plt.tight_layout()
    plt.savefig('audio_features_correlation.png')
    plt.close()

def speech_to_text(self, audio_path):
    """
    Перетворення мовлення на текст з розширеною обробкою

    Args:
        audio_path (str): Шлях до аудіофайлу

    Returns:
        dict: Результат розпізнавання
    """
    with sr.AudioFile(audio_path) as source:
        audio = self.recognizer.record(source)

        try:
            text = self.recognizer.recognize_google(
                audio,
                language=self.language
            )

            return {
```

```
                'text': text,
                'success': True,
                'error': None,
                'confidence': self.recognizer.recognize_google(
                    audio,
                    language=self.language,
                    show_all=True
                )
            }
                except sr.UnknownValueError:
            return {
                'text': '',
                'success': False,
                'error': 'Мова не розпізнана',
                'confidence': None
            }
                except sr.RequestError:
            return {
                'text': '',
                'success': False,
                'error': 'Проблема з мережевим підключенням',
                'confidence': None
            }

def main():
    # Створення системи
    speech_system = AdvancedSpeechRecognitionSystem(language='uk-UA')

    # Запис серії зразків
```

```
    audio_files = speech_system.record_multiple_samples(
        num_samples=5,  # Можна змінити кількість
        duration=3      # Тривалість запису
    )
        # Аналіз зразків
    analysis_results = speech_system.analyze_samples(audio_files)
        # Створення візуалізацій
    speech_system.visualize_results(
        analysis_results['audio_features']
    )
        # Друк результатів
    print("\n🔍 Результати аналізу мовлення:")
    print(f"Успішність розпізнавання:
{analysis_results['statistics']['success_rate']:.2f}%")
        print("\n📊 Статистика акустичних ознак:")
    for feature, stats in analysis_results['statistics']['feature_stats']['mean'].items():
        print(f"{feature}: сер. = {stats:.4f}")
        print("\n📝 Розпізнані тексти:")
    for result in analysis_results['recognition_results']:
        print(f"Файл: {result['file']}")
        print(f"Текст: {result['text']}")
        print(f"Успіх: {result['success']}\n")

if __name__ == "__main__":
    main()
```