Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Кібербезпеки
(повна назва)

Кафедра Інформаційно-мережної інженерії
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіті перший (бакалаврський)

Дослідження нейромережевих методів розпізнавання мовлення за аудіосигналом та візуальними даними обличчя диктора
(тема)

Виконав:
здобувач 3 року навчання,
групи ТРІМІу-23-1
Владислав Максименко
(власне ім'я, прізвище)

Спеціальність 172 Електронні комунікації та радіотехніка
(код і повна назва спеціальності)

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

Освітня програма Інформаційно-мережна інженерія
(повна назва освітньої програми)

Керівник: доц. каф. ІМІ Сергій Омельченко
(посада, власне ім'я, прізвище)

Допускається до захисту

Зав. кафедри _____ Микола Москалець
(підпис) (власне ім'я, прізвище)

2026 р.

Не містить відомостей, заборонених до відкритого публікування

Студент _____ / Владислав Максименко /
(підпис) (власне ім'я, прізвище)

Керівник _____ / Сергій Омельченко /
(підпис) (власне ім'я, прізвище)

Харківський національний університет радіоелектроніки

Факультет Кібербезпеки

Кафедра Інформаційно-мережної інженерії

Рівень вищої освіти перший (бакалаврський)

Спеціальність 172 Електронні комунікації та радіотехніка

(код і повна назва)

Тип програми освітньо-професійна

Освітня програма Інформаційно-мережна інженерія

(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
« » травня 2026 р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Максименко Владиславу Юрійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження нейромережевих методів розпізнавання мовлення за аудіосигналом та візуальними даними обличчя диктора

затверджена наказом по університету від« 20» травня 2026 р. № 415 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 20 червня 2026 р.

3. Вхідні дані до роботи Виконати огляд методів розпізнавання мовлення людини. Розглянути особливості оцінювання мовленевих та візуальних ознак. Розглянути методи класифікації. Дослідити системи розпізнавання.

4. Перелік питань, що потрібно опрацювати у роботі Вступ

1. Огляд методів розпізнавання мовлення
2. Методи розпізнаванням слів мови
3. Розробка програмного модуля аудіовізуальної обробки мовлення
4. Тестування та аналіз результатів роботи программного модуля

Висновки

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри) слайди презентації в форматі Power Point (назва та мета роботи, методи, алгоритм, висновки)

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

Найменування розділу	Консультант (посада, власне ім'я, прізвище)	Позначка консультанта про виконання розділу	
		підпис	дата
Основна частина	доцент. каф. ІМІ Сергій Омельченко		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Вступ	23.05	виконано
2	Огляд методів розпізнавання мовлення	23.05-05.06	виконано
3	Методи розпізнаванням слів мови	06.06-12.06	виконано
4	Розробка програмного модуля аудіовізуальної обробки мовлення	13.06-15.06	виконано
5	Тестування та аналіз результатів роботи программного модуля	15.06-17.06	виконано
6	Висновки	17.06-19.06	виконано
7	Оформлення пояснювальної записки	20.06	виконано

Дата видачі завдання 20 травня 2026 р.

Здобувач _____
(підпис)

Керівник роботи _____ доц. каф. ІМІ Сергій Омельченко
(підпис) (посада, власне ім'я, прізвище)

РЕФЕРАТ

Пояснювальна записка: 77 с., 10 рис., 15 джерел, 3 додатка.

Об’єкт дослідження — процес автоматичного розпізнавання мовлення з використанням аудіо- та відеоданих.

Мета роботи — дослідження нейромережевих методів розпізнавання мовлення за аудіосигналом та візуальними даними обличчя диктора, а також розробка програмного модуля аудіовізуальної обробки мовлення.

Досліджуються сучасні нейромережеві методи автоматичного розпізнавання мовлення, підходи до візуального розпізнавання мовлення за рухом губ та аудіовізуальні методи обробки мовної інформації. У роботі розглянуто моделі DeepSpeech, DeepSpeech 2, wav2vec 2.0, Whisper, LipNet та AV-HuBERT, а також засоби OpenCV, MediaPipe Face Mesh і MoviePy для програмної реалізації.

У результаті роботи розроблено програмний модуль мовою Python, який витягує аудіодоріжку з відеофайлу, виконує розпізнавання мовлення за допомогою Whisper, обробляє відеокадри, визначає ключові точки обличчя та виділяє область губ диктора. Тестування показало, що модуль коректно формує текстовий результат за аудіоканалом і зберігає візуальні фрагменти області губ для подальшого аналізу.

АВТОМАТИЧНЕ РОЗПІЗНАВАННЯ МОВЛЕННЯ, НЕЙРОННІ МЕРЕЖІ, WHISPER, OPENCV, MEDIAPIPE FACE MESH, АУДІОВІЗУАЛЬНА ОБРОБКА, ОБЛАСТЬ ГУБ.

THE ABSTRACT

Explanatory note: 77 p., 10 fig., 15 sources, 3 app.

The object of the study is the process of automatic speech recognition using audio and video data.

The purpose of the work is to study neural network methods for speech recognition based on an audio signal and visual data of the speaker’s face, as well as to develop a software module for audiovisual speech processing.

The work studies modern neural network methods for automatic speech recognition, approaches to visual speech recognition based on lip movements, and audiovisual methods for processing speech information. The paper considers DeepSpeech, DeepSpeech 2, wav2vec 2.0, Whisper, LipNet and AV-HuBERT models, as well as OpenCV, MediaPipe Face Mesh and MoviePy tools for software implementation.

As a result of the work, a Python software module was developed. It extracts an audio track from a video file, performs speech recognition using Whisper, processes video frames, detects facial landmarks and extracts the speaker’s lip region. Testing showed that the module correctly forms a text result through the audio channel and saves visual fragments of the lip region for further analysis.

AUTOMATIC SPEECH RECOGNITION, NEURAL NETWORKS, WHISPER, OPENCV, MEDIAPIPE FACE MESH, AUDIOVISUAL PROCESSING, LIP REGION.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ 8
ВСТУП 9
1 ОГЛЯД МЕТОДІВ РОЗПІЗНАВАННЯ МОВЛЕННЯ 11
1.1 Основні принципи автоматичного розпізнавання мовлення 11
1.2 Нейромережеві методи розпізнавання мовлення 15
1.3 Методи візуального розпізнавання мовлення за рухом губ 19
2 МЕТОДИ РОЗПІЗНАВАННЯМ СЛІВ МОВИ 23
2.1 Аналіз сучасних моделей автоматичного розпізнавання мовлення 23
2.2 Аналіз методів обробки візуальних даних обличчя диктора 28
2.3 Порівняння аудіальних, візуальних та аудіовізуальних підходів 32
3 РОЗРОБКА ПРОГРАМНОГО МОДУЛЯ АУДІОВІЗУАЛЬНОЇ ОБРОБКИ МОВЛЕННЯ 38
3.1 Розробка архітектури програмної системи розпізнавання мовлення 38
3.2 Реалізація програмного модуля на Python 41
4 ТЕСТУВАННЯ ТА АНАЛІЗ РЕЗУЛЬТАТІВ РОБОТИ ПРОГРАМНОГО МОДУЛЯ 44
4.1 Організація тестування програмного модуля 44
4.2 Аналіз результатів тестування програмного модуля 47
ВИСНОВКИ 51
ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ 55
ДОДАТОК А. СЛАЙДИ ПРЕЗЕНТАЇ 57
ДОДАТОК Б. КОД 65
ДОДАТОК В. ТЕЗИ КОНФЕРЕНЦІЇ 70

ПЕРЕЛІК СКОРОЧЕНЬ

ASR – (Automatic Speech Recognition) – автоматичне розпізнавання мовлення;

VSR – (Visual Speech Recognition) – візуальне розпізнавання мовлення;

AVSR – (Audio-Visual Speech Recognition) – аудіовізуальне розпізнавання мовлення;

CNN – (Convolutional Neural Network) – згорткова нейронна мережа;

RNN – (Recurrent Neural Network) – рекурентна нейронна мережа;

LSTM – (Long Short-Term Memory) – мережа довгої короткочасної пам’яті;

GRU – (Gated Recurrent Unit) – рекурентний блок із вентилями;

CTC – (Connectionist Temporal Classification) – класифікація часових послідовностей;

MFCC – (Mel-Frequency Cepstral Coefficients) – мел-частотні кепстральні коефіцієнти;

ROI – (Region of Interest) – область інтересу;

OpenCV – (Open Source Computer Vision Library) – бібліотека комп’ютерного зору;

AV-HuBERT – (Audio-Visual Hidden-Unit BERT) – аудіовізуальна модель для формування спільного представлення мовлення.

ВСТУП

Технології автоматичного розпізнавання мовлення протягом останніх років стали невід'ємною частиною багатьох програмних продуктів і технічних рішень. Голосові помічники, системи керування мобільними пристроями, засоби автоматичного створення текстових записів та інтелектуальні інформаційні сервіси активно використовують алгоритми аналізу мовлення для взаємодії з користувачем. Зростання обсягів цифрової інформації та розвиток методів штучного інтелекту сприяли суттєвому підвищенню можливостей таких систем.

У більшості випадків розпізнавання мовлення здійснюється на основі аналізу аудіосигналу. За сприятливих умов цей підхід забезпечує високу точність розпізнавання слів і фраз мови. Разом з тим на практиці нерідко виникають ситуації, коли якість звуку погіршується через сторонній акустичний шум, особливості приміщення або технічні обмеження обладнання. У результаті ефективність системи розпізнавання може суттєво знижуватися.

Одним зі способів підвищення надійності розпізнавання є використання додаткових джерел інформації від диктора. Людина під час сприйняття мовни людини звертає увагу не лише на звук, але й на артикуляцію, рухи губ та міміку обличчя. Подібний підхід може бути реалізований і в комп'ютерних системах розпізнавання. Аналіз візуальних даних дозволяє отримати додаткові ознаки мовлення, які можуть використовуватися в системі сумісно з аудіосигналом.

Окремий інтерес для обробки та прийняття рішень становлять сучасні нейромережеві методи. Застосування згорткових мереж, рекурентних архітектур і трансформерів дає можливість суттево покращити якість автоматичного розпізнавання мовлення порівняно з традиційними методами. Крім того, останніми роками активно розвиваються мультимодальні системи, де поєднується обробка аудіо та відео в одній моделі.

Актуальність теми пов’язана з розвитком нейромережевих систем розпізнавання мовлення та потребою у використанні додаткових каналів інформації під час обробки та розпізнавання. У реальних умовах аудіозапис не

завжди має достатню якість та повноту інформації, а тому аналіз візуальних даних обличчя диктора може доповнювати аудіальний канал і надавати додаткову інформацію про артикуляцію диктора.

Метою кваліфікаційної роботи є дослідження нейромережевих методів розпізнавання мовлення за аудіосигналом та візуальними даними обличчя диктора, а також розробка програмного забезпечення для обробки відеофайлу з мовленням.

Для досягнення поставленої мети необхідно розглянути особливості методів автоматичного розпізнавання мовлення, проаналізувати існуючі сучасні нейромережеві архітектури та дослідити підходи до обробки аудіосигналів і візуальних даних обличчя людини. Також потрібно виконати існуючі методи аудіального, візуального та аудіовізуального підходів і реалізувати програмний код засобами мови Python.

Об’єктом дослідження є процес автоматичного розпізнавання мовлення із використанням аудіо- та відеоданих.

Предметом дослідження є нейромережеві методи, алгоритми та програмні засоби, що застосовуються для розпізнавання мовлення за аудіосигналом і візуальними даними обличчя диктора.

Під час виконання роботи використовувалися методи аналізу наукових джерел, порівняння сучасних підходів до розпізнавання мовлення, методи машинного навчання, комп’ютерного зору та програмної реалізації алгоритмів засобами Python.

Практичне значення отриманих результатів полягає у створенні програмного модуля, який обробляє вхідний відеофайл, виділяє з нього аудіодоріжку, виконує розпізнавання мовлення за допомогою Whisper та обробляє відеокадри для визначення області губ диктора. Такий модуль може бути використаний для подальшої перевірки якості роботи аудіального і візуального каналів під час обробки мовної інформації.

1 ОГЛЯД МЕТОДІВ РОЗПІЗНАВАННЯ МОВЛЕННЯ

1.1 Основні принципи автоматичного розпізнавання мовлення

Автоматичне розпізнавання мовлення є одним із найбільш поширених напрямів розвитку сучасних інтелектуальних інформаційних систем. Основне призначення таких систем полягає у перетворенні усного мовлення людини в текстову форму, придатну для подальшого зберігання, аналізу або використання іншими програмними засобами. В останні роки інтерес до даної технології значно зріс, так як вона стала важливим елементом голосових асистентів, систем дистанційного керування, сервісів автоматичного перекладу та багатьох інших прикладних рішень.

Процес розпізнавання мовлення являє собою складну задачу, під час якої комп'ютерна система повинна визначити, які саме слова були вимовлені користувачем. На відміну від текстової інформації, мовний сигнал є безперервним у часі, хоча мовленева активність диктора обмежена сукупністю інтервалів часу та сигнал є випадковим і містить значну кількість додаткових факторів, в тому числі акустичні завади та шуми в каналі передачі. На результат розпізнавання можуть впливати індивідуальні особливості голосу, темп мовлення, емоційний стан людини, якість використаного мікрофона, рівень навколишнього шуму та інші обставини.

Перші системи автоматичного розпізнавання мовлення будувалися на основі статистичних методів та порівняння звукових шаблонів. Подібні рішення могли працювати лише з обмеженим словниковим запасом і вимагали ретельного налаштування для кожного користувача. З розвитком обчислювальної техніки та накопиченням великих обсягів мовних даних ситуація поступово змінилася. Сучасні системи здатні обробляти природне мовлення великої кількості людей без попереднього навчання під конкретного диктора.

У загальному випадку роботу системи розпізнавання мовлення можна подати як послідовність кількох взаємопов'язаних етапів. Спочатку виконується отримання аудіосигналу з мікрофона або готового аудіофайлу. Після цього сигнал проходить попередню обробку, під час якої можуть виконуватися фільтрація шумів, нормалізація гучності та інші операції, спрямовані на

покращення якості даних. Далі система виділяє характерні ознаки мовлення, які використовуються для подальшого аналізу алгоритмами машинного навчання.

Важливим елементом сучасних систем є етап формування ознак. Безпосереднє використання сирого аудіосигналу часто є неефективним через великий обсяг даних та наявність надлишкової інформації. З цієї причини застосовуються спеціальні методи перетворення сигналу, які дозволяють виділити найбільш інформативні характеристики мовлення. Отримані ознаки передаються до моделі розпізнавання, яка визначає найбільш ймовірну текстову послідовність.

Суттєвий вплив на розвиток технологій розпізнавання мовлення справило використання нейронних мереж. На відміну від класичних статистичних підходів, нейромережеві моделі здатні автоматично виявляти складні закономірності у великих масивах даних. Це дозволило значно підвищити точність розпізнавання навіть у випадках, коли мовлення містить сторонні шуми або має складну структуру.

Особливого поширення набули моделі глибокого навчання, які можуть працювати з довгими мовними послідовностями та враховувати контекст висловлювання. Якщо ранні системи часто аналізували окремі слова або короткі фрази, то сучасні моделі здатні обробляти цілі речення та діалоги. Завдяки цьому результати розпізнавання стали більш природними та наближеними до людського сприйняття мовлення.

Незважаючи на значні досягнення у цій сфері, проблема автоматичного розпізнавання мовлення залишається актуальною. Складності виникають під час роботи в шумному середовищі, за наявності кількох співрозмовників або при використанні мовлення з різними акцентами. Саме тому останнім часом активно розвиваються мультимодальні підходи, де аудіодані доповнюються інформацією про рухи губ та інші візуальні ознаки мовця.

Автоматичне розпізнавання мовлення поєднує цифрову обробку сигналів, машинне навчання та елементи комп’ютерного зору, якщо система додатково використовує візуальні дані диктора. Загальна логіка роботи таких систем зводиться до послідовної обробки вхідного сигналу, виділення ознак і формування текстового результату. Узагальнену послідовність роботи сучасної системи автоматичного розпізнавання мовлення наведено на рисунку 1.1.

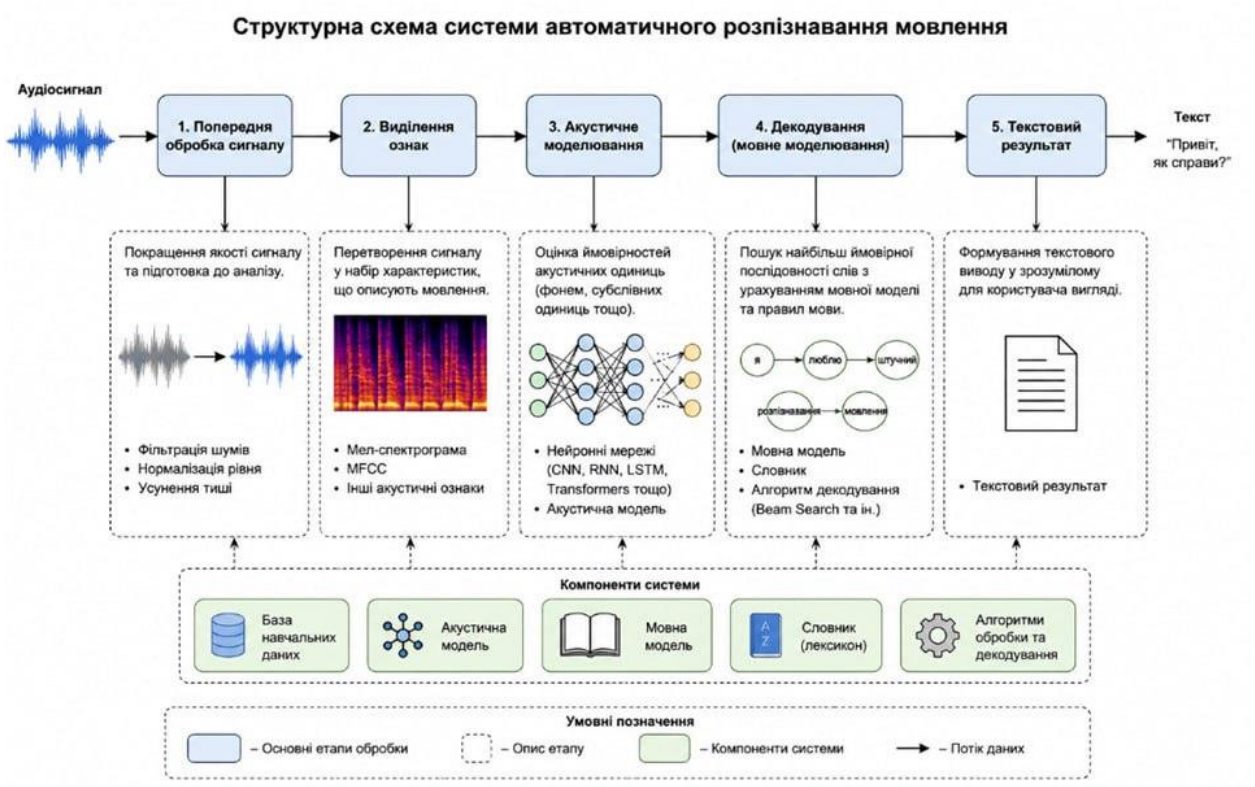


Рисунок 1.1 – Узагальнена послідовність роботи системи автоматичного розпізнавання мовлення

Як видно з рисунку 1.1, процес автоматичного розпізнавання мовлення має послідовну структуру. Спочатку система отримує аудіосигнал з мікрофона або готового аудіофайлу. Далі виконується попередня обробка сигналу, під час якої можуть застосовуватися фільтрація шумів, нормалізація рівня гучності та усунення зайвих фрагментів. Це дозволяє підготувати вхідні дані до подальшого аналізу.

Після попередньої обробки здійснюється виділення ознак мовного сигналу. На цьому етапі аудіосигнал перетворюється у набір числових характеристик, які описують його важливі параметри. До таких ознак можуть належати спектральні характеристики, мел-спектрограми, MFCC та інші параметри, що використовуються для подальшого розпізнавання.

Наступним етапом є акустичне моделювання та декодування мовлення. Для цього можуть використовуватися статистичні методи, а в сучасних системах все частіше застосовуються нейромережеві алгоритми. Вони визначають найбільш ймовірну текстову послідовність на основі отриманих характеристик мовного сигналу. Результатом роботи системи є текстове представлення

мовлення, яке може бути використане для пошуку інформації, керування програмним забезпеченням або взаємодії користувача з комп’ютерною системою.

Загальний принцип автоматичного розпізнавання мовлення полягає у виборі такої послідовності слів, яка найбільш імовірно відповідає отриманим акустичним ознакам мовного сигналу. У спрощеному вигляді цей процес можна подати так:

$$W_{rec} = \arg\max_{w} P(W \mid X) \qquad (1.1)$$

де W_rec — розпізнана послідовність слів; W — можлива послідовність слів; X — набір акустичних ознак, отриманих із мовного сигналу; P(W | X) — ймовірність відповідності певної послідовності слів вхідним акустичним ознакам. Такий запис відображає сутність задачі автоматичного розпізнавання мовлення, оскільки результатом роботи системи є текстова послідовність, яка найкраще відповідає вхідним даним.

Важливою характеристикою сучасних систем автоматичного розпізнавання мовлення є їх здатність працювати в реальних умовах експлуатації. На практиці мовний сигнал часто містить фонові шуми, відлуння приміщення або спотворення, пов’язані з якістю обладнання. Крім того, на точність розпізнавання впливають індивідуальні особливості диктора, темп мовлення, акцент та інтонація.

Отже, автоматичне розпізнавання мовлення можна розглядати як багатоступеневий процес, у якому якість результату залежить від підготовки вхідного сигналу, вибору ознак і роботи моделі розпізнавання. У реальних умовах на цей процес впливають шуми, якість запису, особливості вимови та інші зовнішні фактори. Саме тому подальший розвиток таких систем пов’язаний не лише з удосконаленням аудіальних моделей, а й з використанням додаткових джерел інформації, зокрема візуальних даних обличчя диктора.

1.2 Нейромережеві методи розпізнавання мовлення

Подальший розвиток автоматичного розпізнавання мовлення значною мірою пов’язаний із використанням нейромережевих методів. Це пояснюється тим, що мовний сигнал має складну структуру і залежить не тільки від окремих звуків, а й від їх послідовності, темпу вимови, інтонації та особливостей диктора. У класичних системах значна частина ознак задавалася наперед, тому якість роботи суттєво залежала від того, наскільки правильно були обрані параметри сигналу. Нейронні мережі змінили цей підхід, оскільки вони здатні самостійно знаходити корисні закономірності у великому обсязі даних.

У задачах розпізнавання мовлення нейромережа зазвичай працює не з самим аудіосигналом у початковому вигляді, а з його числовим представленням. Для цього сигнал перетворюють у набір ознак, наприклад у спектрограму або мел-спектрограму. Таке подання дозволяє бачити, як змінюється енергія сигналу в різних частотних діапазонах протягом часу. Для людини це виглядає як зображення, але для моделі це набір чисел, за якими можна визначати мовні одиниці.

Одним із поширених підходів є використання згорткових нейронних мереж. Спочатку CNN активно застосовувалися у комп’ютерному зорі, однак згодом їх почали використовувати і для обробки мовлення. Причина полягає в тому, що спектрограму аудіосигналу можна аналізувати подібно до зображення. У такому випадку згорткові шари виділяють локальні закономірності у частотно-часовій області. Це можуть бути зміни інтенсивності сигналу, переходи між звуками, характерні фрагменти вимови та інші акустичні ознаки.

Перевага згорткових мереж полягає в тому, що вони не потребують повністю ручного опису всіх ознак. На нижчих рівнях модель виявляє прості елементи сигналу, а на глибших шарах формує складніші представлення. Такий підхід добре підходить для первинного аналізу мовних даних. Водночас CNN не завжди достатньо добре враховують довгі часові залежності, тому в системах розпізнавання мовлення їх часто поєднують з іншими архітектурами [1].

Для аналізу послідовностей тривалий час використовувалися рекурентні нейронні мережі. Їх особливість полягає у тому, що вони обробляють дані послідовно і можуть враховувати попередній контекст. Для мовлення це важливо, оскільки звук, склад або слово не існують окремо від загальної фрази.

Наприклад, правильне розпізнавання певного слова може залежати від того, які слова були вимовлені перед ним.

Прості рекурентні мережі мають обмеження при роботі з довгими послідовностями. Через це у практичних системах частіше застосовуються LSTM та GRU. Мережі LSTM мають спеціальні механізми пам'яті, які дозволяють зберігати важливу інформацію протягом більшого проміжку часу. GRU мають дещо простішу структуру, але також здатні ефективно працювати з часовими залежностями. У задачах розпізнавання мовлення такі моделі можуть аналізувати послідовність акустичних ознак і визначати, які мовні одиниці найбільш імовірно відповідають вхідному сигналу.

На практиці часто використовуються двонаправлені рекурентні мережі. Вони аналізують послідовність не тільки з початку до кінця, а й у зворотному напрямку. Це дає змогу враховувати ширший контекст фрази. Такий підхід корисний у випадках, коли значення фрагмента мовлення стає зрозумілим лише після аналізу наступних слів. Проте рекурентні архітектури мають і недоліки. Вони можуть бути повільнішими під час обробки великих послідовностей, оскільки залежать від послідовного проходження даних.

Суттєвим етапом розвитку стали трансформерні архітектури. На відміну від рекурентних мереж, Transformer не обробляє вхідні дані строго по черзі. Основою його роботи є механізм уваги, який дозволяє моделі визначати, які частини послідовності мають найбільше значення для поточного результату. Це важливо для мовлення, тому що окремі звуки, слова або частини фрази можуть бути пов'язані між собою на різній відстані.

Механізм уваги дав змогу ефективніше працювати з довгими фразами та великими обсягами даних. Завдяки цьому трансформери стали основою багатьох сучасних систем обробки природної мови і розпізнавання мовлення. Вони краще масштабуються, можуть навчатися на великих наборах аудіоданих і показують високу якість у складних умовах. Саме тому сучасні моделі розпізнавання все частіше використовують архітектури, побудовані на Transformer [2].

Окремо варто розглянути проблему зіставлення аудіосигналу з текстом. У навчальних даних зазвичай відомо, яка фраза була вимовлена, але не завжди відомо, який саме фрагмент сигналу відповідає конкретній букві, фонемі або слову. Ручна розмітка таких відповідностей є складною і потребує багато часу. Для розв'язання цієї проблеми використовується функція втрат CTC, тобто Connectionist Temporal Classification.

CTC дозволяє навчати модель без точної покадрової розмітки. Модель отримує послідовність ознак і текстовий результат, а під час навчання самостійно визначає найбільш імовірне вирівнювання між ними. Це особливо зручно для розпізнавання зв'язного мовлення, де тривалість звуків може змінюватися залежно від диктора, темпу вимови та умов запису. Через це CTC часто застосовується у системах, які працюють з послідовностями різної довжини [3].

Сучасні системи розпізнавання мовлення все частіше будуються за принципом end-to-end. У таких системах модель отримує на вхід аудіосигнал або його ознаки, а на виході формує текст. Це зменшує кількість проміжних етапів і дозволяє моделі самостійно оптимізувати внутрішнє представлення даних. У порівнянні з класичними підходами такий варіант є більш гнучким, хоча він потребує значних обсягів навчальних даних і обчислювальних ресурсів.

Прикладом сучасного нейромережевого підходу є модель Whisper. Вона належить до моделей автоматичного розпізнавання мовлення і використовує трансформерну архітектуру. Її важливою перевагою є те, що вона доступна у вигляді попередньо навченої моделі. Тобто для прикладної роботи не потрібно самостійно навчати велику нейромережу з нуля. Достатньо використати готову модель, подати на вхід аудіофайл і отримати текстовий результат [4].

У програмній частині роботи Whisper використовується як готовий нейромережевий компонент для перетворення аудіосигналу в текст. Такий варіант дає змогу перевірити роботу сучасної моделі розпізнавання мовлення на власному відеофайлі та поєднати аудіальний результат з окремою обробкою відеоданих обличчя диктора.

Важливо зазначити, що нейромережеві методи можуть застосовуватися не тільки для аудіосигналу. У системах аудіовізуального розпізнавання мовлення вони також використовуються для аналізу відеоряду. Наприклад, рухи губ можна подати як послідовність кадрів або як набір ознак, отриманих після виділення області рота. У такому випадку нейромережа аналізує не звук, а зміну форми губ у часі. Це особливо корисно в умовах шуму, коли аудіосигнал має низьку якість.

Поєднання аудіального та візуального каналів є перспективним напрямом, оскільки ці джерела інформації доповнюють одне одного. Якщо звук спотворений або частково втрачений, рух губ може допомогти уточнити результат. Якщо ж відео має низьку якість або обличчя диктора частково закрите, основну роль відіграє аудіоканал. Саме тому мультимодальні системи

розпізнавання мовлення розглядаються як більш стійкі до реальних умов експлуатації [5].

Таким чином, нейромережеві методи є основою сучасних систем розпізнавання мовлення. Згорткові мережі ефективно виділяють локальні ознаки, рекурентні моделі враховують часові залежності, а трансформери дають змогу аналізувати ширший контекст мовлення. Використання попередньо навчених моделей, зокрема Whisper, робить такі технології доступними для практичної реалізації. Нейромережевий підхід використовується як основа для аудіорозпізнавання та подальшої обробки візуальних даних обличчя диктора.

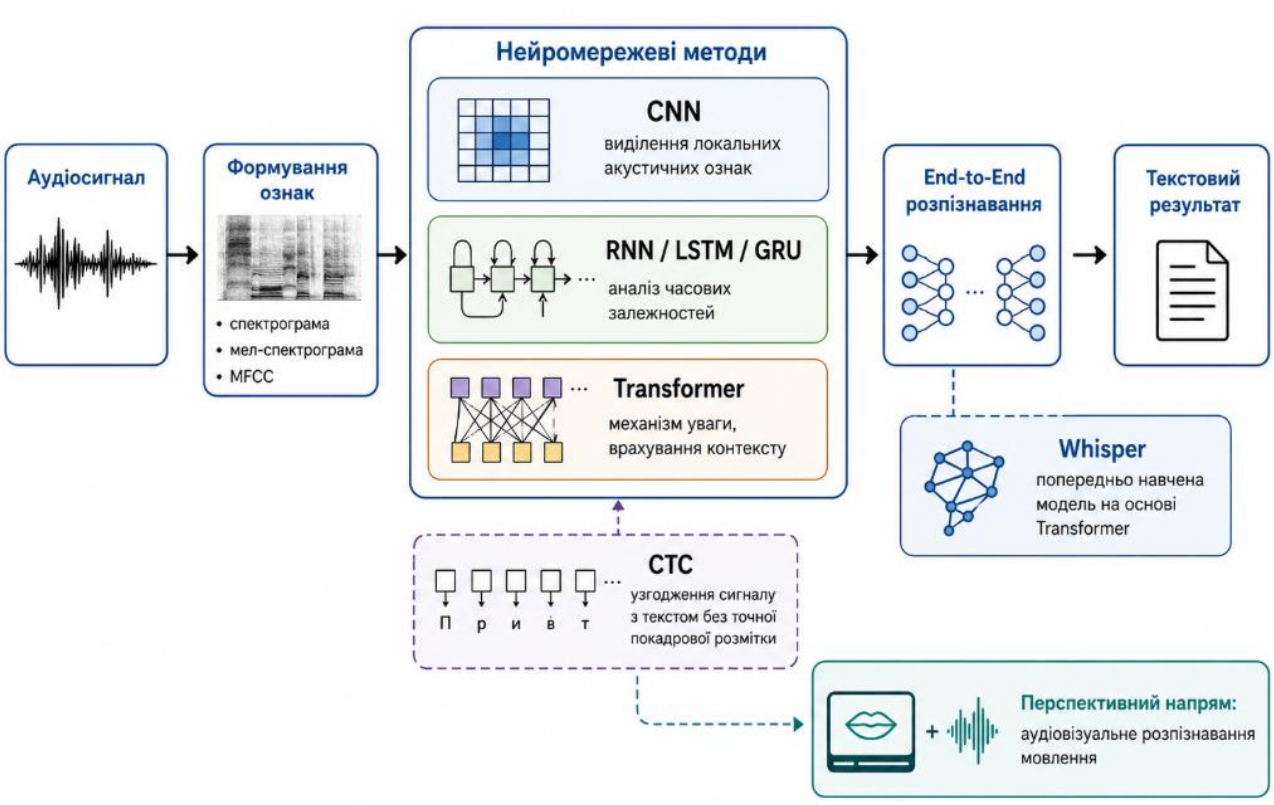


Рисунок 1.2 – Основні нейромережеві методи розпізнавання мовлення

Як видно з рисунку 1.2, нейромережеві методи розпізнавання мовлення можуть поєднувати різні підходи до аналізу аудіоданих. Згорткові мережі використовуються для виділення локальних ознак, рекурентні архітектури враховують часові залежності, а трансформерні моделі дають змогу аналізувати ширший контекст мовлення. Саме таке поєднання підходів створює основу для сучасних систем автоматичного та аудіовізуального розпізнавання мовлення

1.3 Методи візуального розпізнавання мовлення за рухом губ

Візуальне розпізнавання мовлення є окремим напрямом у галузі автоматичного розпізнавання мовлення. Його основна ідея полягає в тому, що мовна інформація може бути отримана не тільки з аудіосигналу, але й з відеоряду обличчя диктора. Під час вимови людина змінює положення губ, щелепи, язика та м’язів обличчя. Ці зміни пов’язані з артикуляцією звуків, тому можуть використовуватися для аналізу мовлення.

Такий підхід часто називають читанням з губ або Visual Speech Recognition. На відміну від класичного автоматичного розпізнавання мовлення, VSR не залежить від якості мікрофона та рівня навколишнього шуму. Це робить його корисним у ситуаціях, коли аудіосигнал має низьку якість або взагалі відсутній. Наприклад, візуальний канал може бути важливим у шумному транспорті, на виробництві, під час відеоспостереження або в системах допомоги людям з порушеннями слуху.

Водночас розпізнавання мовлення лише за рухом губ є складною задачею. Одна з головних проблем полягає у тому, що різні звуки можуть мати дуже схоже візуальне відображення. Такі групи звуків називають віземами. Наприклад, під час вимови окремих приголосних положення губ може виглядати майже однаково, хоча з погляду фонетики це різні звуки. Через це система не завжди може однозначно визначити, яке саме слово було вимовлене, якщо використовує тільки зображення губ.

На якість візуального розпізнавання також впливають умови зйомки. Якщо обличчя диктора повернуте вбік, губи частково закриті, освітлення нерівномірне або відео має низьку роздільну здатність, точність роботи моделі знижується. Проблеми можуть створювати борода, вуса, тіні на обличчі, різна швидкість мовлення та індивідуальні особливості артикуляції. Саме тому для побудови таких систем важливе значення має попередня обробка відеоданих.

Першим етапом у більшості систем VSR є виявлення обличчя та області рота. З усього відеокадру для подальшого аналізу зазвичай використовується не все зображення, а тільки область інтересу, яка містить губи. Це дозволяє зменшити кількість зайвої інформації та зосередити модель на тих ознаках, які безпосередньо пов’язані з вимовою. Після цього область рота нормалізують за

розміром, положенням і яскравістю, щоб послідовність кадрів була більш стабільною.

Для виділення області обличчя та губ можуть застосовуватися різні алгоритми комп'ютерного зору. Одним із практичних інструментів є MediaPipe Face Mesh. Ця технологія дозволяє визначати набір ключових точок на обличчі людини, зокрема точки навколо губ. За координатами цих точок можна автоматично знайти область рота та вирізати її з кожного кадру відео. Такий підхід зручний для програмної реалізації, оскільки він не потребує ручного виділення губ на кожному зображенні [6].

Після отримання послідовності кадрів області губ система переходить до етапу виділення візуальних ознак. У ранніх підходах для цього могли використовуватися геометричні параметри, наприклад ширина відкриття рота, форма губ або відстані між окремими точками. Проте такі ознаки не завжди достатньо добре описують реальну артикуляцію, бо рух губ є складним і змінюється в часі. Через це сучасні системи частіше використовують нейромережеві методи.

Для аналізу окремих кадрів або коротких відеофрагментів можуть застосовуватися згорткові нейронні мережі. Вони здатні автоматично знаходити просторові ознаки, пов'язані з формою губ, контуром рота та зміною положення артикуляційного апарату. Якщо модель аналізує не один кадр, а кілька сусідніх кадрів одночасно, вона може враховувати ще й динаміку руху. Для цього використовують тривимірні згорткові мережі, які працюють із просторово-часовими даними.

Однак лише просторових ознак недостатньо. Мовлення є послідовним процесом, тому важливо враховувати, як змінюється положення губ протягом часу. Для цього в системах візуального розпізнавання мовлення можуть використовуватися рекурентні мережі, LSTM, GRU або трансформерні архітектури. Вони аналізують послідовність візуальних ознак і допомагають моделі пов'язати окремі рухи губ із фразою або словом.

Одним із відомих прикладів нейромережевого підходу до читання з губ є модель LipNet. Вона була запропонована як end-to-end система для розпізнавання речень за відео рухів губ. Особливість такого підходу полягає в тому, що модель одночасно навчається виділяти просторово-часові візуальні ознаки та формувати текстову послідовність. У цій моделі використовуються згорткові шари,

рекурентна частина та функція втрат CTC, яка дозволяє працювати з послідовностями без точної покадрової розмітки [7].

Подальший розвиток цього напряму пов’язаний із використанням більших наборів даних та складніших моделей. Наприклад, у дослідженнях на основі відеозаписів реального мовлення розглядається задача розпізнавання не тільки окремих слів, а й цілих фраз та речень. Такі умови є складнішими, оскільки диктори мають різну вимову, відео записується в різних умовах, а мовлення не обмежується невеликим набором команд. Саме тому для розвитку VSR важливими є відкриті датасети, які містять велику кількість прикладів природного мовлення [8].

Окрему увагу в сучасних дослідженнях приділяють аудіовізуальному розпізнаванню мовлення. У таких системах аудіоканал і відеоканал використовуються разом. Це дає змогу компенсувати недоліки кожного з них. Якщо звук спотворений шумом, система може врахувати рух губ. Якщо відео має низьку якість, більшу роль відіграє аудіосигнал. У результаті мультимодальний підхід може бути стійкішим у реальних умовах експлуатації.

Прикладом сучасного мультимодального підходу є AV-HuBERT. Така модель використовує аудіальну та візуальну інформацію для формування спільного представлення мовлення. Це дозволяє застосовувати її як для задач читання з губ, так і для автоматичного розпізнавання мовлення. Важливо, що подібні підходи не розглядають аудіо та відео як повністю незалежні джерела, а намагаються знайти зв’язок між звуком і артикуляцією диктора [5].

Поєднання аудіосигналу та візуальних даних відповідає логіці аудіовізуального розпізнавання мовлення. Повністю візуальний підхід має обмеження, пов’язані з віземами, освітленням і якістю відео. Аудіальний канал також може втрачати точність у шумному середовищі. Тому в подальшій програмній реалізації аудіальний канал може використовуватися для отримання текстового результату, а відеоканал для пошуку обличчя, виділення області губ і збереження візуальних фрагментів. Такий розподіл дозволяє не навчати окрему модель читання з губ, але все одно врахувати візуальні дані диктора під час побудови системи.

Основну послідовність обробки відеоданих у системах візуального розпізнавання мовлення наведено на рисунку 1.3.

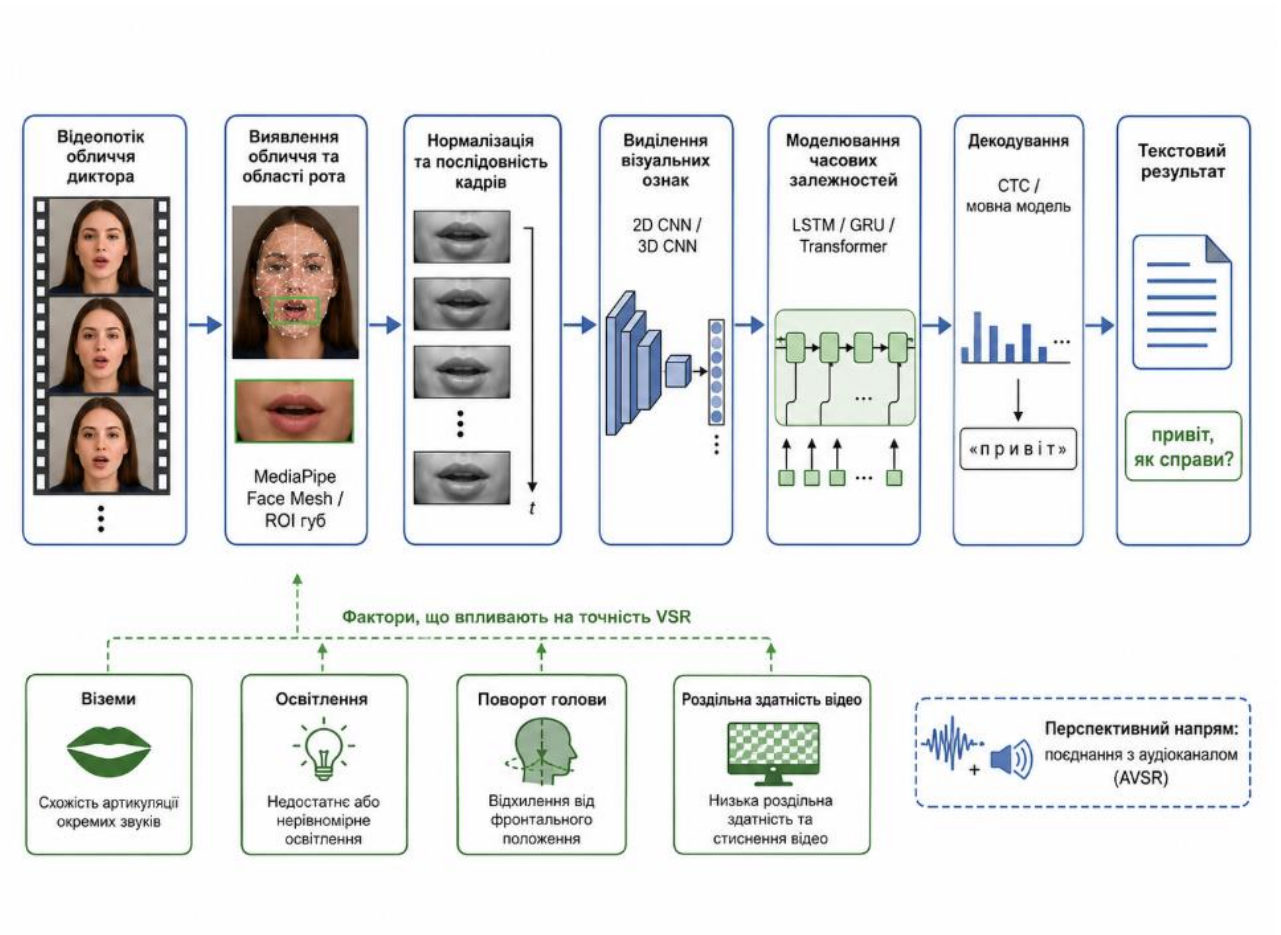


Рисунок 1.3 – Структурна схема системи візуального розпізнавання мовлення за рухом губ

Як видно з рисунку 1.3, процес візуального розпізнавання мовлення включає виявлення області обличчя та губ, формування послідовності кадрів, виділення візуальних ознак, аналіз їх зміни в часі та декодування текстового результату. Для практичної реалізації важливим є саме етап коректного визначення області губ, оскільки він дозволяє підготувати відеодані диктора для подальшої обробки.

2 МЕТОДИ РОЗПІЗНАВАННЯМ СЛІВ МОВИ

2.1 Аналіз сучасних моделей автоматичного розпізнавання мовлення

Сучасні системи автоматичного розпізнавання мовлення суттєво відрізняються від ранніх підходів. У перших системах окремо розглядалися акустична модель, мовна модель, словник та блок декодування. Така структура була досить складною, оскільки кожен етап потрібно було налаштовувати окремо. Помилка на одному з етапів могла впливати на весь подальший результат.

З розвитком глибинного навчання підхід до розпізнавання мовлення поступово змінився. Більше уваги почали приділяти моделям, які можуть обробляти мовний сигнал більш цілісно. У таких системах частина ознак формується автоматично, а якість результату залежить не тільки від ручного налаштування алгоритмів, а й від архітектури моделі, обсягу навчальних даних та умов запису.

У традиційних системах значну роль відігравала ручна підготовка ознак. Розробник мав заздалегідь визначити, які параметри мовного сигналу будуть корисними для розпізнавання. Це ускладнювало створення універсальних рішень, оскільки мовлення різних людей може суттєво відрізнятися за темпом, вимовою, інтонацією та якістю запису. Нейромережеві моделі частково зменшили цю проблему, бо здатні самостійно знаходити важливі закономірності у великій кількості даних.

Одним із відомих прикладів переходу до end-to-end розпізнавання стала система DeepSpeech. Її ідея полягала в тому, щоб спростити традиційний конвеєр і передати більшу частину обробки нейронній мережі. Модель отримує акустичні ознаки мовного сигналу та формує текстову послідовність. Для узгодження вхідного сигналу з текстом у таких системах може використовуватися функція втрат CTC, яка дозволяє працювати з послідовностями різної довжини [3], [9].

Для свого часу DeepSpeech був важливим прикладом того, що глибокі нейронні мережі можуть застосовуватися не тільки для класифікації окремих звуків, а й для повноцінного розпізнавання мовлення. Модель показала, що частину етапів, які раніше виконувалися окремими алгоритмами, можна об’єднати в одну систему.

Таке рішення зробило структуру системи більш компактною. Водночас воно не усунуло всі складності, пов’язані з якістю даних. Якщо модель навчалася переважно на чистому звуку, то в реальних умовах із шумом або спотворенням сигналу результат може бути гіршим. Це важливо враховувати при виборі моделі для практичного використання.

Подальшим розвитком цього напряму стала система DeepSpeech 2. У ній використовувалися більші обсяги навчальних даних і складніша нейромережева архітектура. Важливим є те, що DeepSpeech 2 розглядалася не лише як рішення для англійської мови. Автори показали можливість застосування end-to-end підходу для різних мовних середовищ, зокрема для англійської та китайської мов [10].

Цей момент важливий для подальшого вибору моделі розпізнавання мовлення. Розпізнавання мовлення не можна розглядати окремо від мови, акценту та умов запису. Одна й та сама модель може давати різні результати залежно від того, наскільки її навчальні дані схожі на реальні вхідні записи. Якщо в навчальній вибірці мало шумних записів або мало прикладів різної вимови, якість роботи системи може знижуватися.

Моделі DeepSpeech добре підходять для пояснення загальної логіки нейромережевого розпізнавання. Їх можна описати через послідовність обробки аудіосигналу, формування ознак, нейромережеве моделювання та отримання тексту. Для навчальної або дослідницької роботи така структура є зрозумілою. Проте для обраної програмної реалізації самостійне навчання або глибоке налаштування подібної системи є надмірно ресурсомістким, оскільки потребує великого набору розмічених аудіоданих і значного часу на підготовку моделі.

Навчання або якісне донавчання таких моделей потребує значних обсягів розмічених аудіоданих, обчислювальних ресурсів і часу на підготовку датасету. Через це для обраної програмної реалізації використовується готова модель розпізнавання мовлення, яка може бути інтегрована у Python-програму та перевірена на власному відеофайлі.

Іншим важливим напрямом розвитку стали самонавчальні моделі. Однією з найвідоміших серед них є wav2vec 2.0. Її головна особливість полягає в тому, що модель може попередньо навчатися на великій кількості нерозміченого аудіо. На першому етапі не обов’язково мати точну текстову розмітку для кожного запису, що є суттєвою перевагою для багатьох задач.

Модель wav2vec 2.0 працює з мовним сигналом і формує внутрішні представлення, які надалі можуть використовуватися для розпізнавання. Під час попереднього навчання частина інформації приховується, а модель намагається правильно визначити приховані фрагменти. Завдяки цьому вона навчається знаходити корисні закономірності у мовленні без повної ручної розмітки [11].

Такий підхід особливо цінний для мов, для яких складно зібрати великі якісні набори даних. У традиційних системах нестача розмічених записів є серйозною проблемою. Самонавчання частково її зменшує, оскільки дозволяє використати більший обсяг звичайних аудіозаписів. Це робить wav2vec 2.0 важливим етапом розвитку сучасних систем автоматичного розпізнавання мовлення.

При цьому wav2vec 2.0 не можна розглядати як повністю готове універсальне рішення для будь-якої практичної задачі. У багатьох випадках модель усе одно потрібно донавчати або налаштовувати під конкретні умови. Це може бути інша мова, технічна лексика, неякісний мікрофон, фоновий шум або особливості вимови диктора. Тому при виборі моделі важливо враховувати не тільки її наукову цінність, а й зручність використання у конкретній програмній реалізації.

Модель wav2vec 2.0 розглядається як важливий приклад сучасного самонавчального підходу до розпізнавання мовлення. Вона показує, як сучасні

моделі можуть використовувати великі обсяги нерозміченого аудіо і формувати мовні представлення без повної ручної розмітки. Проте для програмної реалізації, пов’язаної з обробкою власного відеофайлу, зручніше використати модель, яка вже має готову реалізацію, простий запуск у Python та не потребує окремого етапу донавчання.

Окреме місце серед сучасних моделей займає Whisper. Це попередньо навчена модель автоматичного розпізнавання мовлення, побудована на трансформерній архітектурі. На відміну від моделей, які потребують окремого навчання або донавчання під конкретний набір даних, Whisper може використовуватися як готовий компонент для обробки аудіофайлів. Модель приймає на вхід звуковий файл і формує текстовий результат, що робить її зручною для прикладних задач розпізнавання мовлення [4].

Для розроблюваного програмного модуля важливим є не лише рівень сучасності моделі, а й можливість її практичного використання. DeepSpeech, DeepSpeech 2 і wav2vec 2.0 добре показують розвиток нейромережевих підходів, однак їх застосування часто пов’язане з додатковим налаштуванням, підготовкою даних або складнішою інтеграцією. Whisper у цьому випадку є більш придатним варіантом, оскільки дозволяє зосередитися на побудові повного циклу обробки відеофайлу, а не на навчанні моделі розпізнавання з нуля.

У програмній реалізації Whisper використовується для обробки аудіодоріжки, отриманої з відеофайлу. Після витягування звуку модель виконує розпізнавання мовлення, а отриманий текст зберігається як окремий результат роботи аудіального каналу. Такий підхід дає змогу перевірити роботу сучасної ASR-моделі на власному відеозаписі та пов’язати її результат з подальшою обробкою відеокадрів.

Водночас Whisper не виконує аналіз руху губ і не використовує зображення обличчя диктора. Тому відеоканал у програмному модулі обробляється окремо. Для нього використовуються засоби комп’ютерного зору, які дозволяють визначити обличчя, знайти ключові точки та виділити область

губ. У результаті аудіальна частина відповідає за перетворення мовлення в текст, а візуальна частина готує фрагменти області губ для подальшого аналізу.

Якщо порівняти розглянуті моделі, можна простежити поступовий розвиток підходів до автоматичного розпізнавання мовлення. DeepSpeech показує можливості end-to-end розпізнавання, DeepSpeech 2 розширює цей підхід за рахунок більших обсягів даних і складнішої архітектури, wav2vec 2.0 демонструє переваги самонавчання на нерозміченому аудіо, а Whisper є прикладом готової трансформерної моделі, яку можна безпосередньо використати у прикладній програмі.

Еволюцію сучасних моделей автоматичного розпізнавання мовлення наведено на рисунку 2.1.



Рисунок 2.1 – Еволюція сучасних моделей автоматичного розпізнавання мовлення

Як видно з рисунку 2.1, розвиток сучасних моделей автоматичного розпізнавання мовлення відбувався від перших end-to-end систем до самонавчальних і трансформерних підходів. DeepSpeech і DeepSpeech 2

показують перехід до більш цілісної обробки мовного сигналу, wav2vec 2.0 демонструє можливості попереднього навчання на нерозмічених аудіоданих, а Whisper є прикладом готової моделі, яку можна безпосередньо використати для розпізнавання аудіодоріжки відеофайлу.

2.2 Аналіз методів обробки візуальних даних обличчя диктора

Обробка візуальних даних обличчя диктора є важливою частиною систем аудіовізуального розпізнавання мовлення. Якщо аудіоканал дозволяє аналізувати звукову складову мовлення, то відеоканал дає змогу врахувати артикуляційні рухи людини. Насамперед це стосується губ, нижньої щелепи та частини мімічних м’язів, які змінюють положення під час вимови окремих звуків і слів.

На відміну від звичайної задачі розпізнавання обличчя, у даному випадку важливим є не встановлення особи диктора, а аналіз руху окремих ділянок обличчя. Система має знайти обличчя у кадрі, визначити область рота, відстежити її зміну в часі та підготувати ці дані для подальшої обробки. Якщо цей етап виконано неточно, наступна нейромережева модель отримає спотворені або зайві дані.

Першим кроком зазвичай є виявлення обличчя на відеокадрі. Це потрібно для того, щоб відокремити обличчя диктора від фону, інших людей або об’єктів у кадрі. У реальних умовах така задача не завжди є простою. Освітлення може бути нерівномірним, камера може мати невисоку якість, а обличчя людини іноді розташоване не по центру кадру.

Після виявлення обличчя система переходить до визначення ключових точок. Такі точки називають facial landmarks. Вони описують положення очей, носа, брів, контуру обличчя та рота. Для задачі розпізнавання мовлення найбільше значення мають точки навколо губ, оскільки саме вони дозволяють виділити область інтересу для подальшого аналізу.

Одним із відомих інструментів для аналізу обличчя є OpenFace. Він використовується для визначення лицевих орієнтирів, оцінки положення голови та аналізу мімічної активності. Його можна розглядати як приклад підходу, де обробка обличчя виконується не тільки на рівні виявлення самого об’єкта, а й на рівні більш детального аналізу його геометрії [12].

Для задачі розпізнавання мовлення важливо також враховувати положення голови. Якщо диктор дивиться прямо в камеру, область губ виділяється значно простіше. Коли голова повернута вбік, частина рота може бути менш помітною, а форма губ у кадрі спотворюється. Через це система може неправильно оцінити артикуляцію або втратити частину важливої інформації.

У практичних програмних системах зручним інструментом для роботи з обличчям є MediaPipe Face Mesh. Цей підхід дозволяє визначати велику кількість ключових точок на обличчі та працювати з відеопотоком у реальному часі. Це корисно для програмної реалізації, оскільки за координатами точок навколо рота можна автоматично знайти область губ і вирізати її з кожного кадру відео [6].

Використання MediaPipe спрощує практичну реалізацію обробки відеокадрів. Програма отримує координати потрібних точок, визначає межі прямокутної області навколо губ та передає цей фрагмент на подальше збереження або аналіз. При цьому не потрібно вручну позначати область рота на кожному кадрі відео.

Після виділення області губ формується так звана область інтересу, або ROI. У ній залишається тільки та частина кадру, яка безпосередньо пов’язана з артикуляцією. Це дозволяє зменшити обсяг зайвих даних. Наприклад, фон, волосся, одяг або інші об’єкти в кадрі не мають прямого значення для аналізу вимови, тому їх бажано не передавати до моделі.

Область губ зазвичай приводиться до єдиного розміру. Це потрібно для того, щоб усі кадри мали однаковий формат і могли бути оброблені програмою або нейромережею. Також може виконуватися нормалізація яскравості, зміна контрасту та згладжування зображення. Ці дії не розпізнають мовлення самі по собі, але вони підвищують якість підготовки даних.

Важливим є не тільки сам кадр, а й послідовність кадрів. Рух губ має часовий характер, тому окреме статичне зображення не завжди дає достатньо інформації. Наприклад, однакова форма губ може зустрічатися у різних звуках, але попередній і наступний рухи допомагають уточнити значення. Тому відеодані потрібно розглядати як послідовність, а не як набір незалежних фотографій.

У системах візуального розпізнавання мовлення можуть застосовуватися різні варіанти представлення візуальних даних. Один підхід полягає у використанні самих кадрів області губ. Інший підхід передбачає використання координат ключових точок. У першому випадку модель аналізує зображення, у другому працює з геометричними параметрами руху губ. Обидва підходи мають практичне значення.

Аналіз кадрів області губ дозволяє враховувати форму рота, контури губ, відкриття щелепи та інші візуальні ознаки. Такий варіант добре підходить для згорткових нейронних мереж, які можуть знаходити локальні закономірності на зображенні. Проте він потребує якісного відео, оскільки розмиття або низька роздільна здатність можуть суттєво погіршити результат.

Використання координат ключових точок має іншу перевагу. Воно дозволяє зменшити обсяг даних, оскільки замість повного зображення система працює з набором числових координат. За такими координатами можна оцінити ширину відкриття рота, зміну форми губ, відстані між окремими точками та динаміку руху. Для простішої програмної моделі це може бути зручним рішенням.

Однак самі координати не завжди передають усю інформацію. Вони не враховують текстуру губ, тіні, дрібні зміни форми рота та інші особливості зображення. Тому в більш складних системах часто використовується поєднання різних ознак. Наприклад, можна одночасно аналізувати область губ як зображення та координати ключових точок як додаткову геометричну інформацію.

Для навчання і тестування моделей візуального розпізнавання важливе значення мають відкриті датасети. Одним із таких наборів є Lip Reading in the Wild. Він містить короткі відеофрагменти з вимовою окремих слів різними дикторами. Особливість такого датасету полягає в тому, що відео отримані з реальних умов, а не тільки з лабораторного запису. Це робить задачу складнішою, але ближчою до практичного використання [13].

Іншим важливим напрямом є розпізнавання не окремих слів, а цілих фраз і речень. У цьому випадку система повинна враховувати не тільки артикуляцію окремих звуків, а й контекст висловлювання. Дослідження Lip Reading Sentences in the Wild показало, що задача читання з губ може розглядатися як розпізнавання природних речень у реальних умовах, а не лише як класифікація обмеженого набору команд [8].

Розпізнавання мовлення за візуальними даними не можна звести тільки до пошуку губ у кадрі. На результат впливають якість відео, положення голови, освітлення, швидкість мовлення та індивідуальні особливості артикуляції. Через це обробка відеоданих має бути окремим етапом, від якого залежить якість подальшого аналізу.

Однією з основних складностей залишається візуальна схожість окремих звуків. Рухи губ не завжди дозволяють точно відрізнити одну фонему від іншої, тому для таких задач важливо аналізувати не один кадр, а послідовність рухів. Саме тому в системах VSR застосовуються згорткові мережі, рекурентні архітектури, трансформери та комбіновані підходи, які враховують зміну візуальних ознак у часі [7].

У розроблюваній системі відеоканал використовується для підготовки візуальних даних, а не для повного розпізнавання мовлення за рухом губ. Його завдання полягає у виявленні обличчя диктора, визначенні ключових точок, виділенні області губ і збереженні відповідних фрагментів. Для цього використовується MediaPipe Face Mesh, оскільки цей інструмент дає змогу автоматично отримувати координати точок обличчя без ручної розмітки кожного кадру.

Таким чином, обробка візуальних даних обличчя диктора є підготовчим, але важливим етапом аудіовізуальної системи. Вона дозволяє отримати послідовність фрагментів області губ, які можуть бути використані для подальшого аналізу разом із результатами аудіального розпізнавання.

2.3 Порівняння аудіальних, візуальних та аудіовізуальних підходів

Після розгляду сучасних моделей автоматичного розпізнавання мовлення та методів обробки візуальних даних можна перейти до порівняння основних підходів до побудови таких систем. Для порівняння обрано три варіанти: розпізнавання лише за аудіосигналом, розпізнавання лише за візуальними даними обличчя диктора та комбінований аудіовізуальний підхід. Кожен із них має власну сферу застосування, переваги та обмеження.

Аудіальний підхід є найбільш поширеним у сучасних системах розпізнавання мовлення. Він використовує звуковий сигнал, отриманий з мікрофона або аудіодоріжки відеофайлу. Саме на цьому принципі працює більшість голосових асистентів, систем автоматичного створення субтитрів, сервісів диктування тексту та інструментів транскрибування. Його головна перевага полягає в тому, що мовлення безпосередньо представлено у звуковій формі, тому система отримує повнішу інформацію про фонеми, інтонацію, паузи та темп вимови.

Сучасні аудіальні моделі, зокрема Whisper, демонструють високу якість розпізнавання у багатьох практичних задачах. Їх можна використовувати без самостійного навчання з нуля, що значно спрощує розробку прикладних програмних систем. [4].

Разом з тим аудіальний підхід має помітну залежність від якості звуку. Якщо запис містить сильний фоновий шум, відлуння, перекриття голосів або спотворення мікрофона, точність розпізнавання може знижуватися. У реальних умовах це трапляється досить часто. Наприклад, мовлення в транспорті, на

вулиці, у виробничому приміщенні або під час відеозв’язку не завжди має достатньо чисту аудіодоріжку.

Візуальний підхід використовує інше джерело інформації. У цьому випадку система аналізує не звук, а рухи губ, щелепи та окремих ділянок обличчя диктора. Такий варіант особливо цікавий у ситуаціях, коли аудіосигнал недоступний або має низьку якість. Наприклад, якщо звук пошкоджений, відсутній або перекритий шумом, відеодані можуть залишатися додатковим джерелом мовної інформації.

Методи візуального розпізнавання мовлення мають практичне значення і для систем допомоги людям з порушеннями слуху. Також вони можуть використовуватися у відеоаналітиці, людино-машинній взаємодії та мультимодальних інтерфейсах. Прикладом нейромережевого підходу до читання з губ є LipNet, де відеодані рухів губ використовуються для формування текстової послідовності [7].

Однак повністю візуальне розпізнавання мовлення є значно складнішим, ніж може здаватися на перший погляд. Основна проблема полягає в тому, що багато звуків мають схоже візуальне відображення. Рухи губ не завжди дозволяють однозначно відрізнити одну фонему від іншої. Через це система може правильно визначити загальний характер артикуляції, але помилитися у конкретному слові.

На якість візуального підходу також впливають умови зйомки. Якщо обличчя диктора погано освітлене, повернуте вбік або частково закрите, система отримує неповні дані. Низька роздільна здатність відео, розмиття кадру та швидка вимова також ускладнюють аналіз. Саме тому датасети на кшталт LRS і LRW мають важливе значення для дослідження таких моделей, оскільки вони містять приклади мовлення у більш реалістичних умовах [8, 13].

Комбінований аудіовізуальний підхід намагається використати переваги обох джерел інформації. Його ідея полягає в тому, що аудіосигнал і відеоряд обличчя диктора доповнюють один одного. Якщо звук якісний, основну роль може відігравати аудіоканал. Якщо аудіо погіршується через шум або

спотворення, візуальний канал може допомогти уточнити результат. Такий підхід є ближчим до того, як людина сама сприймає мовлення у складних умовах.

Аудіовізуальні системи не обов’язково мають повністю замінювати класичне аудіальне розпізнавання. Частіше вони розглядаються як спосіб підвищення стійкості системи. У реальних умовах складно гарантувати, що аудіо або відео завжди будуть ідеальними. Тому поєднання двох каналів дозволяє зробити систему гнучкішою.

Сучасні дослідження мультимодального розпізнавання мовлення показують, що аудіо та відео можуть використовуватися не просто як два окремі потоки даних. Вони можуть формувати спільне представлення мовлення. Саме така ідея лежить в основі підходів типу AV-HuBERT, де модель навчається враховувати зв’язок між аудіальними та візуальними ознаками [5].

У подальшій реалізації використовується комбінована схема обробки аудіальних і візуальних даних. Аудіоканал передається до готової нейромережевої моделі Whisper для формування текстового результату, а відеоканал обробляється окремо для визначення обличчя диктора та виділення області губ. Такий поділ дозволяє узгодити теоретичний аналіз аудіовізуального розпізнавання мовлення з реальною програмною реалізацією на Python.

Порівняльну характеристику трьох підходів наведено в таблиці 2.1.

Таблиця 2.1 – Порівняння підходів до розпізнавання мовлення

Підхід	Основні дані	Переваги	Обмеження	Використання у програмній системі
Аудіальний	Аудіосигнал мовлення	Висока інформативність звуку, доступність готових моделей, зручність реалізації	Залежність від шуму, якості мікрофона та умов запису	Використовується як основний канал розпізнавання за допомогою Whisper
Візуальний	Відеоряд обличчя та рухів губ	Може працювати без звуку, корисний при шумі, дозволяє аналізувати артикуляцію	Проблема візем, залежність від освітлення, ракурсу та якості відео	Використовується для виділення області губ і аналізу візуальних ознак
Аудіовізуальний	Аудіосигнал і відеодані обличчя	Більш стійкий до реальних умов, поєднує два джерела інформації	Складніша реалізація, потреба в синхронізації аудіо та відео	Є основним напрямом побудови програмної системи в роботі

Як видно з таблиці 2.1, аудіальний, візуальний та аудіовізуальний підходи розв’язують одну задачу, але спираються на різні джерела інформації. Аудіальний канал є основним для отримання текстового результату, оскільки саме звук містить повну мовну інформацію. Водночас його якість може знижуватися за наявності шуму, тихого мовлення або технічних спотворень запису.

Візуальний канал має інше призначення. Він не дає такого повного мовного сигналу, як аудіо, але дозволяє враховувати рухи губ та інші

артикуляційні ознаки. Його результат залежить від освітлення, положення обличчя, видимості області рота та якості відео. Через це повністю візуальне розпізнавання є складним, але відеодані можуть бути корисними як додаткове джерело інформації.

Для програмної реалізації обрано комбіновану схему, у якій аудіальний і візуальний канали виконують різні функції. Аудіодоріжка передається до Whisper для формування текстового результату. Відеоряд обробляється окремо: програма визначає обличчя диктора, виділяє область губ і зберігає відповідні фрагменти. Такий варіант не потребує навчання окремої моделі читання з губ, але дозволяє працювати з двома складовими одного відеофайлу.

Під час поєднання аудіо та відео важливим залишається питання узгодженості цих даних. Якщо звук і зображення не відповідають одне одному в часі, аналіз може бути менш точним. У розроблюваному модулі ця проблема зменшується завдяки тому, що аудіодоріжка і відеокадри беруться з одного вхідного файлу.

Загальну логіку використання трьох підходів до розпізнавання мовлення наведено на рисунку 2.2.



Рисунок 2.2 – Порівняння аудіального, візуального та аудіовізуального підходів до розпізнавання мовлення

Як видно з рисунку 2.2, аудіальний підхід використовується для перетворення мовлення в текст, візуальний підхід пов’язаний з аналізом області губ, а аудіовізуальний підхід поєднує ці два джерела даних. Саме така логіка використовується в програмному модулі: Whisper обробляє аудіодоріжку, а OpenCV і MediaPipe Face Mesh застосовуються для обробки відеокадрів та виділення області губ диктора.

3 РОЗРОБКА ПРОГРАМНОГО МОДУЛЯ АУДІОВІЗУАЛЬНОЇ ОБРОБКИ МОВЛЕННЯ

3.1 Розробка архітектури програмної системи розпізнавання мовлення

Вхідними даними програмного модуля є відеофайл із записом мовлення диктора. У ньому одночасно міститься звукова доріжка та послідовність кадрів із зображенням обличчя. Тому обробка в модулі поділяється на два напрями: аудіальний канал використовується для отримання текстового результату, а відеоканал — для роботи з кадрами обличчя та виділення області губ.

Робота модуля починається із завантаження вхідного відеофайлу. Програма отримує шлях до файлу, відкриває його та готує дані до подальшої обробки. Після цього з відео виділяється аудіодоріжка, яка передається до модуля розпізнавання мовлення. Окремо виконується покадрове читання відео, необхідне для аналізу обличчя диктора.

Аудіальний канал призначений для перетворення мовлення в текст. На цьому етапі звукова доріжка зберігається в аудіофайл і передається до Whisper. Модель обробляє отриманий аудіосигнал, після чого програма зберігає результат розпізнавання у текстовому вигляді [4]. У межах цієї архітектури Whisper відповідає саме за аудіальну частину, тоді як відеокадри обробляються окремо.

Відеоканал працює з кадрами вхідного файлу. Для цього відео відкривається засобами OpenCV, після чого кадри послідовно зчитуються в циклі [14]. Кожен кадр готується до подальшої обробки та передається до MediaPipe Face Mesh, який визначає ключові точки обличчя, зокрема точки навколо рота [6].

На основі координат цих точок програма визначає межі області губ у кадрі. Далі формується прямокутна область інтересу, яка охоплює рот диктора з невеликим відступом від країв. Такий фрагмент може бути збережений окремо або використаний для перевірки коректності роботи відеоканалу.

Після формування області інтересу з кадру вирізається фрагмент із зображенням губ диктора. Такий фрагмент може бути збережений як окреме зображення або використаний для візуальної перевірки роботи модуля. У процесі обробки також може формуватися кадр із позначеною областю рота, що дозволяє побачити, чи правильно програма визначає потрібну частину обличчя.

Виділення області губ потрібне для того, щоб відокремити важливу частину кадру від другорядних елементів. Повне зображення може містити фон, одяг, волосся, плечі або інші об’єкти, які не мають прямого значення для аналізу артикуляції. Область губ пов’язана з рухами мовного апарату, тому саме вона використовується як основний візуальний фрагмент у відеоканалі.

У цій архітектурі відеоканал не виконує повне розпізнавання мовлення за рухом губ. Його завдання полягає у підготовці візуальних даних: пошуку обличчя, визначенні ключових точок і виділенні області губ. Текстовий результат формується окремо, за аудіодоріжкою відеофайлу. Такий розподіл дозволяє поєднати розпізнавання мовлення за аудіосигналом із підготовкою візуальних даних обличчя диктора.

Структурно програмний модуль складається з кількох взаємопов’язаних частин: завантаження відеофайлу, виділення аудіодоріжки, розпізнавання мовлення, покадрової обробки відео, визначення ключових точок обличчя та формування області губ. Вихідними результатами роботи модуля є розпізнаний текст, аудіофайл, кадр із позначеною областю рота та окремі фрагменти із зображенням губ.

Узагальнену архітектуру програмного модуля наведено на рисунку 3.1.

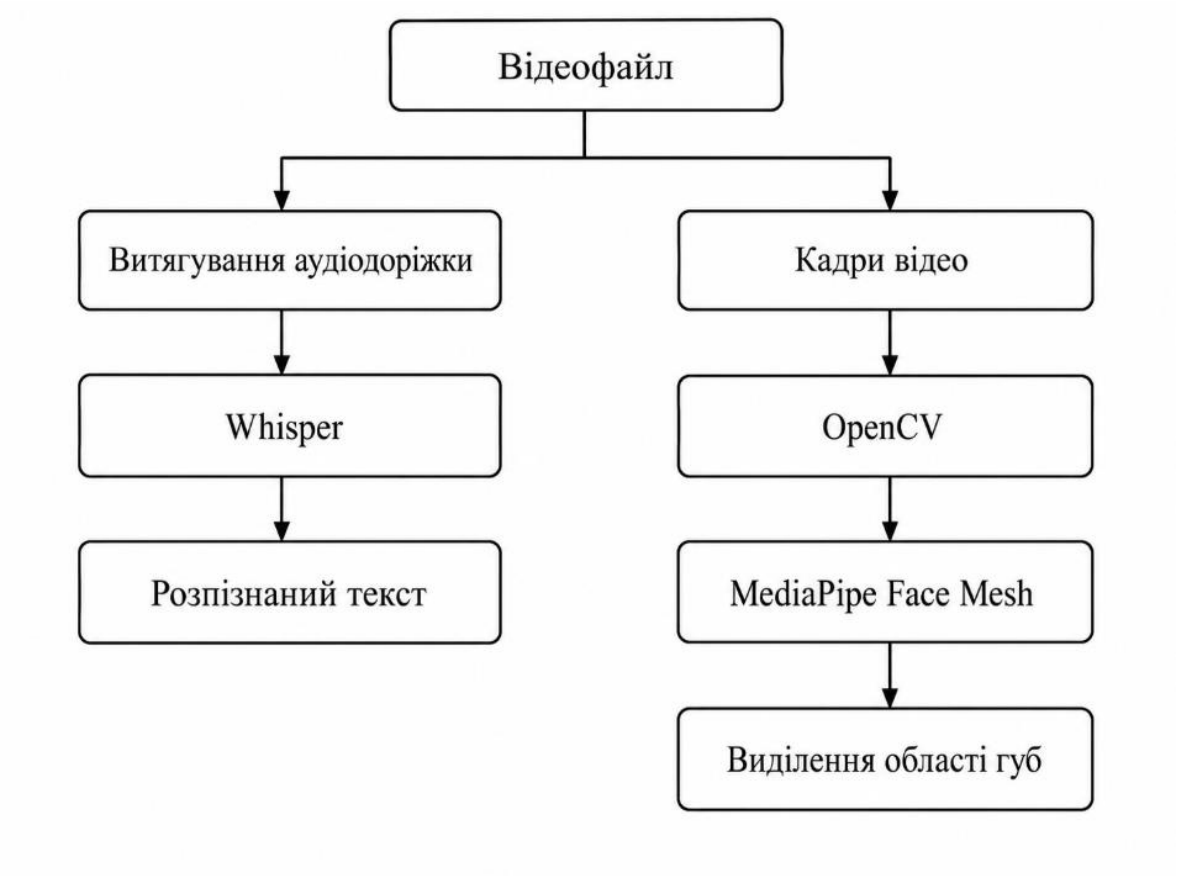


Рисунок 3.1 – Архітектура програмної системи аудіовізуальної обробки мовлення

Як видно з рисунку 3.1, обробка починається з одного вхідного відеофайлу, після чого дані розділяються на аудіальний і візуальний канали. Аудіальний канал завершується формуванням текстового результату, а візуальний канал — виділенням області губ на кадрах відео. Така схема показує, як у межах одного модуля поєднуються два різні напрями обробки мовлення.

На рівні програмних засобів архітектура поєднує MoviePy, Whisper, OpenCV та MediaPipe Face Mesh. MoviePy використовується для отримання аудіодоріжки з відеофайлу, Whisper — для розпізнавання мовлення, OpenCV — для читання кадрів, а MediaPipe Face Mesh — для визначення ключових точок обличчя [15]. Кожен із цих компонентів виконує окрему функцію, але всі вони працюють з даними, отриманими з одного вхідного відеофайлу.

Збереження проміжних результатів потрібне для подальшої перевірки роботи модуля. Розпізнаний текст дозволяє оцінити аудіальний канал, а кадр із позначеною областю рота та фрагменти губ дають змогу перевірити коректність відеообробки. Детальний аналіз цих результатів виконується під час тестування програмного модуля.

Архітектура програмного модуля побудована навколо обробки одного вхідного відеофайлу. З нього отримується аудіосигнал для розпізнавання мовлення та послідовність відеокадрів для аналізу обличчя диктора. Подальший підрозділ присвячено реалізації цих етапів у вигляді Python-програми.

3.2 Реалізація програмного модуля на Python

Програмний модуль реалізовано як Python-скрипт, який послідовно обробляє один вхідний відеофайл із записом мовлення диктора. У коді задається шлях до відеофайлу, назви проміжних результатів і папка для збереження фрагментів області губ. Після запуску програма виконує підготовлену послідовність операцій: створює аудіофайл, запускає розпізнавання мовлення, читає кадри відео та зберігає результати відеообробки.

На першому етапі з відеофайлу витягується звукова доріжка, яка зберігається у файл extracted_audio.wav. Далі цей файл передається до Whisper, а отриманий результат розпізнавання записується у recognized_text.txt. У цій частині програми Whisper виконує конкретну функцію обробки підготовленого аудіофайлу та формування текстового результату.

Відеочастина обробляється окремо. OpenCV відкриває вхідний файл, послідовно зчитує кадри та передає їх до MediaPipe Face Mesh для визначення ключових точок обличчя. Після цього програма відбирає точки, що відповідають області рота, обчислює межі прямокутної області, вирізає фрагмент губ і зберігає його у папку lip_frames. Для контролю також формується кадр frame_with_lips.jpg, на якому область губ позначена прямокутником.

Межі прямокутної області губ формуються на основі координат ключових точок. Щоб фрагмент не обрізався занадто близько до контуру рота, до розрахованих меж додається невеликий відступ. Межі області губ можна подати так:

$$x_{min} = \min(x_i) - d, \quad x_{max} = \max(x_i) + d$$
$$y_{min} = \min(y_i) - d, \quad y_{max} = \max(y_i) + d \tag{3.1}$$

де x_i та y_i — координати ключових точок губ; d — додатковий відступ від меж області; x_min, x_max, y_min, y_max — межі прямокутної області, яка містить губи диктора. У програмній реалізації такий спосіб визначення меж дозволяє зберігати не лише сам контур губ, а й невелику прилеглу частину обличчя. Це зменшує ризик обрізання важливих візуальних ознак під час покадрової обробки відео.

Після обчислення меж області інтересу програма вирізає відповідний фрагмент із кадру та зберігає його в папку з результатами. Окремо може зберігатися кадр із позначеною областю рота, що дає змогу перевірити правильність визначення губ без перегляду всього відеофайлу. Таким чином, у процесі виконання модуля формуються як текстові, так і візуальні проміжні результати.

Послідовність основних етапів роботи програмного модуля наведено на рисунку 3.2. Схема відображає рух даних від моменту завантаження вхідного відеофайлу до формування результатів аудіального та візуального каналів.



Рисунок 3.2 – Послідовність роботи програмного модуля під час обробки відеофайлу

Як видно з рисунку 3.2, робота програми має послідовний характер. Спочатку виконується підготовка вхідного відеофайлу, після чого окремо формується результат аудіального каналу та результат відеообробки. Аудіальна частина завершується збереженням розпізнаного тексту, а візуальна частина — збереженням фрагментів області губ і кадру з позначеною областю рота.

Реалізований модуль формує набір результатів, які можна використати для перевірки його роботи. Розпізнаний текст дає змогу оцінити аудіальну частину, а збережені зображення області губ дозволяють перевірити коректність відеообробки. Детальний аналіз цих результатів на тестових відеофайлах наведено в наступному розділі.

4 ТЕСТУВАННЯ ТА АНАЛІЗ РЕЗУЛЬТАТІВ РОБОТИ ПРОГРАМНОГО МОДУЛЯ

4.1 Організація тестування програмного модуля

Після розроблення програмного модуля було проведено його тестування на коротких відеофайлах із записом мовлення диктора. На цьому етапі потрібно було перевірити, чи правильно програма виконує основні дії: відкриває вхідне відео, виділяє аудіодоріжку, формує текстовий результат розпізнавання мовлення та зберігає фрагменти області губ.

Тестування виконувалося з урахуванням того, що робота аудіального та візуального каналів може змінюватися залежно від умов запису. Для цього використовувалися відеофайли з різним освітленням, положенням обличчя диктора та рівнем звуку. Така перевірка дозволяє оцінити не лише сам факт запуску програми, а й стабільність її роботи під час обробки різних вхідних даних.

Для першої перевірки було використано відеофайл, записаний за найсприятливіших умов: обличчя диктора знаходилося у фронтальному положенні, освітлення було достатнім, а мовлення вимовлялося чітко. Такий тест потрібний для фіксації базового результату роботи програмного модуля без додаткового ускладнення умов запису. Запуск програми виконувався з робочої папки проєкту через середовище PowerShell. Результат виконання програмного модуля для першого тестового відеофайлу наведено на рисунку 4.1.

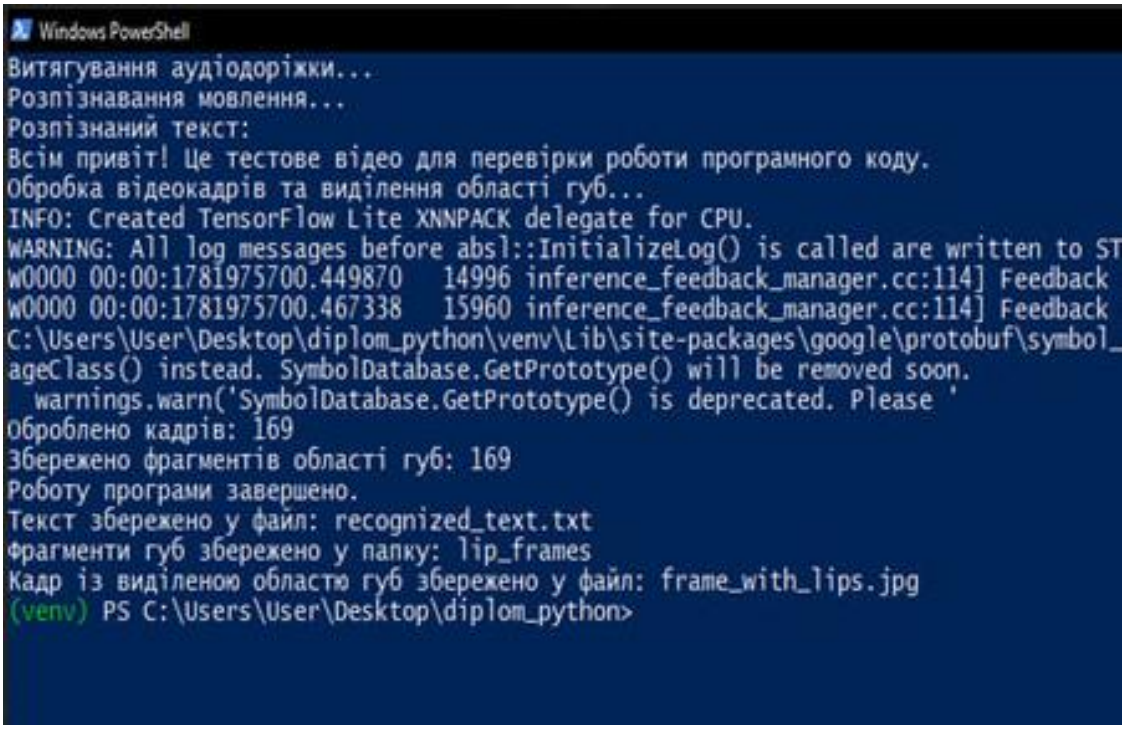


Рисунок 4.1 – Результат виконання програмного модуля у середовищі PowerShell

Як видно з рисунку 4.1, за нормального освітлення, фронтального положення обличчя та чіткого мовлення програма виконала всі основні етапи обробки відеофайлу. Було витягнуто аудіодоріжку, виконано розпізнавання мовлення, оброблено відеокадри та збережено фрагменти області губ диктора. Текстовий результат у цьому тесті був сформований коректно, без помітних змістових спотворень, що свідчить про стабільну роботу модуля за сприятливих умов запису.

Першим етапом роботи модуля є обробка аудіальної частини відеофайлу. Для цього з вхідного відео витягується звукова доріжка, яка зберігається у форматі WAV. Після цього аудіофайл передається до моделі Whisper, що виконує автоматичне розпізнавання мовлення. У першому тесті розпізнаний

текст відповідав змісту вимовленої фрази, тому аудіальний канал можна вважати коректно відпрацьованим для умов нормального запису.

Після обробки аудіодоріжки програма переходить до відеокадрів. Відеофайл відкривається засобами OpenCV, після чого кадри послідовно передаються на обробку до MediaPipe Face Mesh. Цей інструмент визначає ключові точки обличчя диктора, серед яких окремо використовуються координати, що відповідають області губ. На основі цих координат формується прямокутна область, яка вирізається з кадру та зберігається як окремий фрагмент.

Один із результатів роботи візуального каналу наведено на рисунку 4.2.

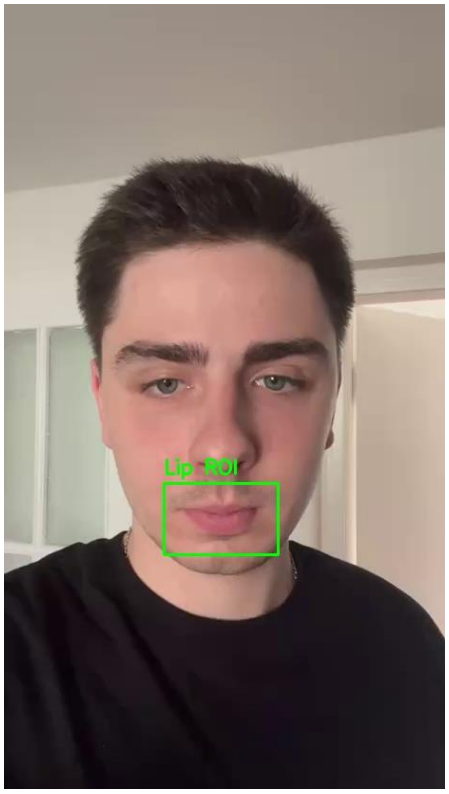


Рисунок 4.2 – Кадр відео з виділеною областю губ диктора

На рисунку 4.2 показано кадр першого тестового відео, у якому програма визначила область губ диктора. Виділена прямокутна зона охоплює губи та невелику прилеглу частину обличчя. Такий відступ потрібний для того, щоб під

час збереження фрагмента не втрачалися важливі візуальні ознаки, пов’язані з рухом губ.

4.2 Аналіз результатів тестування програмного модуля

Окремо було виконано перевірку роботи програмного модуля за умов дуже зниженого освітлення. Такий тест дає змогу оцінити, чи зберігається працездатність візуального каналу тоді, коли якість зображення погіршується. Результат запуску програми для другого тестового відеофайлу наведено на рисунку 4.3.

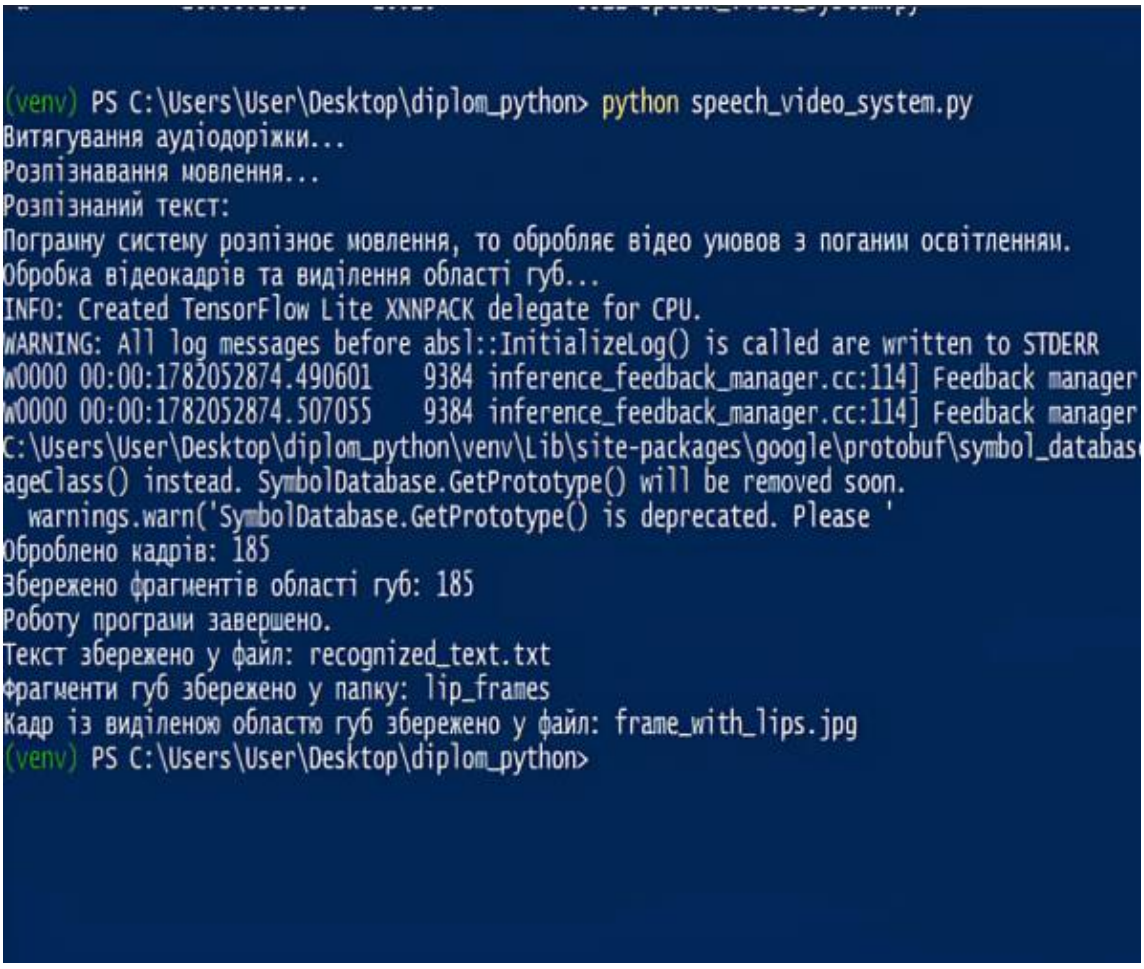


Рисунок 4.3 – Результат виконання програмного модуля для відео зі зниженим освітленням

Як видно з рисунку 4.3, програма завершила обробку відеофайлу, розпізнала мовлення та зберегла фрагменти області губ диктора. Для цього тесту було оброблено 185 кадрів і збережено 185 фрагментів області губ. Отже, зниження освітлення не зупинило роботу візуального модуля, хоча в текстовому результаті розпізнавання з’явилися окремі неточності.

Для третьої перевірки було використано відеофайл зі складнішими умовами запису. Знижене освітлення поєднувалося з тихішим мовленням і незначним поворотом голови диктора. Результат виконання програмного модуля для третього тестового відеофайлу наведено на рисунку 4.4.

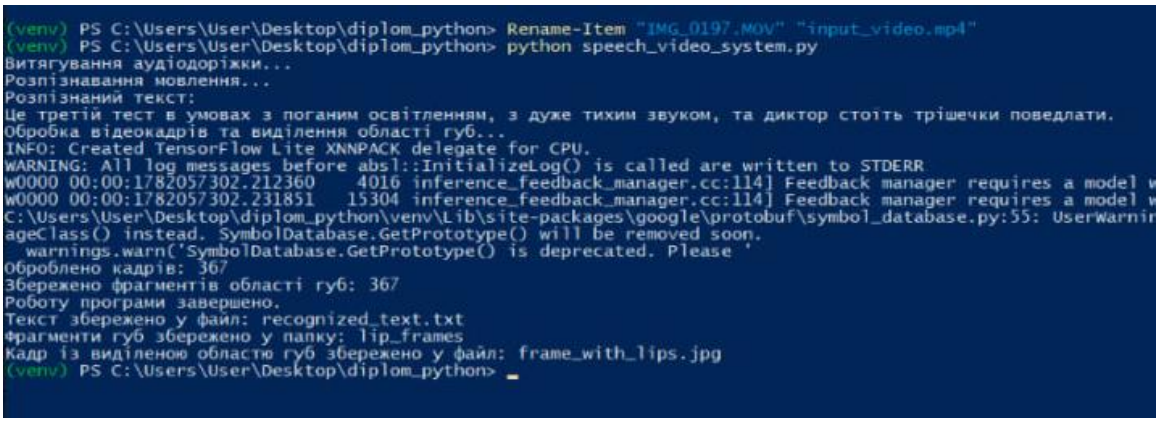


Рисунок 4.4 – Результат виконання програмного модуля для третього тестового відеофайлу

Результат третього запуску показав, що програма продовжила обробку відеокадрів і змогла зберегти фрагменти області губ навіть за складніших умов. Було оброблено 367 кадрів і збережено 367 фрагментів області губ. Незначний поворот голови та знижене освітлення не завадили роботі візуального модуля, оскільки обличчя диктора залишалося видимим для обробки. Текстовий результат у цьому тесті зберіг основний зміст вимовленої фрази, хоча умови запису були складнішими, ніж у першому тесті.

Результати всіх трьох тестів зведено в таблицю 4.1.

Таблиця 4.1 – Результати тестування програмної системи

№ тесту	Умови запису відео	Розпізнаний текст	Кількість оброблених кадрів	Результат виділення області губ
1	Нормальне освітлення, фронтальне положення обличчя, чітке мовлення	Всім привіт! Це тестове відео для перевірки роботи програмного коду.	169	Область губ визначена та збережена стабільно
2	Знижене освітлення, фронтальне положення обличчя, мовлення чітке	Програму систему розпізнює мовлення, то обробляє відео умовов з поганим освітленням.	185	Область губ визначена та збережена; у текстовому результаті наявні окремі неточності
3	Знижене освітлення, тихіше мовлення, незначний поворот голови диктора	Це третій тест в умовах з поганим освітленням, з дуже тихим звуком, та диктор стоїть трішечки повернута.	367	Область губ визначена та збережена; складніші умови не зупинили роботу візуального модуля

Для оцінювання стабільності роботи візуального модуля було використано показник повноти визначення області губ у відеокадрах:

$$K_{lip} = (N_{lip} / N_{frames}) \cdot 100\% \quad (4.1)$$

де K_lip — показник повноти визначення області губ; N_lip — кількість кадрів, у яких було збережено фрагмент області губ; N_frames — загальна кількість оброблених кадрів відео.

ВИСНОВКИ

Дослідження було спрямоване на вивчення нейромережевих методів розпізнавання мовлення за аудіосигналом та візуальними даними обличчя диктора, а також на створення програмного модуля для первинної аудіовізуальної обробки відеофайлу. Актуальність теми пов’язана з активним використанням мовних технологій у системах транскрибування, голосових інтерфейсах, автоматичному створенні субтитрів, перекладі та людино-машинній взаємодії. Особливе значення має поєднання аудіального та візуального каналів, оскільки в реальних умовах якість звуку та зображення може змінюватися.

Під час розгляду теоретичних основ встановлено, що автоматичне розпізнавання мовлення є складним процесом, який включає отримання мовного сигналу, його попередню обробку, виділення ознак і формування текстового результату. Якість такого розпізнавання залежить не лише від обраної моделі, а й від умов запису, рівня шуму, гучності мовлення, чіткості вимови, якості мікрофона та індивідуальних особливостей диктора. Через це одна й та сама система може показувати різні результати залежно від вхідних даних.

У роботі розглянуто розвиток нейромережевих підходів до розпізнавання мовлення. Класичні методи поступово поступилися місцем моделям глибокого навчання, оскільки нейронні мережі краще працюють зі складними залежностями у мовному сигналі та можуть обробляти великі обсяги даних. Було проаналізовано такі моделі, як DeepSpeech, DeepSpeech 2, wav2vec 2.0 та Whisper. DeepSpeech і DeepSpeech 2. Ці моделі демонструють розвиток end-to-end підходу, та модель wav2vec 2.0 показує можливості самонавчання на нерозмічених аудіоданих, а Whisper є прикладом готової трансформерної моделі, придатної для практичного використання.

Окрему увагу приділено візуальним даним обличчя диктора. Встановлено, що рухи губ, положення нижньої щелепи та зміна області рота можуть містити додаткову інформацію про вимовлене слово або фразу. Водночас візуальний

канал не є повною заміною аудіального, оскільки багато звуків мають схоже візуальне відображення. Для його використання необхідно коректно визначити обличчя, знайти ключові точки та виділити область губ. Саме цей етап є важливим для підготовки відеоданих до подальшого аналізу.

Порівняння аудіального, візуального та аудіовізуального підходів показало, що кожен із них має власні переваги й обмеження. Аудіальний підхід забезпечує основний текстовий результат, але залежить від якості звуку. Візуальний підхід дозволяє враховувати артикуляційні ознаки, однак залежить від освітлення, ракурсу, якості відео та видимості області губ людини. Аудіовізуальний підхід є більш гнучким та надійним, оскільки використовує два джерела даних одного запису. Мною було обрано схему, у якій аудіальний канал відповідає за розпізнавання мовлення, а відеоканал — за виділення області губ диктора.

На основі проведеного аналізу розроблено програмний модуль мовою Python. Його робота побудована навколо одного вхідного відеофайлу, який містить аудіодоріжку та послідовність відеокадрів із зображенням обличчя диктора. Для виділення аудіодоріжки використовується MoviePy, для розпізнавання мовлення — Whisper, для покадрової обробки відео — OpenCV, а для визначення ключових точок обличчя — MediaPipe Face Mesh. У результаті програма формує текстовий результат розпізнавання, кадр із позначеною областю губ та окремі фрагменти області губ.

Під час реалізації було враховано важливе обмеження: створений модуль не виконує повноцінне розпізнавання мовлення за рухом губ. Візуальний канал використовується для пошуку обличчя, визначення ключових точок і виділення області губ, а не для самостійного перетворення артикуляції в текст. Текстовий результат формується за аудіодоріжкою за допомогою Whisper. Такий підхід відповідає фактичній реалізації програми та не завищує її функціональні можливості.

Тестування програмного модуля було проведено на трьох відеофайлах із різними умовами запису. Перший тест виконувався за нормального освітлення,

фронтального положення обличчя та чіткого мовлення. У цьому випадку програма обробила 169 кадрів, зберегла 169 фрагментів області губ і сформувала коректний текстовий результат. Це підтвердило стабільну роботу аудіального та візуального каналів за сприятливих умов.

Другий тест проводився за зниженого освітлення. Програма обробила 185 кадрів і зберегла 185 фрагментів області губ. Візуальний канал продовжив коректно визначати область рота, оскільки обличчя диктора залишалося видимим у кадрі. Водночас у текстовому результаті з’явилися окремі неточності, що свідчить про залежність аудіального розпізнавання від якості запису та чіткості мовлення.

Третій тест включав складніші умови: знижене освітлення, тихіше мовлення та незначний поворот голови диктора. За таких умов програма обробила 367 кадрів і зберегла 367 фрагментів області губ. Це показало, що відеоканал може працювати стабільно за умови, що обличчя не перекрите і залишається в кадрі. Якість текстового розпізнавання при цьому залежить від аудіосигналу, зокрема від гучності та чіткості вимови.

Порівняння результатів трьох тестів показало, що програмний модуль коректно виконує основну послідовність дій: виділяє аудіодоріжку, формує текстовий результат, обробляє відеокадри та зберігає фрагменти області губ. Найкращий результат було отримано у першому тесті, де умови запису були найбільш сприятливими. Найбільші неточності в текстовому розпізнаванні проявилися у другому тесті, записаному за зниженого освітлення та гіршої якості зображення. Водночас відеоканал у цьому випадку не припинив роботу і зміг зберегти фрагменти області губ. Це свідчить про те, що модуль здатний виконувати відеообробку навіть за ускладнених умов, хоча якість загального результату все одно залежить від умов запису та характеристик вхідного файлу.

Отримані результати дозволили визначити основні обмеження реалізованого рішення. Для аудіального каналу важливими є якість запису, гучність голосу, відсутність сильного шуму та чіткість вимови. Для відеоканалу значення мають освітлення, положення обличчя, відсутність перекриття області

губ і достатня якість кадру. Якщо обличчя сильно повернуте, частково закрите або розмите, точність виділення області губ може знизитися.

Практичний результат роботи полягає у створенні програмного модуля, який поєднує аудіальну та візуальну обробку мовлення в одному Python-рішенні. Модуль дозволяє отримувати текст за аудіодоріжкою відеофайлу та паралельно формувати візуальні дані з області губ диктора. Це може бути основою для подальших досліджень, пов’язаних із глибшим поєднанням аудіо- та відеоознак.

Подальше вдосконалення модуля може бути пов’язане з покращенням попередньої обробки аудіосигналу, зменшенням шумів, стабілізацією відеокадрів і точнішим виділенням області губ людини. Перспективним напрямом є також використання окремої моделі для аналізу руху губ, яка могла б працювати разом з аудіальним розпізнаванням. У цьому випадку візуальний канал міг би використовуватися не лише для підготовки зображень фрагментів губ, а й як повноцінне джерело інформації під час розпізнавання мовлення.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. LeCun Y., Bengio Y., Hinton G. Deep Learning. Nature. 2015. Vol. 521. P. 436–444.

2. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. Attention Is All You Need. Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. P. 5998–6008.

3. Graves A., Fernández S., Gomez F., Schmidhuber J. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. Proceedings of the 23rd International Conference on Machine Learning. 2006. P. 369–376.

4. Radford A., Kim J. W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv preprint arXiv:2212.04356. 2022. URL: https://arxiv.org/abs/2212.04356

5. Shi B., Hsu W.-N., Lakhotia K., Mohamed A. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction. arXiv preprint arXiv:2201.02184. 2022. URL: https://arxiv.org/abs/2201.02184

6. MediaPipe Face Mesh. MediaPipe Documentation. URL: https://mediapipe.readthedocs.io/en/latest/solutions/face_mesh.html

7. Assael Y. M., Shillingford B., Whiteson S., de Freitas N. LipNet: End-to-End Sentence-level Lipreading. arXiv preprint arXiv:1611.01599. 2016. URL: https://arxiv.org/abs/1611.01599

8. Chung J. S., Senior A., Vinyals O., Zisserman A. Lip Reading Sentences in the Wild. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. P. 6447–6456. URL: https://arxiv.org/abs/1611.05358

9. Hannun A., Case C., Casper J., Catanzaro B., Diamos G., Elsen E., Prenger R., Satheesh S., Sengupta S., Coates A., Ng A. Y. Deep Speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567. 2014. URL: https://arxiv.org/abs/1412.5567

10. Amodei D., Anubhai R., Battenberg E., Case C., Casper J., Catanzaro B., Chen J., Chrzanowski M., Coates A., Diamos G., Elsen E., Ng A. Y. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. arXiv preprint arXiv:1512.02595. 2015. URL: https://arxiv.org/abs/1512.02595

11. Baevski A., Zhou H., Mohamed A., Auli M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Advances in Neural Information Processing Systems. 2020. URL: https://arxiv.org/abs/2006.11477

12. Baltrušaitis T., Zadeh A., Lim Y. C., Morency L.-P. OpenFace 2.0: Facial Behavior Analysis Toolkit. 13th IEEE International Conference on Automatic Face & Gesture Recognition. 2018.

13. Chung J. S., Zisserman A. Lip Reading in the Wild. Computer Vision – ACCV 2016. Lecture Notes in Computer Science. Vol. 10112. Springer, 2017. P. 87–103.

14. OpenCV Documentation. URL: https://docs.opencv.org/

15. MoviePy Documentation. URL: https://zulko.github.io/moviepy/