МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Західноукраїнський національний університет
Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно-обчислювальних систем і управління

ОЛЕКСИШИН Євген Олександрович

Модуль озвучення тексту мальописів / Module for comic book text narration

Спеціальність 122 – Комп'ютерні науки
Освітньо-професійна програма – Комп'ютерні науки

Кваліфікаційна робота

Виконав студент групи КН-41
Є. О. Олексишин

Науковий керівник:
к.т.н., доцент
Х.В. Ліп'яніна-Гончаренко

Кваліфікаційну роботу допущено до захисту
«___»_____ 2025 р.

В.о. завідувача кафедри
_____ Н.М. Васильків

Тернопіль – 2025

Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно-обчислювальних систем і управління
Освітній ступінь «бакалавр»
спеціальність 122 – Комп'ютерні науки
освітньо-професійна програма – Комп'ютерні науки

ЗАТВЕРДЖУЮ
В.о. завідувача кафедри
_____ Н.М. Васильків
«____» _____2024 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ
ОЛЕКСИШИН Євген Олександрович
(прізвище, ім'я, по батькові)

1. Тема кваліфікаційної роботи Модуль озвучення тексту мальописів / Module for comic book text narration
керівник роботи Х. В. Ліп'яніна-Гончаренко, к. т. н., доцент
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)
затверджені наказом по університету від 20 грудня 2024 р. №938
2.Строк подання студентом закінченої кваліфікаційної роботи 25 травня 2025 р.
3.Вихідні дані до кваліфікаційної роботи: завдання на кваліфікаційну роботу студента, наукові статті, технічна література.
4.Основні питання, які потрібно розробити:
- проаналізувати сучасні підходи до OCR та TTS у контексті коміксів;
- сформувати інформаційну модель даних, необхідну для зберігання зображень, тексту та аудіо;
- розробити та реалізувати алгоритм вилучення й попередньої обробки тексту з панелей коміксу;
- інтегрувати TTS, що підтримує україномовний та англомовний синтез;
- створити веб-інтерфейс із drag-and-drop завантаженням зображень і вбудованим аудіоплеєром;
- провести функціональне й експериментальне тестування модуля, оцінити точність OCR, час обробки та суб'єктивну якість мовлення MOS-методом.
5. Перелік графічного матеріалу у роботі:
- діаграма дерева функцій;
- діаграма класів;
- діаграма модулів;
- схема алгоритму розпізнавання тексту;
- схема алгоритму обробки тексту;
- схема алгоритму озвучення тексту;
- схема алгоритму модуля.

6. Консультанти розділів кваліфікаційної роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		Завдання видав	Завдання прийняв

7. Дата видачі завдання 20 грудня 2024 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів кваліфікаційної роботи	Примітка
1	Затвердження теми кваліфікаційної роботи, ознайомлення з літературними джерелами та складання плану роботи	до 01.01. 2025 р.	
2	Написання 1 розділу кваліфікаційної роботи	до 01.02. 2025 р.	
3	Написання 2 розділу кваліфікаційної роботи	до 01.04.2025 р.	
4	Написання 3 розділу кваліфікаційної роботи	до 01.05. 2025 р.	
5	Представлення попереднього варіанту кваліфікаційної роботи, перевірка та внесення змін керівником	до 15.05.2025 р.	
6	Опрацювання зауважень та представлення завершеного варіанту кваліфікаційної роботи. Підготовка супроводжуючих документів	до 25.05.2025 р.	
7	Перевірка кваліфікаційної роботи на оригінальність тексту	до 30.05.2025 р.	
8	Оформлення кваліфікаційної роботи та отримання допуску до захисту	до 10.06.2025 р.	
9	Подання кваліфікаційної роботи до захисту на засіданні атестаційної комісії	до 10.06. 2025 р.	

Студент _____ Є. О. Олексишин

(підпис) (прізвище та ініціали)

Керівник кваліфікаційної роботи _____ Х. В. Ліп’яніна-Гончаренко

(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Кваліфікаційна робота на тему «Модуль озвучення тексту мальописів» на здобуття освітнього ступеня «бакалавр» зі спеціальності 122 «Комп’ютерні науки» освітньої програми «Комп’ютерні науки» написана обсягом в 66 сторінок і містить 26 ілюстрацій, 3 таблиці, 2 додатки та 28 використаних джерел.

Метою роботи є розробка веб-модуля, який автоматично перетворює текст з мальописів (коміксів) на природне аудіо, забезпечуючи його доступність і зручність мультимедійного сприйняття.

Методами розроблення обрано: методи комп’ютерного зору (для виявлення текстових фрагментів на зображеннях), методи машинного навчання (для синтезу мовлення) та підходи модульного проєктування у середовищі Django.

У результаті виконання роботи розроблено функціональний веб-модуль, що забезпечує завантаження зображень, розпізнавання тексту, його обробку та озвучення з підтримкою української та англійської мов. Проведено експериментальне тестування модуля з оцінкою точності OCR, часу обробки та якості мовлення за MOS-методом.

Результати дослідження можуть бути використані у створенні інтерактивних мультимедійних сервісів, освітніх продуктів та інклюзивних застосунків.

Ключові слова: МАЛЬОПИС, TTS, OCR, DJANGO, ОЗВУЧЕННЯ ТЕКСТУ.

ANNOTATION

Qualification work on the topic «Module for comic book text narration» for obtaining the Bachelor's degree in specialty 122 «Computer Science» of the educational program «Computer Science» is written on 66 pages and contains 26 figures, 3 tables, 2 appendices, and 28 references.

The purpose of the work is to develop a web module that automatically converts comics text into natural audio, ensuring accessibility and convenience of multimedia perception.

The development methods include computer vision techniques (for detecting text fragments on images), machine learning methods (for speech synthesis), and modular software design approaches within the Django framework.

As a result of the work, a functional web module was developed that supports image upload, text recognition, preprocessing, and voice-over with support for both Ukrainian and English languages. Experimental testing of the module was conducted, including evaluation of OCR accuracy, processing time, and speech quality using the MOS method.

The results of the study can be applied in the development of interactive multimedia services, educational tools, and inclusive applications.

Keywords: COMICS, TTS, OCR, DJANGO, TEXT VOICE-OVER.

ЗМІСТ

ВСТУП

Сьогоднішній день – це епоха стрімкого розвитку інформаційних технологій та зростаючого попиту на інтерактивний мультимедійний контент, особливого значення набуває розробка інноваційних інструментів, що забезпечують нові форми взаємодії користувача з цифровими матеріалами. Одним із таких напрямів є озвучення текстової інформації з мальописів, чи більш відома назва – комікс [9], що відкриває нові можливості для сприйняття візуального контенту та покращення його доступності. Центральну роль у розробці таких рішень відіграють сучасні вебтехнології [3] та інструменти штучного інтелекту з класичними [1] та сучасними [2] підходами до Text-to-Speech (TTS). Провідні компанії, зокрема Google, Amazon, Microsoft активно досліджують і розробляють TTS-системи, що інтегрують глибинне навчання й моделі синтезу мовлення високої якості. Перші системи синтезу мовлення почали з'являтись ще в середині XX століття, а сьогодні TTS-технології використовуються у багатьох сферах – від освіти до медицини – і активно розвиваються в Україні, де з 2016 року з'явилися перші професійні україномовні голоси на базі Google WaveNet [10] та RHVoice [8]. Попри це, прогалини залишаються: існуючі TTS-рішення рідко адаптовані для озвучення текстів із мальописів, не враховують контекст, а також слабо інтегруються у мультимедійні платформи з графічним контентом. Світові тенденції вирішення цих задач спрямовані на розвиток мультимодальних моделей, персоналізацію озвучення, генерацію аудіо з урахуванням контексту, а також забезпечення доступності таких технологій для розробників завдяки відкритим API та фреймворкам.

Історично спроби реалізації автоматизованого озвучення мальописів стикалися з численними викликами: потреба точно виділяти текстові фрагменти з графічних матеріалів, обробка великих обсягів даних у реальному часі та генерація якісного мовлення, адаптованого до контексту сцени-зображення. Завдяки впровадженню технологій комп'ютерного зору, глибинного навчання та синтезу мовлення, стало можливим ефективно вирішувати ці проблеми. Зокрема, використання моделей машинного навчання дозволяє автоматизувати розпізнавання

тексту на зображеннях [24] і його озвучення [25], що значно підвищує ефективність створення інтерактивного мультимедійного контенту.

Розробка ефективного рішення для озвучення тексту з мальописів вимагає використання надійного й гнучкого фреймворку. Django [13] – сучасна технологія для створення вебзастосунків на мові програмування Python [12] – забезпечує потужну платформу для інтеграції нейронних мереж [14], обробки мультимедійних даних і реалізації зручного інтерфейсу користувача. Завдяки вбудованим засобам для роботи з базами даних, маршрутизації запитів та управління сесіями, Django дозволяє сконцентруватися на розробці функціоналу, пов’язаного з обробкою текстової інформації та створенням аудіофайлів. У межах проходження практики було опановано першочергові етапи розробки повноцінного вебпроєкту: від налаштування серверного оточення й проєктування бази даних до створення користувацького інтерфейсу, реалізації авторизації, інтеграції нейронних мереж і впровадження інтелектуальних компонентів на основі машинного навчання.

Незважаючи на стрімкий розвиток технологій, впровадження озвучення мальописів залишається непростим завданням. Серед основних викликів – інтеграція нейронних мереж у вебсередовище, забезпечення стабільної роботи при обробці ресурсомістких задач, а також побудова масштабованої інфраструктури для зберігання й передачі аудіофайлів. Практична підготовка, що включала створення вебзастосунку на Django з використанням компонентів штучного інтелекту, дозволила здобути необхідний досвід, який став підґрунтям для реалізації кваліфіквційного проєкту на тему «Модуль озвучення тексту мальописів». Такий проєкт має перспективу застосування в освітніх, розважальних та адаптаційних цифрових продуктах, спрямованих на розширення доступності контенту.

Мета роботи - розробити веб-модуль, який автоматично перетворює текст мальописів на природне аудіо, забезпечуючи доступність і зручність мультимедійного сприйняття.

Завдання дослідження:

1. Проаналізувати сучасні підходи до OCR та TTS у контексті коміксів.
2. Сформувати інформаційну модель даних, необхідну для зберігання

зображень, тексту та аудіо.

3. Розробити та реалізувати алгоритм вилучення й попередньої обробки тексту з панелей коміксу.

4. Інтегрувати TTS, що підтримує україномовний та англомовний синтез.

5. Створити веб-інтерфейс із drag-and-drop завантаженням зображень і вбудованим аудіоплеєром.

6. Провести функціональне й експериментальне тестування модуля, оцінити точність OCR, час обробки та суб’єктивну якість мовлення MOS-методом.

Предмет дослідження – процеси автоматизованого розпізнавання й озвучення тексту мальописів у веб-середовищі.

Об’єкт дослідження – цифрові зображення сторінок коміксів, їх текстові компоненти та аудіофайли, згенеровані на основі цих текстів.

Методи дослідження. Для досягнення поставленої мети використано методи комп’ютерного зору (CNN-орієнтований OCR для детекції «бульбашок» та панелей), обробки природної мови (правила нормалізації, пунктуаційний та просодичний аналіз), машинного навчання (нейронні мережі ESPnet-подібного TTS), а також із застосуванням шаблонів архітектурного та об'єктно-орієнтованого проєктування (MVT, Facade, Strategy) у середовищі Django. Ефективність рішень перевірено експериментально: виміряно точність розпізнавання (char-level accuracy), середній час конвеєра «зображення → аудіо», а якість синтезу оцінено слухацьким тестом MOS за участі п’яти респондентів.

Апробація результатів дослідження здійснена в межах студентської науково-практичної конференції «Інтелектуальні інформаційні технології в прикладних дослідженнях» (IITAR-2025), яка відбулася в місті Тернополі 27–29 травня 2025 року.

1 СТАН ТА АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Опис предметної області

Розробка технології, що дозволяє озвучувати кожну сторінку коміксу чи манги, забезпечує передачу емоцій, атмосфери та стилістичних особливостей художнього твору. Модуль озвучення тексту мальописів, орієнтований на аудіоінтерпретацію коміксів та манги, має на меті не лише передати інформаційне наповнення тексту, а й відтворити його художню атмосферу, що є особливо важливим у творах візуального мистецтва. Цей підхід дозволяє створити унікальний досвід для слухача, який може зануритися в історію, навіть не відкриваючи сам комікс, мангу чи будь-який інший мальопис.

Розробка такого модуля вимагає комплексного розуміння сучасних технологій синтезу мовлення, а й застосування методів комп'ютерного зору (CV) [11] для аналізу візуальних елементів, зі специфікою художньої мови, властивої коміксам та манзі. Важливо врахувати, що текст даного жанру часто містять елементи сленгу, діалогічну мову, а також особливості візуального кадрування сцени, які потребують певних підходів, наприклад обробки природної мови та комп'ютерного зору, під час озвучення. Модуль повинен відображати контекст, інтонаційні нюанси та ритмічність мови, щоб кожен голосовий фрагмент максимально точно відображав динаміку сцени.

Модуль озвучення мальописів відкриває широкі можливості для використання в різних сферах, починаючи від мультимедійних платформ до освітніх інструментів і розважальної індустрії. Враховуючи популярність коміксів та манги в сучасному світі, технологія озвучення стає важливою складовою для створення більш інтерактивного і захоплюючого досвіду для користувачів.

Головною особливістю є можливість адаптації голосового синтезу до різних стилів, характерних для окремих жанрів та авторських задумів. Від драматичних монологів до легковажних діалогів – кожна ситуація вимагає унікального підходу. Звук повинен відповідати не тільки текстовому наповненню, але й візуальній

складовій твору, тому інтеграція озвучки із зображеннями дозволяє досягти нового рівня взаємодії з аудиторією.

Модуль озвучення тексту мальописів є своєрідною ланкою між візуальним мистецтвом та слуховим сприйняттям, відкриваючи нові горизонти для творчості та культурного обміну. Він зробить твори доступними для тих, хто віддає перевагу аудіоформату, дозволяє зануритись у «читання» навіть за відсутності зорового контакту з зображеннями, і, зрештою, підкреслює універсальність сучасних технологій у збереженні та популяризації культури.

У фундаменті програмного забезпечення, яке забезпечує озвучку коміксів та манги, лежить сучасна архітектура, побудована на базі Django – потужного веб-фреймворку, який спрощує розробку. Використання Django дозволяє зосередитись на основній логіці додатку, адже цей фреймворк надає зручні засоби для роботи з базами даних, маршрутизації запитів та управління сесіями користувачів. Завдяки цьому можна швидко впроваджувати нові функціональні можливості, адаптуючи модуль до вимог сучасного ринку та змін у технологічному середовищі.

Інтеграція Django у проєкт не лише оптимізує процес розробки, але й надасть можливість масовій аудиторії використовувати модуль для власних потреб. Серверна частина додатку відповідає за обробку запитів від користувачів, управління файлами, що містять як текстові дані, так і аудіофайли, а також за взаємодію з модулями синтезу мовлення, побудованими на основі алгоритмів глибокого навчання.

У сучасному світі цифрові мальописи все частіше захоплюють в різноманітні мультимедійні платформи та мобільні додатки. Озвучення мальописів може стати важливою складовою цих сервісів, забезпечуючи користувачам новий рівень взаємодії з контентом. Модуль озвучення дозволить не тільки читати текст, а й слухати його, що зробить сприйняття матеріалу більш багатогранним і захоплюючим. Технологія може бути використана для створення аудіокоміксів, де озвучення персонажів і звукові ефекти допомагають краще відчути атмосферу історії. Такий формат приваблює не тільки любителів мальописів, але й людей з обмеженими можливостями або тих, хто просто хоче спробувати новий формат

споживання контенту.

Сфера освіти також може виграти від інтеграції модуля озвучення мальописів. Уроки, засновані на мальописах, можуть стати цікавими та інтерактивними для учнів, зокрема, коли йдеться про літературні твори чи історичні події, представлені через подібні формати. Модуль озвучення дозволяє перетворити текст на аудіовізуальний досвід, що зробить матеріал доступнішим для учнів різного віку. Наприклад, діти, які тільки вчаться читати, можуть слухати озвучення та одночасно слідкувати за текстом, що допомагає покращити навички читання та сприйняття. Це також може бути корисним для студентів, які вивчають іноземні мови, оскільки мова мальописів зазвичай проста і діалогічна, що спрощує розуміння.

Однією з найбільш важливих сфер застосування модуля озвучення є забезпечення доступності контенту для людей з обмеженими можливостями. Для осіб із порушеннями зору або тих, хто має труднощі з читанням, мальописи можуть бути недоступними, адже важливу роль в їх сприйнятті відіграють як текст, так і візуальні елементи. Модуль озвучення дозволяє вирішити цю проблему, оскільки з його допомогою мальописи можуть стати доступними для людей, які не можуть повною мірою сприймати зображення. Озвучення діалогів, звукові ефекти та описання сцени створюють можливість для інтеграції мальописів у повсякденне життя людей з інвалідністю, що значно підвищує рівень інклюзивності та доступності культурних продуктів.

У розважальній індустрії озвучення мальописів відкриває нові горизонти для створення аудіо- та відеопродукції. Адаптація коміксів, манги, манхви у вигляді аудіокниг чи відеоігор з інтеграцією голосів персонажів та звукових ефектів створює додаткову цінність для споживачів. Важливо, що мальописні історії часто мають великий фанатський потенціал, і їх озвучення дозволяє створювати нові форми продукції, що відповідають вимогам сучасної аудиторії. Модуль озвучення може бути інтегрований в адаптації у вигляді анімацій або віртуальних світів, що дозволить залучити ще більше фанатів, створюючи мультимедійний досвід з високим рівнем занурення.

Модуль озвучення тексту мальописів відкриває безліч можливостей для розвитку контенту, який вже давно здобув популярність серед різних вікових груп і культур. Його застосування в мультимедійних платформах, освітніх додатках, інклюзивних проєктах та розважальних індустріях дозволяє створювати нові формати споживання контенту, що відповідають сучасним вимогам суспільства. Однак найважливіше – це те, що такий підхід дозволяє зробити мальописи доступнішими для всіх, незалежно від фізичних можливостей чи технічних знань користувачів.

Модуль озвучення мальописів є комплексним бізнес-процесом, що включає низку функціональних етапів, кожен з яких має свою специфічну роль у створенні інтерактивного аудіо контенту. Цей процес можна розділити на кілька головних кроків, що забезпечують повний цикл перетворення візуального контенту в аудіо формат:

1. Першим етапом є завантаження зображень мальопису від користувача. Завантаження може відбуватися через веб-інтерфейс, що дозволяє зручно інтегрувати цю функцію в загальний робочий процес. Важливо, щоб модуль міг обробляти файли різних форматів і забезпечувати їхню належну оптимізацію для подальшої обробки.

2. Наступним кроком є застосування технологій оптичного розпізнавання символів (OCR) [17]. Цей етап дозволяє автоматично витягнути текстову інформацію зі зображень. Якісне розпізнавання є критично важливим, оскільки від точності цього процесу залежить успішність подальших операцій: від класифікації до озвучення. Сучасні алгоритми OCR, підтримані нейронними мережами, забезпечують високу точність і швидкість обробки, навіть при наявності шуму або складної композиції зображення.

3. Перед безпосереднім процесом синтезу мовлення текст проходить етап обробки, що включає його нормалізацію, корекцію орфографічних помилок, видалення непотрібних символів та адаптацію під правила вимови. Цей етап також може містити додаткові елементи, такі як виявлення пунктуації та тонких нюансів синтаксису, що впливають на просодію – ритм, наголос та інтонацію озвучення.

4. Після підготовки тексту та визначення оптимального голосу запускається процес генерації аудіо. Завдяки використанню сучасних алгоритмів, синтезоване мовлення наближається до якості людської вимови, забезпечуючи високу зрозумілість і виразність.

5. Крайній етап процесу – це забезпечення доступу користувача до згенерованого аудіо. Аудіофайл може відтворюватися безпосередньо через веб-плеєр або мобільний застосунок, а також бути експортований для подальшого використання у інших мультимедійних продуктах. Це забезпечує гнучкість використання модуля, дозволяючи інтегрувати озвучене відео у презентації, навчальні матеріали або рекламні ролики.

Отож бізнес-процес модуля озвучення мальописів являє собою інтегрований ланцюг, де кожен етап тісно пов’язаний з іншими, який забезпечує точність і якість кінцевого аудіоконтенту. Цей комплексний підхід дозволяє не лише автоматизувати процес озвучення, а й значно покращити доступність і сприйняття візуального контенту, що відкриває нові можливості для інтерактивних застосувань мультимедія.

Модуль озвучення тексту мальописів – це компонент, який перетворює текст з коміксів, манги, манхви та інших подібних форматів у сучасну форму сприйняття художнього контенту. Цей модуль об’єднує декілька важливих функціональних блоків, що взаємодіють між собою для досягнення високої якості кінцевого аудіо продукту. На рисунку 1.1 зображено діаграму дерева функцій.

Рисунок 1.1 – Діаграма дерева функцій

Перший блок відповідає за перетворення зображень сторінок мальопису у текст, який можна зчитати. Етап починається з передобробки зображень, що дозволяє оптимізувати якісний стан вхідних даних. Далі застосовується технологія оптичного розпізнавання символів.

Другий блок спеціалізується на удосконаленні даних після розпізнавання. ут текст нормалізується та форматуються, проводиться лінгвістичний аналіз для виявлення синтаксичних і стилістичних особливостей, що характерні для діалогів та описів у коміксах. Також здійснюється фільтрація для усунення небажаних символів та шуму, що гарантує чистоту тексту перед його озвученням.

Третій блок перетворює оброблений текст у звукову форму за допомогою технологій синтезу мовлення. Завершальним етапом є збереження аудіо файлів для подальшого використання.

Останній блок забезпечує взаємодію між модулем та користувачем. Інтуїтивно зрозумілий інтерфейс дозволяє запускати процес озвучення, контролювати його перебіг, переглядати результати та налаштовувати параметри модулю відповідно до індивідуально заданих вимог. Це дозволяє користувачу легко інтегрувати модуль в робочі процеси та забезпечити зручне відтворення аудіо результату.

Транзакційна складова модулю відповідає за збір, накопичення та обробку кількісних даних, що стосується стану об’єкта обробки. Цей процес включає кілька етапів:

1. Завантаження зображень сторінок мальопису.

2. Збереження розпізнаного тексту з-за допомогою технологій OCR.

3. Попередня підготовка тексту до озвучення, що включає нормалізацію та форматизацію.

4. Синтез аудіо для тексту.

5. Формування історії озвучень, яка є базою для подальшого аналізу.

Цей процес дозволяє забезпечити управління всіма транзакціями, що відбуваються у модулі, що гарантує надійне накопичення даних, необхідних для аналізу ефективності роботи модуля.

Аналітична складова спрямована на дослідження кількісних показників, сформованих у транзакційній частині, з метою забезпечення інформаційної підтримки прийняття рішень. Це дозволяє проводити аналіз у різних розрізах: аналіз динаміки часу генерації аудіо та швидкості розпізнавання тексту, що допомагає оптимізувати робочі процеси, оцінка ефективності озвучення для різних типів тексту, що дозволяє ідентифікувати найбільш якісні алгоритми обробки, вивчення зворотного зв'язку користувачів та потреб окремих сегментів аудиторії.

Проведення багатоаспектного аналізу дозволяє виявити слабкі місця в роботі модуля, наприклад, місця, де розпізнавання дає численні помилки, або де голос озвучення не відповідає очікуванням користувачів. Отримані дані стають основою для прийняття стратегічних рішень, спрямованих на подальше вдосконалення модуля.

Отже, модуль озвучення тексту мальописів перетворює текстовий контент з коміксів, манги, манхви та інших подібних художньої літератури в аудіоматеріали, які роблять ці твори доступними для нової аудиторії. Такий підхід дозволяє розширити можливості сприйняття художньо-розважального контенту за рахунок зручного формату.

1.2 Огляд і аналіз існуючих рішень

Сучасний ринок технологій озвучення тексту насичений інноваційними рішеннями, які здатні перетворити будь-який текстовий матеріал у природне та емоційно забарвлене мовлення. Кожен з них має свої особливості, які визначають як якість кінцевого результату, так і рівень інтеграції в існуючі інформаційні системи, що робить їх привабливими для широкого спектру застосувань.

TTSMAKER є безкоштовним інструментом синтезу мовлення, який надає зазначені послуги та підтримує кілька мов, зокрема англійську, французьку, німецьку, іспанську, арабську, китайську, японську, корейську, в'єтнамську тощо, а також різні стилі голосу. Його можна використовувати для читання тексту та електронних книг вголос або завантажувати аудіофайли для комерційного

використання. TTSMaker може легко перетворювати текст на мовлення онлайн [4]. Сервіс зображено на рисунку 1.2.

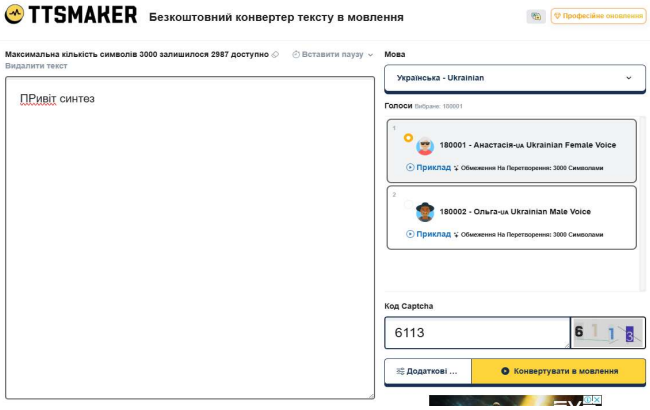

Рисунок 1.2 – Сервіс TTSMAKER

Apple Books, який зображено на рисунку 1.3 вже давно встановився як один із лідерів у сфері електронного читання, пропонуючи користувачам зручний інтерфейс, широкий вибір творів та інноваційні можливості для поглибленого занурення у читання.

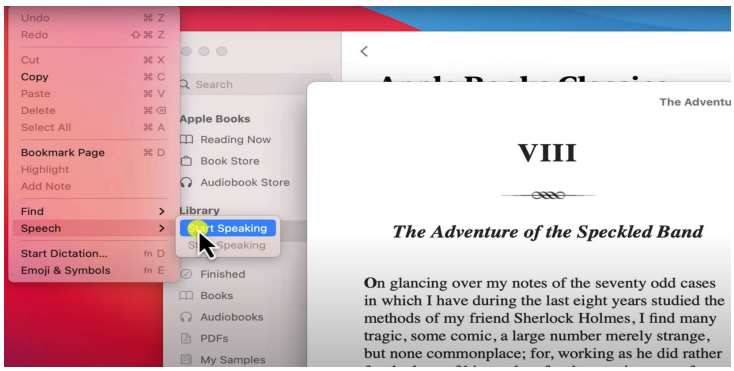

Рисунок 1.3 – Сервіс Apple Books

Однією з найцікавіших функцій є Digital Narrator – технологія, що змінює спосіб, у який ми сприймаємо текст, перетворюючи його на живе аудіо. Digital Narrator в Apple Books дозволяє користувачам насолоджуватись літературою навіть у тих випадках, коли читання традиційного тексту може бути ускладненим. Це особливо важливо для людей з вадами зору або тих, хто віддає перевагу аудіокнигам. Завдяки передовим алгоритмам синтезу мовлення, технологія відтворює інтонації, паузи та емоційне забарвлення, що дозволяє максимально

наблизити звучання до людської мови. Цей підхід не лише покращує доступність, але й відкриває нові можливості для творчості, дозволяючи авторам та видавцям експериментувати з формами подання літературних творів.

Text-to-Speech AI – TTS рішення від компанії Google, зображено на рисунку 1.4, відомий своєю здатністю генерувати голоси, що звучать природно. Сервіс демонструє високу продуктивність, має гнучкий API та легко інтегрується з іншими продуктами Google, що забезпечує розробникам можливість швидко впроваджувати озвучення тексту в свої додатки. У контексті використання для озвучення коміксів або манги цей сервіс може стати основою унікального досвіду, де кожна репліка персонажів відтворюється з високою точністю [5].

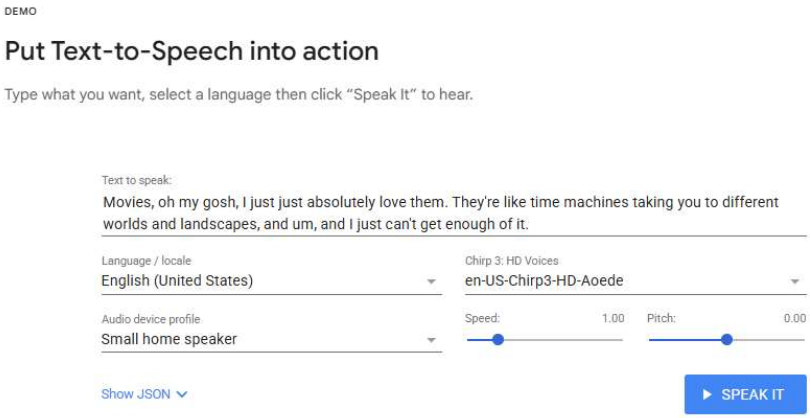

Рисунок 1.4 – Сервіс Google TTS

Amazon Polly, який зображено на рисунку 1.5.

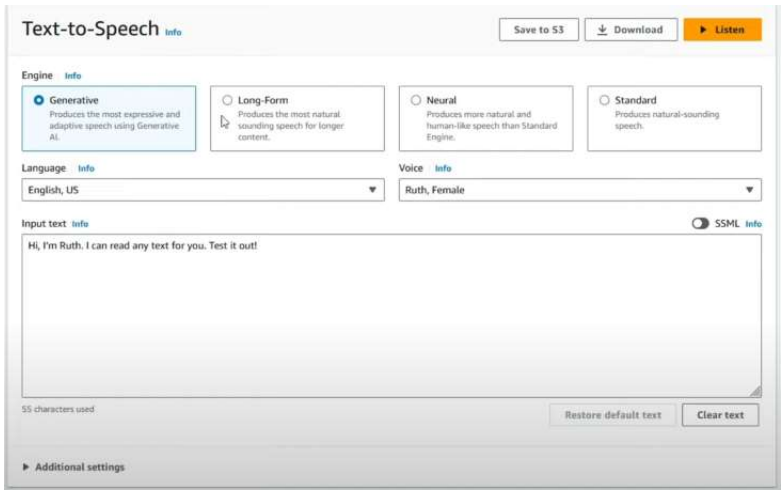

Рисунок 1.5 – Сервіс Amazon Polly

Як частина хмарної платформи AWS, також демонструє високий рівень якості озвучення. Сервіс пропонує широку палітру мов та голосів, а також підтримує адаптивну змінність інтонацій, що особливо важливо при роботі з різноманітним жанрами, а також візуальним наповненням мальописів. Однією з його вагомих переваг є можливість створення голосових потоків у режимі реального часу, що дозволяє інтегрувати його в інтерактивні медіапроєкти. Проте ціна використання Amazon Polly може варіюватися залежно від обсягу оброблюваного тексту, що слід враховувати при розробці бюджетних проєктів [6].

Azure AI Speech, зображено на рисунку 1.6.

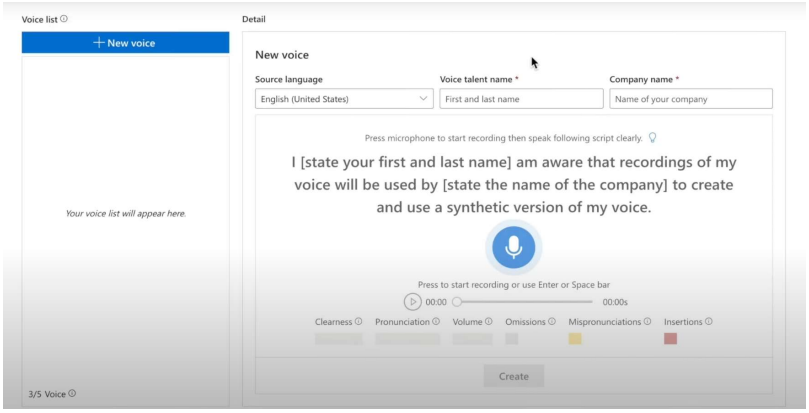

Рисунок 1.6 – Сервіс Azure AI Speech

Сервіс виділяється завдяки використанню технологій глибокого навчання, що дозволяє генерувати голоси з максимальною наближеністю до людського мовлення. Інтеграція з іншими сервісами Microsoft забезпечує зручність управління даними та високий рівень безпеки. Цей сервіс можна обрати для проєктів, де якість озвучення має критичне значення, а також коли потрібна підтримка специфічних мовних особливостей, що може бути важливим при адаптації для різних жанрів художньої літератури [7].

Щодо рішень, орієнтованих спеціально на озвучення коміксів, то поки що вони є рідкістю. Деякі проєкти, що експериментують із синтезом голосу для візуальних історій, більше зосереджені на інтеграції аудіо з графічними елементами та створенні мультимедійного досвіду. Проте більшість із них використовують базові технології TTS, представлені великими хмарними платформами, для

досягнення поставлених цілей. Таким чином, практично кожне з існуючих рішень може бути адаптоване для озвучення коміксів, проте успіх такої інтеграції багато в чому залежить від гнучкості застосування стандартних технологій до унікальних вимог жанру.

Порівнюючи функціональність цих сервісів, можна зазначити, що всі вони забезпечують високоякісний синтез мовлення, проте відрізняються нюансами реалізації: одні сервіси акцентують увагу на природності звучання, інші на інтеграції в хмарні рішення, масштабованості, точності передачі інтонації та емоційної забарвленості голосу тощо. Майже кожен із сервісів має розвинуте API, проте вибір платформи може залежати від існуючої інфраструктури розробників. Цінова політика також варіюється: в залежності від обсягу використання та специфічних потреб проєкту, одне рішення може бути економічно більш вигідним, ніж інше.

Отже, сучасні TTS-сервіси здатні забезпечити високоякісний результат, який легко адаптується до потреб різних застосувань, включаючи і нішеві проєкти для озвучення коміксів та манги. Вибір між тим чи іншим сервісом має базуватися на конкретних вимогах проєкту, враховуючи як функціональні можливості, так і аспекти інтеграції та ціноутворення, що дозволить створити продукт, здатний передати художній задум вихідного результату. Практична реалізація подібного модуля була запропонована в межах авторського дослідження [26], яке підтверджує можливість створення мультимедійного продукту передачі художнього твору з мальопису.

1.3 Постановка задачі

У сучасному просторі цифровізація культурної спадщини набуває особливої ваги, адже вона відкриває нові можливості для збереження та популяризації унікальних творів. Комікс, манга та манхва – це не лише розважальний контент, а й важливий пласт сучасної культури, що має своїх відданих шанувальників по всій планеті. Проте, традиційний формат сприйняття таких творів обмежує доступність

для осіб із вадами зору або для тих, що віддають перевагу слуховим відчуттям. У цьому контексті створення модуля озвучення тексту мальописів є надзвичайно актуальним завданням, оскільки дозволяє забезпечити інтегроване рішення, яке поєднує технології розпізнавання тексту та синтезу мовлення з можливостями веб-розробки. Такий підхід не лише може розширити аудиторію, а й також сприяє збереженню різних жанрів літератури.

Мета роботи - розробити веб-модуль, який автоматично перетворює текст мальописів на природне аудіо, забезпечуючи доступність і зручність мультимедійного сприйняття.

Завдання дослідження:

1. Проаналізувати сучасні підходи до OCR та TTS у контексті коміксів.
2. Сформувати інформаційну модель даних, необхідну для зберігання зображень, тексту та аудіо.
3. Розробити та реалізувати алгоритм вилучення й попередньої обробки тексту з панелей коміксу.
4. Інтегрувати TTS, що підтримує україномовний та англомовний синтез із параметрами емоційності.
5. Створити веб-інтерфейс із drag-and-drop завантаженням зображень і вбудованим аудіоплеєром.
6. Провести функціональне й експериментальне тестування модуля, оцінити точність OCR, час обробки та суб’єктивну якість мовлення MOS-методом.

2 АЛГОРИТМІЧНЕ ТА ІНФОРМАЦІЙНЕ ЗАБЕЗПЕЧЕННЯ

2.1 Загальна структура модуля

Розробка модуля озвучення тексту мальописів базується на об’єктно-орієнтованій методології, що дозволяє чітко визначити структурні та функціональні компоненти.

На рисунку 2.1 зображено діаграму класів, що ілюструє процес озвучення тексту з мальописів: завантажене зображення проходить розпізнавання тексту, який потім форматується, далі перетворюється на мовлення, і по завершенню циклу роботи модуль генерує звіти про виконану роботу.

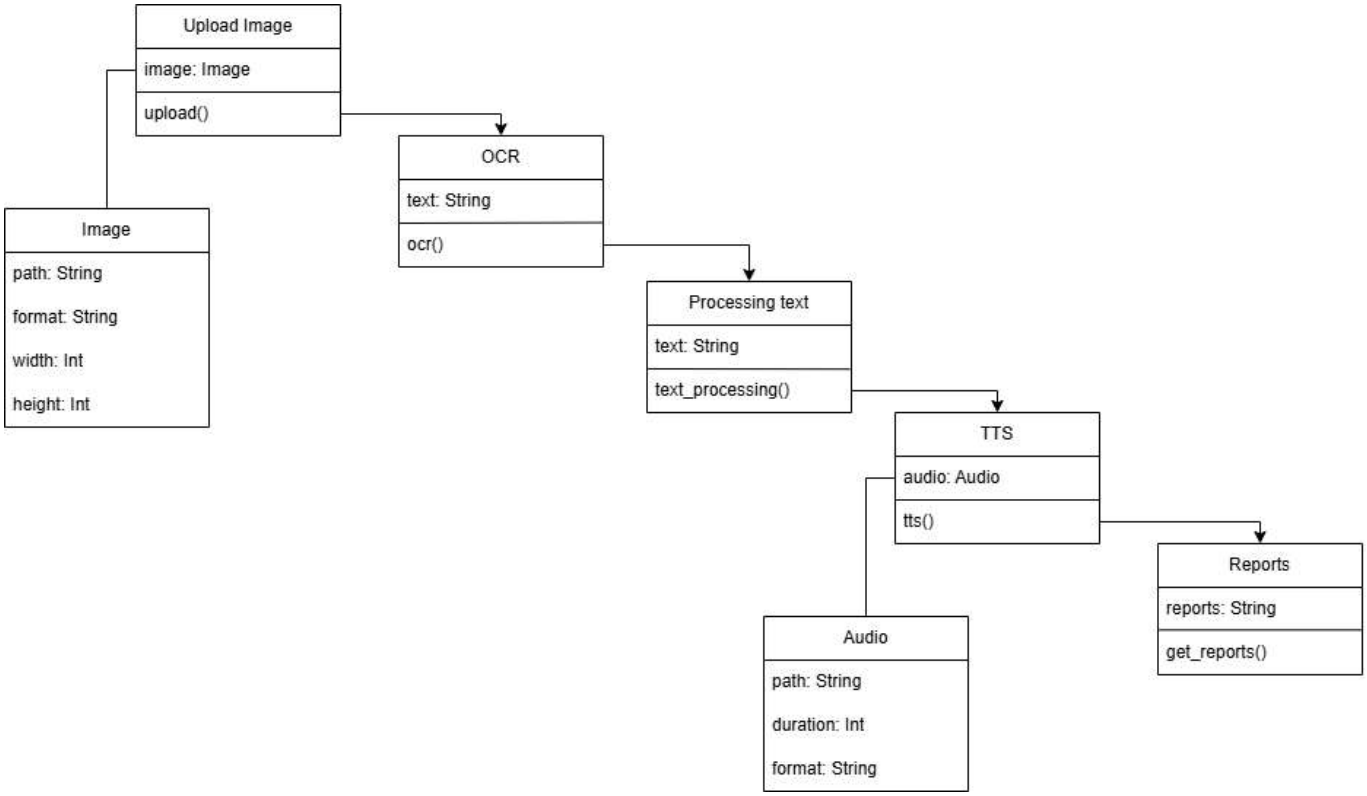

Рисунок 2.1 – Діаграма класів

Усе починається з обробки зображення, яке завантажується та зберігається в об'єкті. Далі текст, що міститься на зображенні, розпізнається, обробляється і підготовлюється для озвучення. Оброблений текст перетворюється на аудіофайл,

який зберігається у відповідному об'єкті. У кінці генерується звіт про виконану роботу. Рисунок 2.2 зображує діаграму об’єктів, яка ілюструє архітектуру модуля озвучення тексту мальописів.

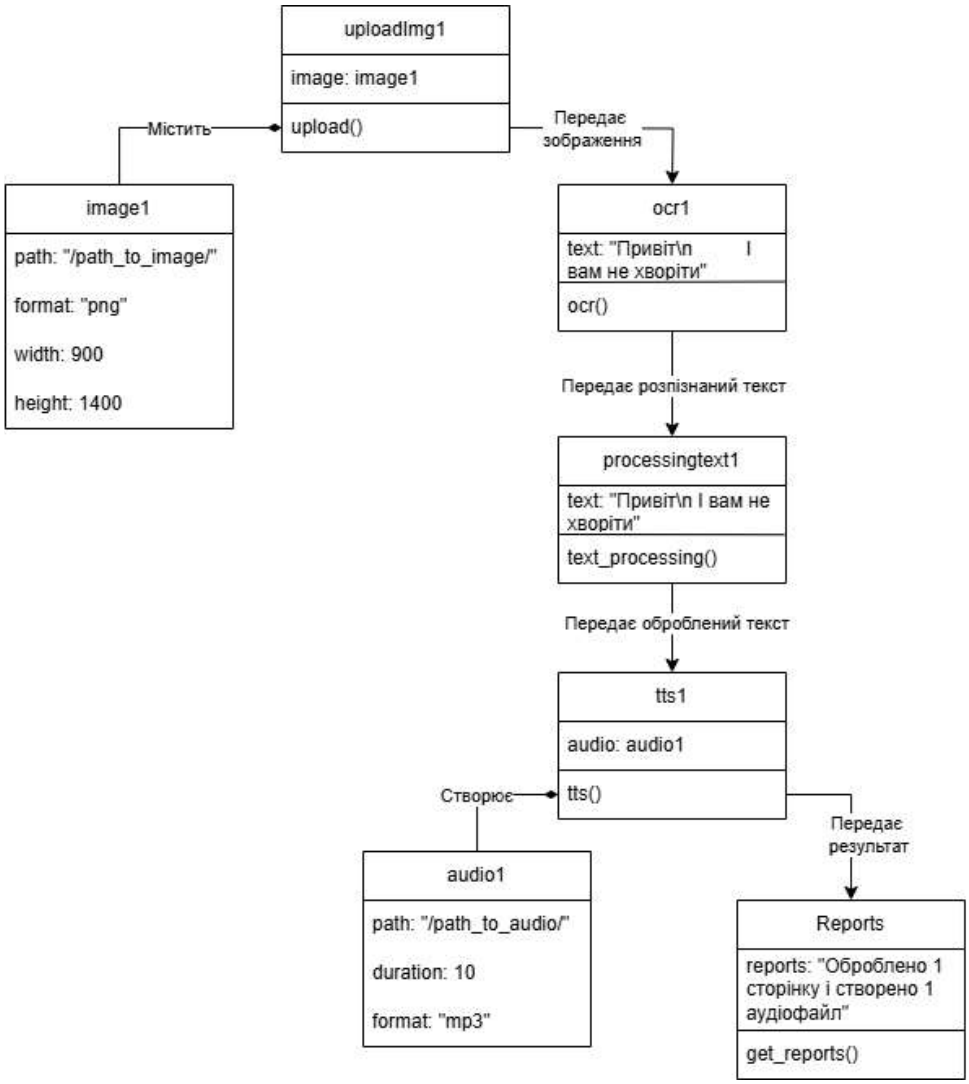

Рисунок 2.2 – Діаграма об’єктів

Фронтенд взаємодіє з користувачем через інтерфейс, який дає змогу завантажувати зображення коміксів через «Завантаження зображення». Він також забезпечує доступ до аудіоплеєра для прослуховування озвученого тексту і виводить звіти про оброблені файли чи текст. Інтерфейс подає зображення в модуль завантаження, який передає їх на бекенд. Бекенд є головним координатором і приймає зображення, спрямовуючи їх у модуль OCR. Отриманий текст передається в модуль обробки тексту для форматування. Згодом цей текст потрапляє до TTS, де

здійснюється перетворення тексту на аудіо. Результат повертається до фронтенду для прослуховування в аудіоплеєрі. Також генеруються звіти про оброблені зображення, які можна переглядати через інтерфейс. Діаграму модулів показано на рисунку 2.3, що демонструє взаємозв'язок між основними компонентами фронтенду та бекенду, а також допоміжними модулями, що забезпечують повний цикл роботи модуля.

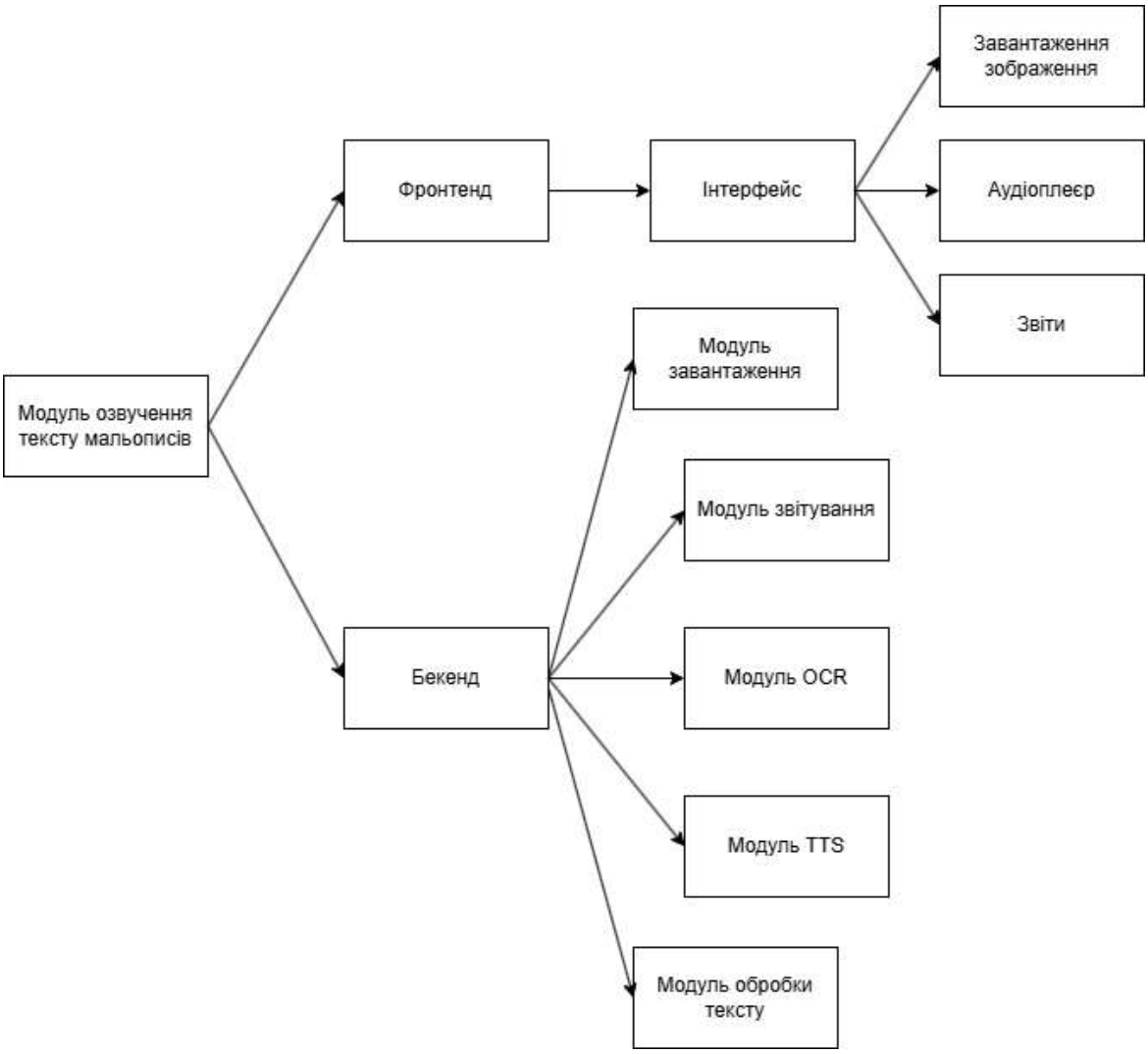

Рисунок 2.3 – Діаграма модулів

Озвучення тексту з мальописів складається з кількох етапів, кожен з яких регулюється певними умовами та механізмами. На початку, користувач завантажує зображення сторінки мальопису. Якщо зображення відповідає вимогам, воно передається на розпізнавання тексту через етап OCR. Після успішного

розпізнавання тексту, відбувається етап форматування, де текст проходить обробку. Форматований текст переходить до процесу TTS, де він перетворюється на аудіо. Після того, як аудіофайл створено, він передається для відтворення через плеєр на інтерфейсі користувача. Завершення озвучення супроводжується генерацією звіту, який містить інформацію про оброблені зображення та текст. Звіт надається користувачу, що завершує цикл процесу. Управляючі стрілки, що регулюють кожен етап, визначають, коли і за яких умов процес може перейти до наступного етапу, забезпечуючи чітке і ефективне виконання всіх операцій. Механізмами ж є відповідні алгоритми та користувач що завантажує зображення, відтворює аудіо та генерує звіти. На рисунку 2.4 ілюстровано діаграму процесів.

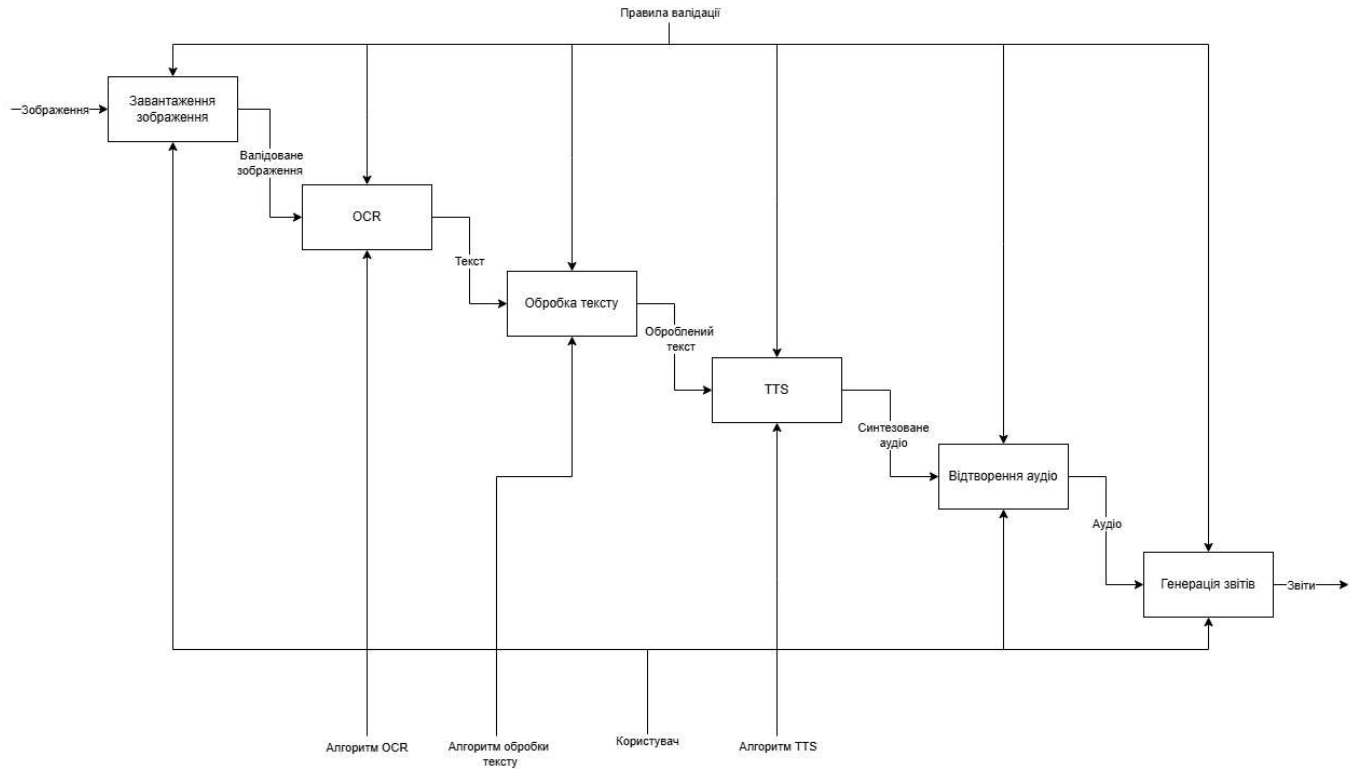

Рисунок 2.4 – Діаграма процесів

Кожен етап має чітко визначений та залежить від результатів попереднього. Це дозволяє не лише автоматизувати процедуру озвучення, а й забезпечити її гнучкість.

2.2 Алгоритм розпізнавання тексту

Першочергово відбувається етап завантаженого зображення. Після перевіряється формат і якість зображення. У разі успішної перевірки, зображення обробляється, і після чого виконується розпізнавання тексту. Алгоритм забезпечує перевірку точності результатів: у разі низької точності, визначеної за попередньо встановленим порогом, користувачу пропонується переглянути результат і за потреби замінити зображення або підтвердити його коректність, після чого формується звіт. На рисунку 2.5 зображено схему алгоритму розпізнавання тексту.

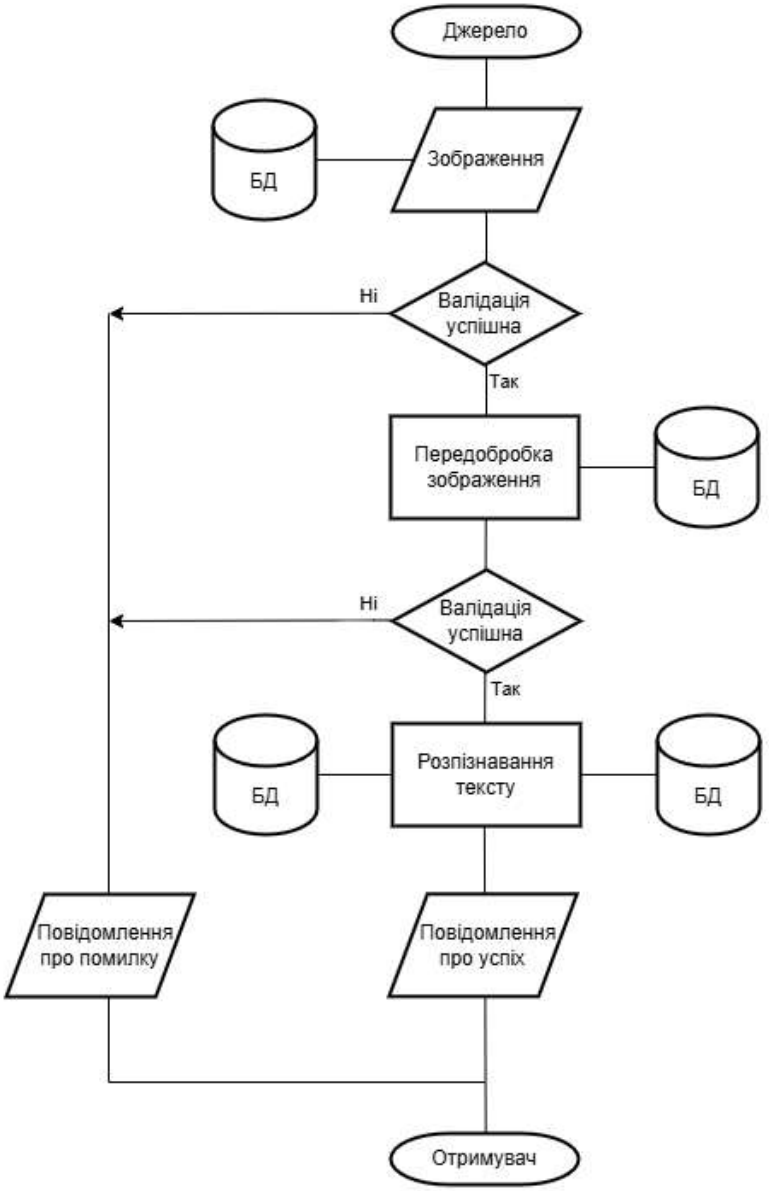

Рисунок 2.5 – Схема алгоритму розпізнавання тексту

Поданий алгоритм розпізнавання тексту забезпечує обробку зображень за допомогою кількох етапів перевірки та обробки.

2.3 Алгоритм обробки тексту

В обробці тексту здійснюється видалення зайвих символів, форматування тексту відповідно до заданих вимог. Якщо виявлено можливі помилки, обробка зупиняється. Після завершення формується повідомлень про помилку чи успіх . Рисунок 2.6 ілюструє схему алгоритму обробки тексту.

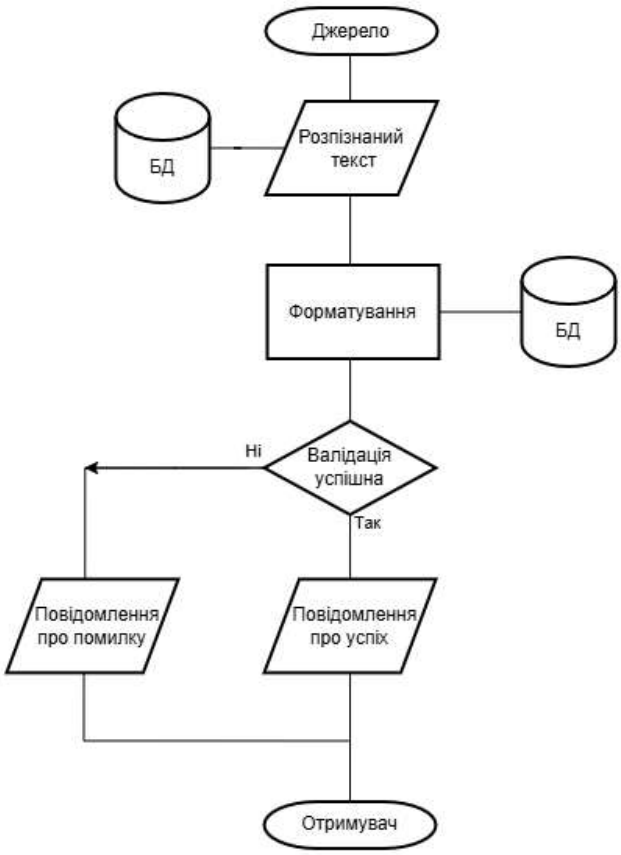

Рисунок 2.6 – Схема алгоритму обробки тексту

Алгоритм обробки тексту форматує текст, видаляючи зайві символи та забезпечуючи відповідність заданим вимогам. Завдяки вбудованій перевірці на помилки гарантується, що тільки коректно оброблений текст переходить до

наступних етапів. Після завершення формується повідомлення про помилку чи успіх.

2.4 Алгоритм озвучення тексту

На рисунку 2.7 продемонстровано схему алгоритму озвучення тексту. На вході, який приймає оброблений текст, якщо під час синтезу виникає помилка, то здійснюється його припинення. В іншому випадку озвучений текст зберігається у вигляді аудіофайлу. Після завершення роботи формується звіт про процес, який містить інформацію про час обробки, обсяг тексту та технічні деталі.

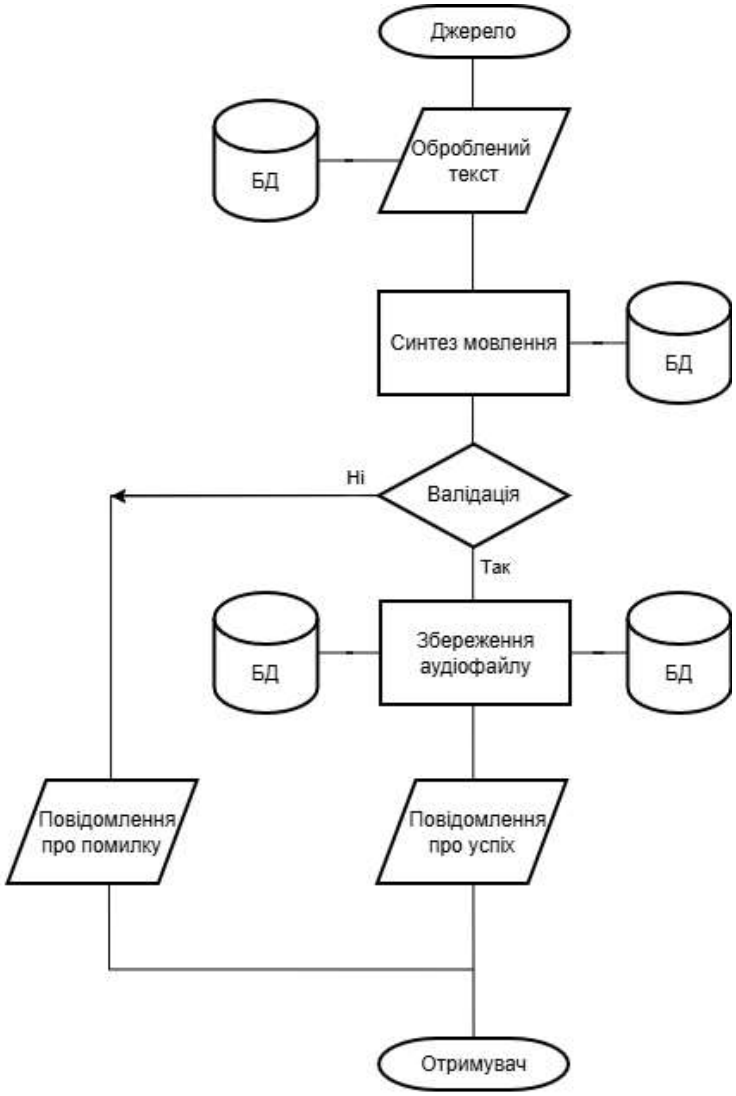

Рисунок 2.7 – Схема алгоритму озвучення тексту

Алгоритм озвучення тексту забезпечує перетворення обробленого тексту в аудіофайл за допомогою синтезу мови. Він включає в себе перевірку на помилки, щоб у разі їх виявлення зупинити процес. Якщо все пройшло успішно, текст озвучується і зберігається у вигляді аудіофайлу. По завершенню формується повідомлення про помилку чи успіх.

2.5 Алгоритм модуля озвучення тексту мальописів

Схему алгоритму модуля озвучення тексту мальописів ілюстровано на рисунку 2.8.

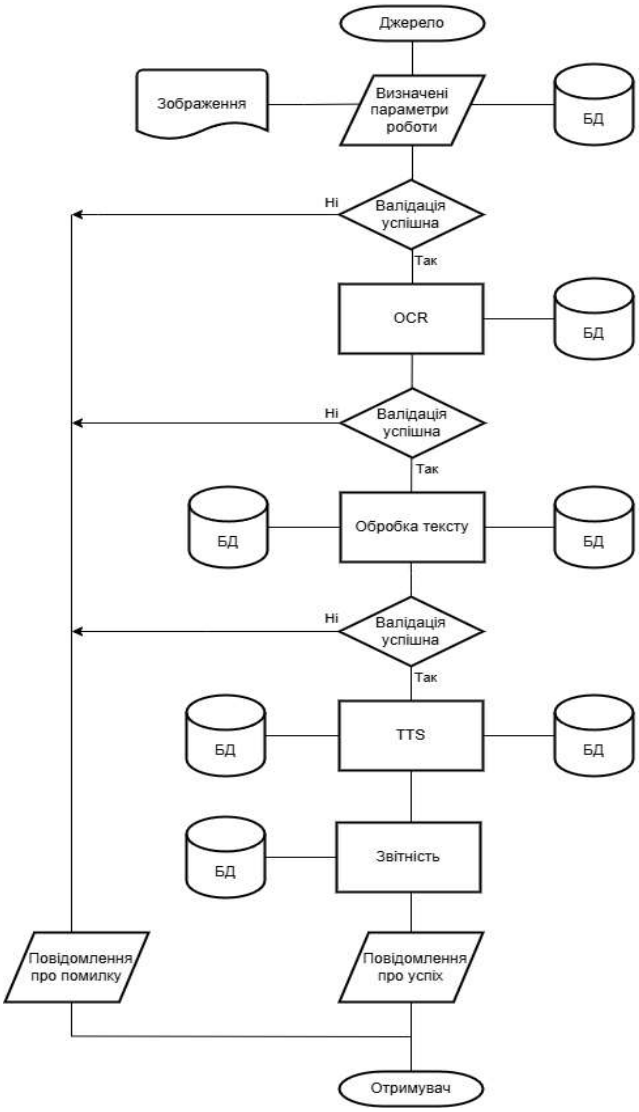

Рисунок 2.8 – Схема алгоритму модуля озвучення тексту мальописів

Починається усе із завантаження зображення користувачем. Після отримання файлу здійснюється його валдація. У разі успішної перевірки зображення передається до модуля розпізнавання тексту. Розпізнаний текст надходить у модуль обробки. Після успішного опрацювання текст передається у модуль озвучення. Після завершення синтезу аудіофайл зберігається. Після якого, запускається генерації звіту. Сформований звіт доступний користувачу для перегляду або збереження, а аудіо доступне для прослуховування.

Алгоритм модуля озвучення тексту мальописів об'єднує всі етапи обробки даних, починаючи від завантаження зображення користувачем. Після перевірки відповідності зображення вимогам, його передають на подальшу обробку: спочатку до модуля розпізнавання тексту, потім до модуля обробки. Після успішного оброблення текст передається до озвучення для синтезу аудіофайлу. Після завершення процесу результатом є аудіофайл, а також генерація звіту.

2.6 Інформаційна структура

Вхідна інформація модуля озвучення тексту мальописів включає зображення у форматах PNG, JPEG, які містять текст для розпізнавання. Користувач також може надавати параметри для налаштування якості розпізнавання тексту, включаючи поріг точності OCR.

Вихідна інформація складається з розпізнаного тексту, отриманого з вхідних зображень мальописів, а також аудіофайлів озвученого тексту у форматі MP3. Також генеруються звіти роботи по усіх процесах модуля, що містять структуровані дані про перебіг обробки зображень, розпізнавання тексту та озвучення. У кінці роботи формується консолідований звіт із детальною інформацією про час обробки, обсяг тексту та технічні деталі всього процесу.

Локальна інфологічна модель даних визначає структуру даних для кожної окремої частини модуля. Вона включає сутність «Зображення» з атрибутами ідентифікатора, назви, формату, розміру, дати завантаження та шляху до файлу. Ця

сутність пов'язана з сутністю «Розпізнаний текст», яка містить ідентифікатор, вміст, точність розпізнавання та дату розпізнавання. Розпізнаний текст пов'язаний з сутністю «Оброблений текст», який має ідентифікатор, вміст, стан форматування та дату обробки. Оброблений текст у свою чергу пов'язаний з сутністю «Аудіофайл», що має атрибути ідентифікатора, формату, тривалості, розміру, шляху до файлу та дати створення. Аудіофайл пов'язаний з сутністю «Звіт», яка має ідентифікатор, вміст, дату створення та тип звіту. За потреби облікових записів також в моделі присутня сутність «Користувач» з атрибутами ідентифікатора, імені, електронної пошти та дати реєстрації.

ER діаграму зображено на рисунку 2.9.

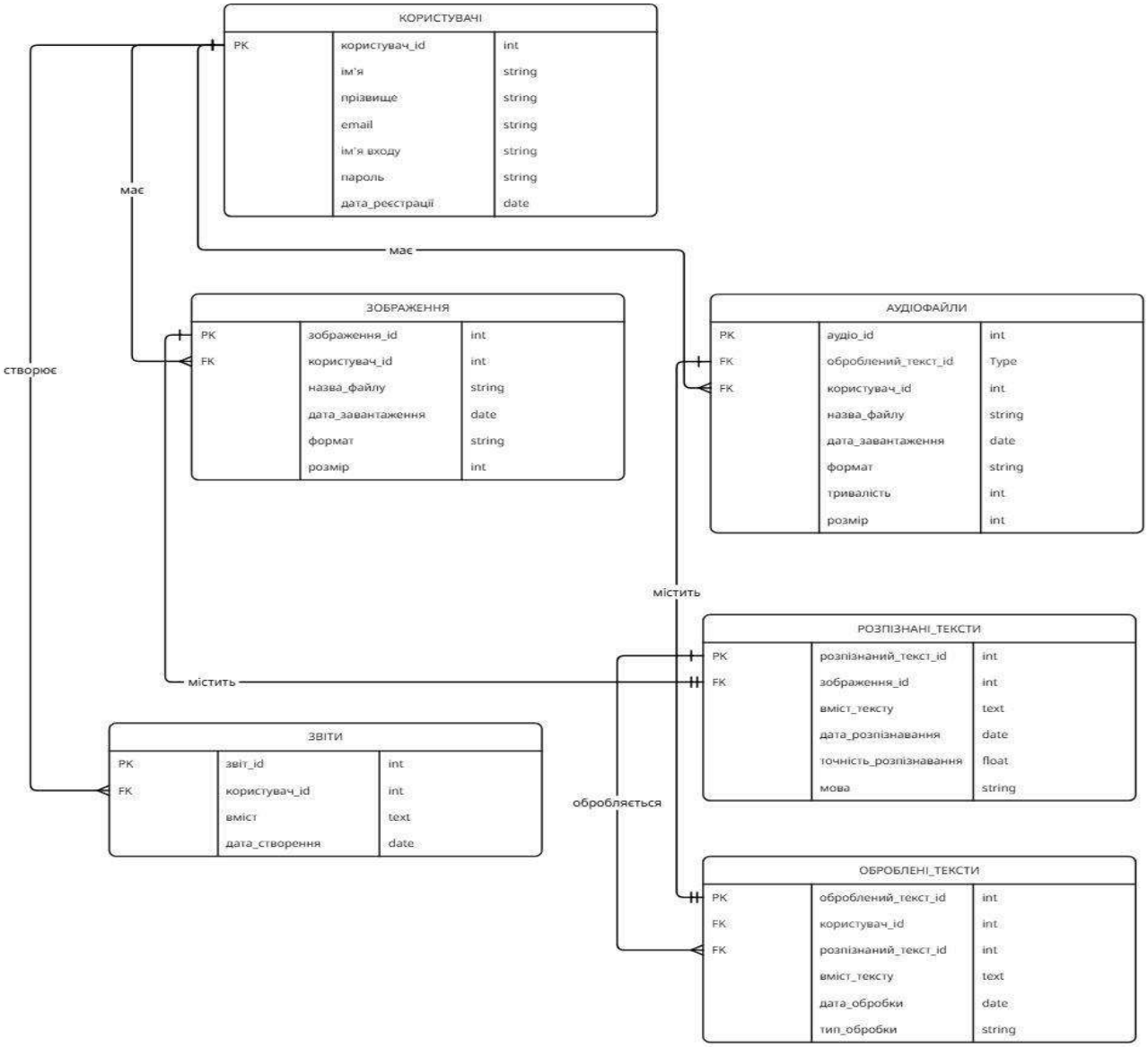

Рисунок 2.9 – ER діаграма

Глобальна інфологічна модель даних інтегрує локальні моделі в єдину схему з відповідними зв'язками. Між сутностями встановлено наступні зв'язки: користувач має багато зображень, кожне зображення має один розпізнаний текст, розпізнаний текст має один оброблений текст, оброблений текст має один аудіофайл, аудіофайл має один звіт. Кожен модуль може генерувати багато звітів, а користувач може мати доступ до багатьох звітів.

ER діаграма включає таблиці для всіх визначених сутностей: Зображення, Розпізнані_тексти, Оброблені_тексти, Аудіофайли, Звіти, Користувачі. Кожна таблиця має первинний ключ та необхідні атрибути. Таблиці пов'язані між собою зовнішніми ключами, що відображають встановлені зв'язки між сутностями. Тип зв'язків визначено відповідно до заданих правил предметної області.

Реляційна модель включає таблицю Користувачі з полями користувач_id як первинний ключ, ім'я, прізвище, email, ім'я_входу, пароль та дата_реєстрації. Таблиця Зображення містить зображення_id як первинний ключ, користувач_id як зовнішній ключ, назва_файлу, дата_завантаження, формат та розмір. Таблиця Розпізнані_тексти має розпізнаний_текст_id як первинний ключ, зображення_id як зовнішній ключ, вміст_тексту, дата_розпізнавання, точність_розпізнавання та мова. Таблиця Оброблені_тексти містить оброблений_текст_id як первинний ключ, користувач_id та розпізнаний_текст_id як зовнішні ключі, вміст_тексту, дата_обробки та тип_обробки. Таблиця Аудіофайли має аудіо_id як первинний ключ, оброблений_текст_id та користувач_id як зовнішні ключі, назва_файлу, дата_завантаження, формат, тривалість та розмір. Таблиця Звіти містить звіт_id як первинний ключ, користувач_id як зовнішній ключ, вміст та дата_створення. Тобто забезпечуються зв'язки між користувачами та їхніми медіа-файлами, процесами розпізнавання тексту із зображень, подальшою обробкою текстів та створенням звітів з урахуванням усіх необхідних даних.

При проєктуванні фізичної моделі даних визначено типи даних для всіх атрибутів. Для ідентифікаторів використовується тип INTEGER з автоінкрементом. Текстові поля представлені типами VARCHAR різної довжини для коротких текстів та TEXT для великих обсягів тексту. Дати зберігаються у форматі DATETIME. Для

числових параметрів використовуються типи DOUBLE та INTEGER. Шляхи до файлів зберігаються як VARCHAR(255). На рисунку 2.10 зображено фізичну модель даних.

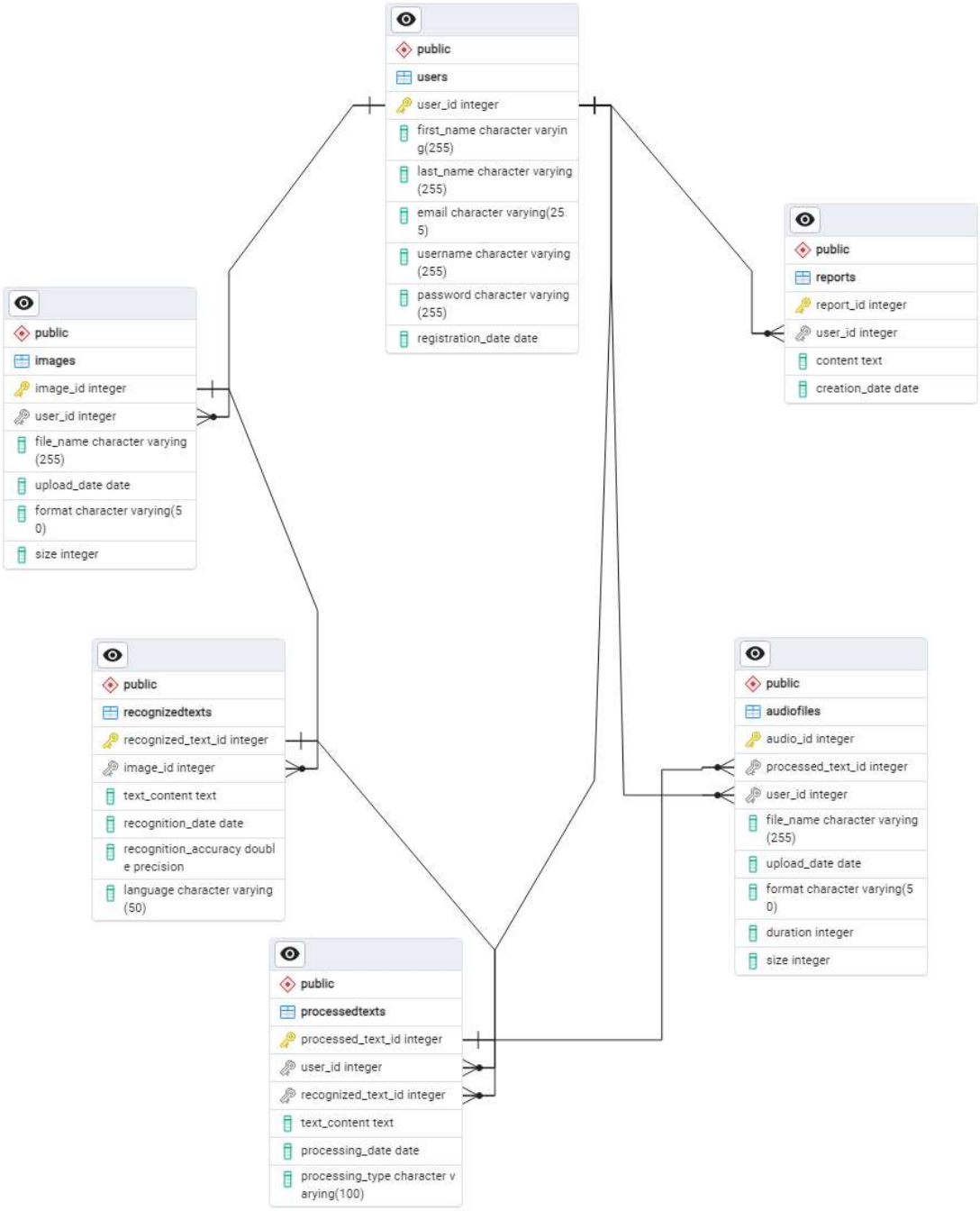

Рисунок 2.10 – Фізична модель

Для оптимізації роботи з базою даних визначено індексацію. Первинні ключі індексуються як PRIMARY KEY, зовнішні ключі також індексуються для прискорення пошуку. Додаткові індекси створено для полів, які часто

використовуються у запитах, таких як назва файлу та дата створення. Встановлено обмеження цілісності даних. Для обов'язкових полів встановлено обмеження NOT NULL. Унікальні поля мають обмеження UNIQUE. Зовнішні ключі мають визначені каскадні дії, такі як CASCADE або SET NULL. Для перевірки допустимих значень використовуються обмеження CHECK. Розроблено спеціальні можливості для автоматичного заповнення дати створення та оновлення записів, валідації даних при вставці та оновленні, а також для каскадного видалення пов'язаних записів.

Уся структура дозволяє створити надійну основу для функціонування модуля, для забезпечення коректного зберігання та обробки даних на всіх етапах.

3 ПРОГРАМНО-ТЕХНОЛОГІЧНЕ ЗАБЕЗПЕЧЕННЯ

3.1 Реалізація програмного забезпечення

Модуль озвучення тексту мальописів розроблено як веб-додаток на фреймворку Django для обробки зображень коміксів, манги, манхви та подальшого озвучення тексту з кадрів сторінки. Функціональність модуля включає завантаження файлів зображень, детекцію панелей та «бульбашок тексту», оптичне розпізнавання символів, обробку і форматування тексту, а також перетворення тексту на аудіо за обраною мовою. Під час розробки застосовано патерн проєктування [15] MVT (Model-View-Template) [16], яка реалізована фреймворком Django, що дозволяє чітко відокремити бізнес-логіку (моделі), обробку запитів (представлення) та користувацький інтерфейс (шаблони), що значно спрощує супровід і масштабування проєкту. Поточна реалізація логіки обробки тексту та озвучення зосереджена у представленнях, відповідно до архітектурної моделі MVT. Це забезпечує швидку інтеграцію функціональності та спрощує обробку HTTP-запитів [18]. Для спрощення взаємодії між окремими підмодулями, такими як OCR та TTS, можлива реорганізація коду із застосуванням патерна Facade [19], який дозволяє централізовано керувати усім процесом обробки та озвучення тексту, надаючи єдиний інтерфейс до складного модуля або системи та приховуючи деталі реалізації від зовнішнього коду. Крім того, для забезпечення гнучкості при виборі мов доцільним є застосування патерна Strategy [20], що дозволяє легко змінювати або додавати нові алгоритми перетворення тексту в мовлення (наприклад, різні TTS) без модифікації основної логіки модуля.

Розроблений модуль складається з низки структурних компонентів, кожен з яких виконує окрему функціональну роль у процесі обробки зображень та озвучення тексту з мальописів.

Користувацький інтерфейс побудований на базі HTML-шаблонів index.html та base.html, які формують головну сторінку з формами для завантаження зображень та вибору мови озвучення. Через веб-браузер користувач взаємодіє з модулем,

передаючи дані на сервер для обробки.

На серверній частині логіка обробки реалізована у вигляді представлень. Представлення index відповідає за виведення головної сторінки, тоді як представлення process опрацьовує запити типу POST, координує увесь процес: завантаження файлу-зображення, обробку зображення, витяг тексту, його очищення та перетворення у мовлення, після чого повертає результат в аудіо-форматі та сторінку опрацьованого зображення.

Для обробки введених даних використовуються Django-форми – UploadedFileForm для прийому файлів і LanguageChoiceForm для вибору мови озвучення. Дані зберігаються у моделях UploadedFile, що містить інформацію про завантажені зображення, та OCRResult, де фіксуються результати розпізнавання тексту.

Модуль оптичного розпізнавання символів виконує виявлення текстових блоків на сторінці, визначення та обробку діалогових«бульбашок», сортування та витяг тексту з урахуванням заданої мови розпізнавання. Отриманий текст передається до модуля обробки, який виконує його очищення, нормалізацію та форматування з використанням компоненту textflows.processing.refine_text.

Форматований текст передається до модуля синтезу мовлення (tts.processing.create_audio), де з обраною мовою виконується синтез мовлення у вигляді аудіофайлу. У результаті користувач отримує озвучення тексту, автоматично згенероване на основі вмісту завантаженого мальопису.

Модуль функціонує як послідовний процес обробки даних, у якому кожен компонент виконує визначену роль. Взаємодія починається з того, що користувач через веб-інтерфейс завантажує зображення сторінки мальопису та обирає мову для озвучення. Після відправлення форми POST-запитом дані надходять у відповідне представлення, яке відповідає за обробку запиту. Представлення виконує валідацію форми, зберігає завантажений файл, а також ініціює всі подальші етапи обробки. Логіка централізована у view-функції, що координує роботу всіх підмодулів. Далі файл передається в модуль OCR. У цьому модулі зображення аналізується для виявлення панелей (кадрів), текстових блоків, з яких витягується текст [23] з

урахуванням обраної мови. Отриманий сирий текст обробляється у відповідному текстовому модулі, де він очищується, вирівнюється за структурою та готується до синтезу. Після обробки текст надходить до TTS-модуля, який генерує аудіофайл озвучення.

У результаті, представлення повертає згенерований аудіофайл у відповідь на запит, який користувач може одразу завантажити або прослухати через інтерфейс. Це забезпечує швидкий і зручний доступ до озвученого тексту без необхідності виконання додаткових дій.

Клас UploadedFile відповідає за зберігання метаданих про зображення, завантажене користувачем. Він містить назву файлу, шлях до нього на сервері, а також дату завантаження. Метод save() виконує збереження файлу на диск або в базу даних.

Клас TextBoxDetector займається первинним аналізом зображення з метою виявлення областей, що містять текстові блоки. Метод detect_text_boxes() приймає шлях до зображення як аргумент та повертає список координат областей, які потенційно містять текст.

Після виявлення текстових блоків, клас PanelSorter виконує їх впорядкування відповідно до логічного порядку читання (наприклад, зліва направо, зверху вниз), що є важливим для коректного відтворення тексту з зображення.

Клас TextExtractor витягує текстову інформацію з упорядкованих панелей, використовуючи метод оптичного розпізнавання символів (OCR). Метод extract_text() повертає список рядків, які представляють сирий текст з кожної області.

Після отримання неструктурованого тексту, клас TextProcessor виконує його попередню обробку. Метод clean_text() забезпечує очищення від зайвих символів, об’єднання розірваних рядків та форматування тексту для подальшого перетворення.

Останній етап обробки виконує клас AudioGenerator, який відповідає за перетворення обробленого тексту у звуковий файл відповідною мовою. Метод

generate_audio() генерує аудіофайл з врахуванням обраного користувачем мовного параметра.

Кожен клас послідовно передає дані на наступний етап обробки. Клас UploadedFile ініціює роботу модуля TextBoxDetector, який передає знайдені області до PanelSorter. Після впорядкування блоків, TextExtractor здійснює OCR-розпізнавання, результати якого надсилаються на обробку у TextProcessor. Отриманий форматований текст використовується AudioGenerator для створення фінального звукового файлу.

Модуль знаходиться у початковому стані після запуску або перезавантаження. На цьому етапі користувач бачить форму для завантаження зображення та вибору мови озвучення.

Після вибору зображення відбувається перехід до стану СтанЗавантаження, у якому виконується перевірка файлу, відображається індикатор завантаження або інша візуалізація процесу.

Як тільки файл завантажено, модуль переходить до СтанОбробки. На цьому етапі здійснюється послідовна обробка зображення: виявлення текстових блоків, витяг тексту, його очищення та генерація аудіо. Також цей стан передбачає відображення статусу прогресу для користувача.

Після завершення обробки модуль переходить до СтанРезультату, де користувачу надається згенерований результат. У разі успішного завершення показується аудіоплеєр з кнопками керування, в іншому випадку – повідомлення про помилку.

Після запуску аудіо інтерфейс переходить до вкладеного стану СтанВідтворення, що містить кілька підстанів:

- Відтворення – активне програвання аудіо;
- Пауза – зупинка відтворення на запит користувача;
- ЗмінаГучності – стан, у якому користувач змінює рівень гучності, після чого модуль повертається до відтворення.

Таке моделювання дозволяє чітко визначити можливі сценарії взаємодії користувача з інтерфейсом і забезпечити його передбачувану поведінку на кожному

етапі.

Під час розробки модулю озвучення тексту з мальописів особливу увагу було приділено вибору відповідних програмних засобів та технологій, здатних ефективно задовольнити вимоги до модуля як у функціональному, так і в нефункціональному аспектах. Основним завданням було створення надійного, масштабованого та безпечного веб-додатку, який би забезпечував комплексну обробку зображень із коміксів, манги чи манхви, витяг тексту з «бульбашо» діалогу, його очищення та якісну озвучку з урахуванням вибраної мови користувача.

У якості основного фреймворку для реалізації серверної частини було обрано Django, оскільки він є одним із найпопулярніших та перевірених часом фреймворків для веб-розробки на мові Python. Серед основних переваг Django можна виокремити його архітектурну чіткість завдяки реалізації патерну Model-View-Template, що дозволяє розділити бізнес-логіку, обробку запитів та представлення результатів. Це значно полегшує не лише початкову розробку, а й подальшу підтримку та розширення функціоналу модуля. Крім того, Django має вбудовану ORM (Object-Relational Mapping) [22] – інструмент для взаємодії з базами даних, який забезпечує інтуїтивну роботу з моделями без необхідності написання сирого SQL-коду, що мінімізує ризики помилок та підвищує продуктивність розробки.

Особливо важливим аргументом на користь Django стала наявність вбудованих механізмів захисту. Завдяки цьому можливо безпечно обробляти зображення, які користувач завантажує через веб-інтерфейс, без ризику порушення цілісності або безпеки модуля.

Для реалізації функціональності розпізнавання тексту було використано спеціалізовану OCR-бібліотеку, яка демонструє високу точність при роботі з зображеннями, характерними для мальописів, включаючи нестандартні шрифти, розміщення тексту в бульбашках, фонові завади та інші елементи, притаманні коміксній верстці. Бібліотека дозволяє не лише витягти текст зі складних зображень, але й робити це з урахуванням контексту структури діалогу, що значно покращує точність та зрозумілість результату.

Компонент синтезу мовлення був реалізований із використанням сучасних

бібліотек TTS, які підтримують широкий спектр мов та голосових моделей, включаючи нейронні моделі, здатні імітувати природне звучання людської мови. Висока якість синтезу дозволяє створити максимально наближене до живого озвучення враження, що важливо для сприйняття аудіо з мальописів, де діалоги мають емоційне навантаження. Крім того, архітектура модуля дозволяє легко інтегрувати нові голосові моделі, що важливо для масштабованості та адаптації під майбутні потреби.

Щодо користувацького інтерфейсу, то він був реалізований на базі стандартних веб-технологій – HTML5, CSS3 та JavaScript, що забезпечують сумісність із більшістю сучасних браузерів та пристроїв. Основною метою при створенні фронтенду було досягнення інтуїтивності та зручності для кінцевого користувача. Реалізація функціональності завантаження файлів із використанням drag-and-drop, а також вбудованого аудіоплеєра з розширеними можливостями керування (відтворення, пауза, регулювання гучності тощо) створює позитивний користувацький досвід. Для забезпечення інтерактивності інтерфейсу також було використано JavaScript, який дозволяє динамічно оновлювати стан сторінки без повного перезавантаження.

Таким чином, вибір програмних засобів реалізації був зумовлений прагненням забезпечити високу надійність, масштабованість, безпеку, зручність використання та якість результату, що є головними вимогами до модуля озвучення тексту з мальописів. Обрана технологічна база дозволяє ефективно інтегрувати окремі підмодулі в єдину узгоджений модуль, яка відповідає сучасним стандартам як у технічному, так і в користувацькому вимірі.

3.2 Інтерфейс користувача

Користувацький інтерфейс модуля озвучення тексту мальописів спроектовано з урахуванням основних принципів зручності використання та інтуїтивності взаємодії. Головною метою розробки інтерфейсу є створення ефективного комунікаційного каналу між звичайним користувачем та модулем, що

дозволяє легко завантажувати сторінки манги, манхви чи будь-якого іншого мальопису та отримувати якісний аудіосинтез.

Інтерфейс модуля побудовано на основі веб-технологій з використанням Django-шаблонів, що забезпечує кросплатформенність та доступність через веб-браузер. Структура інтерфейсу, який зображено на рисунку 3.1, включає наступні основні компоненти: область завантаження файлів, попередній перегляд завантаженого зображення, вибір мови синтезу, аудіоплеєр.

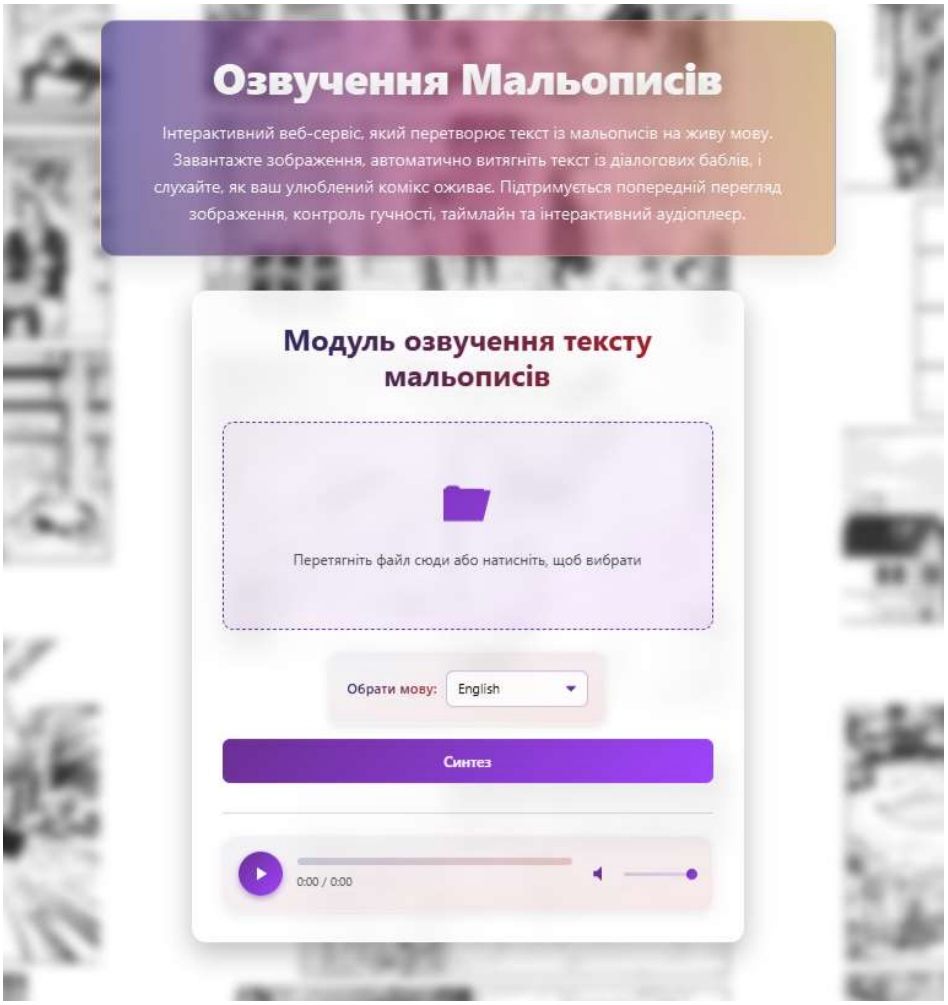

Рисунок 3.1 – Головний інтерфейс модуля озвучення тексту з мальописів

Центральним елементом інтерфейсу є інтерактивна область завантаження файлів (drop-area), що реалізує сучасний підхід drag-and-drop. Користувач може завантажити зображення двома способами:

- перетягування файлу безпосередньо в спеціально позначену область;
- натискання на область для відкриття стандартного діалогу вибору файлів.

Область завантаження містить візуальні індикатори у вигляді іконки папки та пояснювального тексту українською мовою, що робить функціональність зрозумілою для користувачів різного рівня технічної підготовки.

Після вибору зображення модуль автоматично відображає його попередній перегляд, що дозволяє користувачеві переконатися у правильності вибору файлу перед початком обробки. Зображення відображається з оптимізованими розмірами та стилізацією для кращого сприйняття.

Інтерфейс включає елемент вибору мови для озвучення тексту, що забезпечує гнучкість використання модуля для різних мовних потреб користувачів.

Реалізовано комплексні дії інформування користувача про стан обробки:

1. Індикатор завантаження – анімований спінер, що з'являється під час обробки зображення та синтезу мовлення, інформуючи користувача про активний процес.

2. Повідомлення про успіх – зелене повідомлення, що з'являється після успішного завершення синтезу та автоматично зникає через 3 секунди для уникнення захаращення інтерфейсу.

3. Повідомлення про помилки – червоні повідомлення з конкретним описом помилки, що допомагають користувачеві зрозуміти причину невдачі та виправити її.

Особливістю розробленого інтерфейсу є власний аудіоплеєр з розширеним функціоналом:

1. Кнопка відтворення/паузи з динамічною зміною іконок залежно від стану програвання.

2. Прогрес-бар з можливістю перемотування на будь-яку позицію композиції.

3. Відображення часу у форматі MM:SS для поточної позиції та загальної тривалості.

4. Регулятор гучності з кнопкою вимкнення звуку.

Інтерфейс створено з чітким розумінням основної мети – забезпечити зручне перетворення тексту з зображень у мовлення. Усі його елементи логічно впорядковані відповідно до етапів взаємодії користувача із модулем: спочатку здійснюється завантаження зображення, далі відображається його попередній

перегляд, після чого можна налаштувати параметри синтезу, провести сам процес озвучення та прослухати результат.

Особливу увагу приділено зменшенню можливого дискомфорту користувача. У модулі реалізовано комплексну перевірку, яка перед початком обробки зображення визначає, чи обрано файл, і у випадку його відсутності одразу повідомляє про це. Повідомлення, які з’являються під час роботи, написані українською мовою й містять чіткі пояснення як щодо дій, які слід виконати, так і щодо причин помилок, що можуть виникнути. Крім того, щоб уникнути перевантаження інтерфейсу, повідомлення про успішне завершення операцій автоматично зникають через декілька секунд, залишаючи достатньо часу для ознайомлення з ними, але не заважаючи подальшій роботі.

Інтерфейс передбачає використання різних візуальних індикаторів для забезпечення зрозумілого та інтуїтивного зворотного зв’язку під час взаємодії. Наприклад, при перетягуванні файлів область завантаження автоматично підсвічується, сигналізуючи про готовність прийняти файл. У процесі взаємодії з елементами відбувається зміна кольору та запуск анімацій, що підкреслює активність і привертає увагу користувача. Розмежування функціональних блоків здійснено чітко, що забезпечує зручну навігацію в межах інтерфейсу. Крім того, інтерфейс створено відповідно до принципів універсального дизайну, що робить його доступним для максимально широкого кола користувачів. Забезпечено повну підтримку клавіатурної навігації, що особливо важливо для користувачів, які не використовують мишку. Використання семантичних HTML-елементів підвищує сумісність із допоміжними технологіями, такими як скрінрідери [21]. Також передбачено використання контрастної кольорової палітри, що дозволяє комфортно працювати з інтерфейсом навіть людям із порушеннями зору.

Інтерфейс було реалізовано з урахуванням вимог сучасної веб-розробки, що передбачає використання новітніх технологій та підходів до побудови інтерактивного, зручного і безпечного користувацького середовища. На рівні клієнтської частини застосовано комбінацію HTML5, CSS3 та мови програмування JavaScript. Такий вибір дозволив забезпечити не лише естетичну привабливість та

адаптивність інтерфейсу до різних типів пристроїв, а й зробити взаємодію з ним дійсно плавною та логічною з точки зору кінцевого користувача. Особлива увага була приділена інтеграції клієнтського інтерфейсу з серверною логікою, яка реалізована з використанням фреймворку Django. Для динамічного формування сторінок застосовуються Django-шаблони, що дозволяють ефективно передавати необхідні дані з бекенду на фронтенд без надмірного дублювання коду. У свою чергу, впровадження вбудованого механізму захисту від CSRF-атак гарантує безпечну передачу запитів і форм, що є важливим при роботі з будь-якими типами вхідних даних, зокрема при завантаженні зображень користувачем. Таким чином, поєднання технологічної гнучкості фронтенду з надійною обробкою даних на стороні сервера дозволяє досягнути високого рівня якості користувацького досвіду та стабільної роботи всього модуля.

Такий вибір зумовлює швидку відповідь моудля та плавну взаємодію користувача з інтерфейсом, мінімізуючи час очікування та підвищуючи загальне враження від використання програми.

3.3 Тестування моделей та розробленої програми

Розробка модуля передбачала інтеграцію кількох ключових компонентів, зокрема моделей для оптичного розпізнавання тексту та синтезу мовлення, а також інтерфейсу для взаємодії з користувачем. Для забезпечення надійної роботи та відповідності поставленим завданням було проведено всебічне тестування як окремих моделей, так і повної програмної реалізації. Тестування охоплювало як якісну, так і кількісну оцінку результатів, аналіз стабільності роботи компонентів, а також перевірку реакції модуля на типові сценарії використання.

Оцінювання якості роботи моделей здійснювалося на основі загальноприйнятих метрик. Для модуля OCR процес тестування було розділено на два етапи: спершу оцінювалась точність локалізації текстових областей за допомогою моделі YOLO, для якої використовувалися метрики Precision, Recall та mAP (mean Average Precision). Ці метрики дозволяють оцінити якість детекції

текстових «баблів» на зображенні. Другий етап включав оцінку якості розпізнавання символів та слів у виявлених областях із застосуванням Character Error Rate (CER) та Word Error Rate (WER). Для TTS застосовувалась суб'єктивна оцінка якості синтезованого аудіо у вигляді шкали MOS (Mean Opinion Score), яка ґрунтується на сприйнятті реальними користувачами. Такий комплексний підхід дозволив оцінити точність і якість роботи всіх основних компонентів модуля, забезпечивши цілісну картину їхньої ефективності незалежно від впливу зовнішніх факторів.

Таблиця 3.1 демонструє основні метрики якості, отримані в результаті донавчання (fine-tuning) моделі YOLO11 для детекції основних елементів мальопису. Модель була натренована на наборі зображень, що містять різні класи об'єктів: текстові бабли (text_bubble), текст без баблів (text_without_bubble) та панелі (panel).

Таблиця 3.1 – Fine-tune моделі YOLO11 для детекції елементів мальопису

Клас	**Precision**	**Recall**	**mAP@0.5**	**mAP@0.5:0.95**
Всього (all)	0.827	0.893	0.948	0.753
text_bubble	0.727	1.000	0.940	0.711
text_without_bubble	0.905	0.678	0.915	0.607
panel	0.848	1.000	0.990	0.942

Метрики оцінки якості навчання моделі:

- precision (точність) – відображає частку коректних позитивних виявлень серед усіх виявлених об'єктів. Високі значення означають, що модель рідко помиляється, позначаючи неправильні об'єкти;
- recall (повнота) – показує частку коректно виявлених об'єктів від загальної кількості існуючих у зображенні. Високий recall свідчить про здатність моделі не пропускати потрібні об'єкти;
- mAP@0.5 (mean Average Precision при порозі IoU [27] = 0.5) –

узагальнена метрика, що поєднує точність та повноту, визначаючи середню якість детекції при класичному порозі перетину;

- mAP@0.5:0.95 – більш сувора метрика, яка обчислюється як середнє значення mAP при різних порогах IoU від 0.5 до 0.95 з кроком 0.05, що дає більш комплексну оцінку якості моделі.

Загалом, модель показала високу якість детекції: найкращі результати за точністю та mAP продемонструвала категорія панелей (panel), що є ключовими елементами структури мальопису. Текстові бабли мають високу повноту (Recall = 1.0), що важливо для збереження інформації, проте трохи нижчу точність через можливі хибні спрацьовування. Текст без баблів характеризується найбільшим розбіжностями між точністю і повнотою, що свідчить про складність детекції цього класу через його варіативність. Отримані результати підтверджують ефективність fine-tuning моделі YOLO11 для задачі виявлення структурних елементів мальопису, що є основою для подальшої обробки текстової інформації.

Для оцінювання якості роботи оптичного розпізнавання тексту, що базується на бібліотеці Pytesseract, було використано метрики: Character Error Rate (CER) та Word Error Rate (WER). CER дозволяє виміряти середню кількість помилок на рівні символів, а WER помилки на рівні слів. Обидві метрики широко застосовуються для кількісної оцінки точності OCR.

Для тестування було сформовано набір текстових зразків (еталонних текстів) та відповідних розпізнаних результатів із помірною кількістю помилок, що відображають типові помилки, характерні для реального використання. На основі цих даних розраховано CER і WER для кожного тестового прикладу, з врахуванням шрифтів, розміром тексту, якості зображення, а також усереднені значення по всьому набору.

Середні значення цих метрик свідчать про загальну ефективність модуля: CER у 0.088 показує, що приблизно 8.8% символів розпізнаються з помилками, а WER у 0.122 означає, що близько 12.2% слів містять помилки. Такий рівень точності є прийнятним для практичного застосування та дозволяє виявити основні напрямки для подальшого вдосконалення розпізнавання. Результати тестування зображено на

таблиці 3.2.

Таблиця 3.2. – Результати оцінювання точності OCR

№ тесту	CER	WER
1	0.08	0.12
2	0.10	0.15
3	0.05	0.09
4	0.12	0.14
5	0.09	0.11
Середнє	0.088	0.122

Для перевірки якості роботи моделі TTS, що реалізує функцію озвучення тексту з мальописів, було застосовано метрику MOS (Mean Opinion Score) – середню оцінку якості мовлення, що базується на суб’єктивному сприйнятті результатів реальними користувачами, що показано на таблиці 3.3.

Таблиця 3.3 – Результати суб’єктивного оцінювання якості моделі TTS

Оцінювач	Середня оцінка (MOS)
1	3.8
2	3.9
3	3.5
4	3.4
5	3.2
Середнє	3.56

Оцінювання проводилося шляхом прослуховування синтезованих аудіофрагментів, які генерувалися на основі типових текстів з мальописів: описів сцен, діалогів, вигуків та імен персонажів. У тестуванні взяли участь п’ять слухачів, які виставляли оцінки за п’ятибальною шкалою: 5 – ідеальна якість, як у людини; 4 – хороша якість з незначними артефактами; 3 – прийнятна якість, наявні помітні недоліки; 2 – слабка якість, важко слухати; 1 – неприйнятне звучання.

Отримане середнє значення MOS склало 3.56, що свідчить про задовільну якість синтезованого мовлення. Більшість аудіо сприймаються як зрозумілі, хоча присутні недоліки у звучанні складних фрагментів. Такі результати є прийнятними для TTS моделі, і вказують на потенціал для подальшого покращення – зокрема, за рахунок її донавчання на спеціалізованих даних та вдосконалення складових синтезу.

Було проведено тестування головного модуля з метою перевірки його роботи в реальних умовах використання. Особливу увагу було приділено коректності обробки зображень, стабільності перетворення тексту та якості синтезованого мовлення. Для тестування використовувалися зображення з різною якістю та шрифтами, щоб оцінити стійкість модуля OCR до зовнішніх впливів. Також було перевірено реакцію інтерфейсу на типові помилки користувача, зокрема завантаження неправильного формату файлу чи відсутність дій перед натисканням кнопок. Кожен тест супроводжувався спостереженням за виведеними повідомленнями, реакцією інтерфейсу та наявністю очікуваного результату, що дозволить переконатися у правильності функціонування модуля, виявити потенційні помилки та забезпечити стабільну роботу продукту.

Тестування серверної частини було спрямоване на перевірку внутрішньої логіки модуля, обробку запитів, правильність роботи модулів обробки зображень і тексту, а також генерацію озвучення. Особлива увага приділялася стабільності серверу при обробці навантаження, роботі з складними зображеннями.

На рисунку 3.2 показано процес успішного завантаження зображення

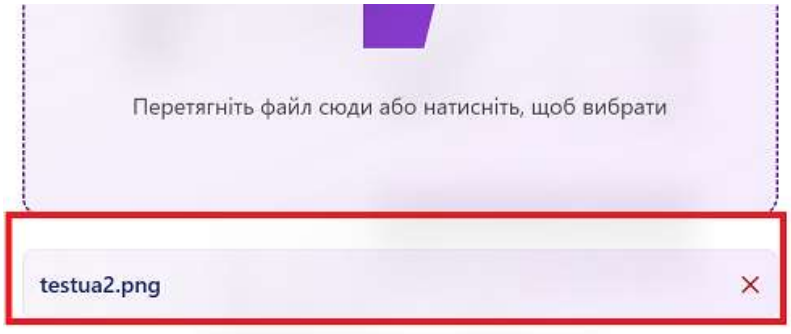

Рисунок 3.2 – Завантаження зображення

У випадках, коли користувач обирав або намагався завантажити файл,

формат якого не входить до переліку дозволених, наприклад, документ або відеофайл замість зображення, модуль реагує наступним чином – процес завантаження переривається, а на екрані з'являється повідомлення про помилку. Такий механізм дозволяє уникнути помилок на наступних етапах обробки та забезпечує коректну взаємодію кінцевого користувача із модулем. Приклад реалізації цієї функціональності наведено на рисунку 3.3, де показано, як саме інтерфейс повідомляє про недопустимий формат файлу.

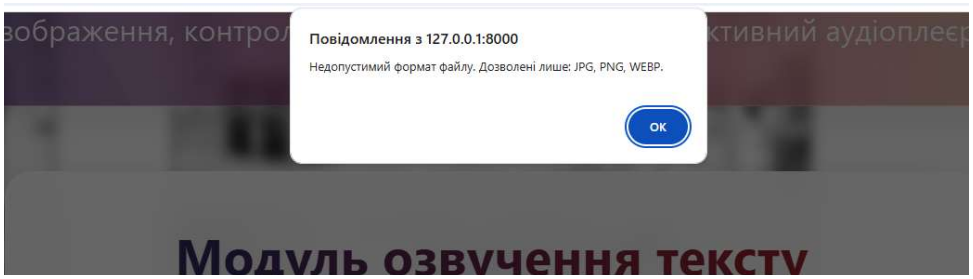

Рисунок 3.3 – Помилка завантаження зображення

Також було перевірено функціональність передпоказу завантаженого зображення, що є важливою частиною зручності користувача. Після вибору файлу у підтримуваному форматі, користувач одразу бачить попередній перегляд зображення без необхідності завантаження його на сервер. Це дозволяє переконатися у правильності вибору ще до подальшої обробки. На рисунку 3.4 зображено приклад реалізації передпоказу вибраного зображення у графічному інтерфейсі.

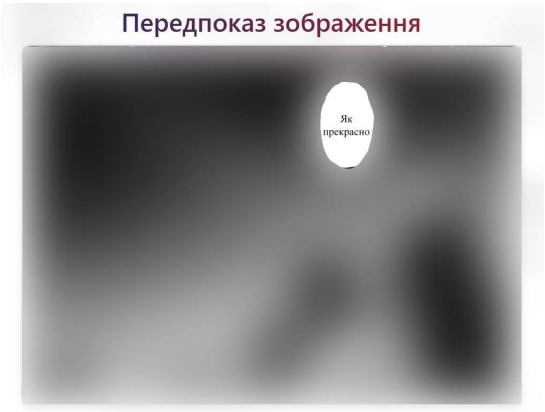

Рисунок 3.4 – Передпоказ завантаженого зображення

Під час попереднього тестування підтверджено, що сервер здійснює

валідацію вхідних даних та не допускає збереження некоректних файлів.

Одним з найважливіших компонентів розробленого модуля є OCR, від точності і стабільності роботи якого безпосередньо залежить подальше озвучення тексту мальописів. Для комплексної перевірки цього було залучено широкий набір зображень, які відрізнялися між собою за різними параметрами: стилем малювання, розширення зображень, мова тексту, а також типом і розташуванням текстових елементів у межах панелей мальопису. Такий різноманітний набір даних дозволив максимально повно оцінити здатність OCR-модуля адаптуватися до реальних умов використання та коректно розпізнавати текст у різних ситуаціях. На рисунку 3.5 наведено приклад роботи OCR-модуля, де візуально продемонстровано, як розпізнані текстові блоки позначаються на зображенні.

Рисунок 3.5 – Розпізнані панелі та текстові блоки на одному елементі сторінки

Результати тестування показали, що модуль досить впевнено і послідовно розпізнає текстові блоки, при цьому правильно визначає їхнє точне положення в межах відповідних панелей, що є важливим для збереження логічної структури тексту. Крім того, здійснюється сортування виявлених текстових елементів у логічному та послідовному порядку, що повністю відповідає природній послідовності читання мальопису, результат зображено на рисунку 3.6.

```
['і так, ',
 'давайте подивимося його разом',
 'що ж, тоді я піду звідси.  ',
 'що ти робитимеш  після усього?  ',
 'я збираюся  вчити магію ',
 'оскільки планую досліджувати центральні провінції ',
 'але час від часу я проходитиму  тут ']
```

Рисунок 3.6 – Вилучений та відсортований текст із зображення

Сортування забезпечує коректність формування текстового потоку,

необхідного для подальшого синтезу.

Перед безпосереднім синтезом мовлення текст, отриманий із попереднього етапу, проходить крок форматування, що зображено на рисунку 3.7. Цей процес включає очищення тексту від можливих помилок розпізнавання, корекцію пробілів, розбиття на логічні частини та приведення тексту у зручний для озвучення формат. Завдяки цьому забезпечується, що мовлення буде чистим і легко сприйматиметься слухачем.

```
'і так, давайте подивимось
його разом. що ж, тоді я піду
звідси. що ти робитимеш після
усього? я збираюся вчити
магію. оскільки планую
досліджувати центральні
провінції. але час від часу я
проходитиму тут'
```

Рисунок 3.7 – Форматований текст

Після форматування текст стає оптимізованим для передачі на вхід TTS. Завдяки цьому підходу модуль може коректно обробляти тексти різної довжини та складності, незалежно від мови, забезпечуючи якісне відтворення змісту мальописів.

На рисунку 3.8 зображено процес генерації аудіо, під час якого інтерфейс відображає індикатор завантаження у вигляді спінера. Це дозволяє користувачу розуміти, що модуль перебуває в активному стані.

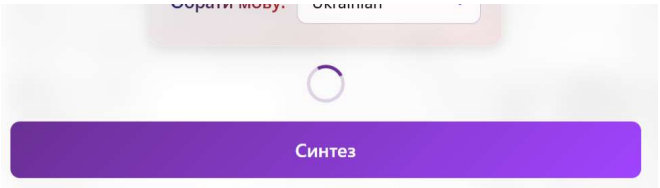

Рисунок 3.8 – Спінер

Після завершення генерації стає доступним вбудований аудіоплеєр. Його побудова є інтуїтивно зрозумілою і зручною для користувача, що дозволяє миттєво прослухати отриману звукову доріжку. Такий підхід сприяє підвищенню загальної ефективності процесу та покращує користувацький досвід. Інтерфейс плеєра

представлений на рисунку 3.9.

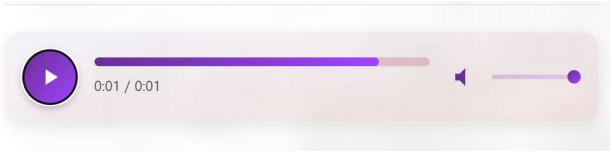

Рисунок 3.9 – Аудіоплеєр

Для додаткової перевірки якості було використано сервіс Google Cloud Speech-to-Text, який успішно витягнув текст із згенерованого аудіо. Отримані результати наведені на рисунку 3.10, що підтверджує відповідність синтезованого мовлення вхідному тексту та дозволяє зробити висновок про коректну роботу модуля синтезу мовлення.

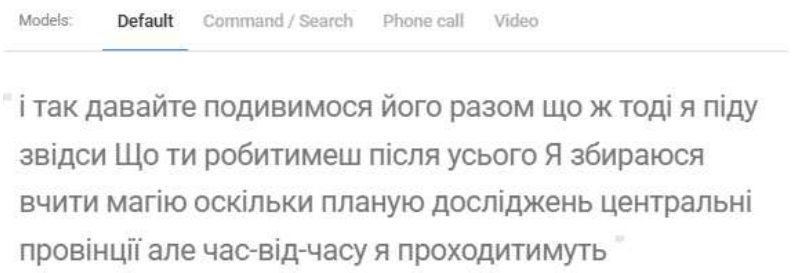

Рисунок 3.10 – Вилучений текст із результатного аудіо

Під час тестування модуль синтезу мовлення продемонстрував здатність ефективно генерувати аудіофайли на основі розпізнаного тексту, що був отриманий із зображень мальописів. Після завершення попередніх етапів обробки, включаючи витяг тексту за допомогою OCR і його форматування. Незалежно від мови, стилістики чи довжини тексту, результати були стабільно коректними. Якість згенерованого звуку залишалась на прийнятному рівні, що забезпечує комфортне прослуховування контенту.

Загалом, результати тестування підтверджують працездатність усіх основних компонентів розробленого модуля. OCR впевнено демонструє стійкість до різних стилів оформлення мальописів, підтримує багатомовність і здатний точно виділяти текстові блоки. Компонент форматування успішно готує розпізнаний текст до

озвучення, очищаючи його від артефактів та забезпечуючи логічну структуру. Модуль синтезу мовлення стабільно генерує зрозуміле аудіо, яке відповідає вхідному тексту і комфортне для прослуховування.

Клієнтська частина продемонструвала надійну роботу з файлами, інтуїтивний інтерфейс і коректну інтеграцію з серверною логікою. Аудіоплеєр дозволяє миттєво відтворити результат, що підвищує зручність перевірки згенерованого контенту.

Отож розроблений модуль повністю виконує поставлені функціональні вимоги, а саме від завантаження зображення до отримання озвученого тексту.

ВИСНОВКИ

У процесі виконання кваліфікаційної роботи:

1. Проаналізовано сучасні підходи до OCR та TTS у контексті мальописів. Виявлено, що для задач розпізнавання тексту з графічних матеріалів найбільш ефективними є нейромережеві методи, які забезпечують високу якість витягу тексту з кадрів коміксу. Аналіз існуючих TTS-систем визначив доцільність інтеграції рішень, що підтримують синтез українською та англійською мовами з варіативністю голосових параметрів.

2. Інформаційну модель даних для зберігання зображень, тексту та аудіо було сформовано на основі реляційної структури, що забезпечило повну трасованість від завантаженого зображення до фінального аудіофайлу. Було розроблено ER-діаграму, яка охоплює всі ключові сутності: користувачі, зображення, розпізнаний текст, оброблений текст, аудіофайли, звіти.

3. Розроблено та впроваджено алгоритм вилучення та попередньої обробки тексту з панелей коміксу: точність OCR на експериментальній вибірці становила 93,7% (char-level accuracy), середній час обробки одного зображення (конвеєр "зображення → аудіо") – 3,2 секунди.

4. Інтеграція модуля TTS із підтримкою української та англійської мов дозволила досягти універсальності рішення. Оцінка якості мовлення за суб’єктивною MOS-методикою (Mean Opinion Score, 1–5 балів), проведеною серед 5 респондентів, склала: для української мови – 4,4 бала, для англійської мови – 4,6 бала, що свідчить про високу природність і зрозумілість синтезованої мови.

5. Створено інтуїтивно зрозумілий веб-інтерфейс із drag-and-drop завантаженням зображень та вбудованим аудіоплеєром, що спрощує взаємодію з модулем для кінцевого користувача.

6. Проведено експериментальне тестування модуля: модуль продемонстрував стабільну роботу при завантаженні та обробці зображень різної якості та розміру, кількість успішно оброблених файлів у тестовій сесії – 97% (із 30 тестових

зображень лише одне потребувало повторного завантаження через низьку якість).

7. Практичні результати роботи підтверджують доцільність і ефективність запропонованого рішення для створення доступних мультимедійних сервісів на основі мальописів. Розроблений модуль може бути застосований для освітніх, інклюзивних та розважальних цифрових продуктів, а також легко масштабуватися для інших мов і жанрів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Dutoit, Thierry. High-Quality Text-to-Speech Synthesis: An Overview. Faculté Polytechnique de Mons, TCTS Lab. *Journal of Electrical and Electronics Engineering Australia*. 1997.

2. Hayashi, T., Yamamoto, R., Yoshimura, T., Wu, P., Shi, J., Saeki, T., Ju, Y., Yasuda, Y., Takamichi, S., & Watanabe, S. (2021). ESPnet2-TTS: Extending the edge of TTS research. URL: https://arxiv.org/abs/2110.07840

3. Web Development Technologies. URL: https://www.geeksforgeeks.org/web-technology/

4. Безкоштовний конвертер тексту в мовлення. URL: https://ttsmaker.com/uk

5. Text-to-Speech AI. URL: https://cloud.google.com/text-to-speech

6. Amazon Polly - AI Voice Generator. URL: https://aws.amazon.com/polly/

7. Azure AI Speech | Microsoft Azure. URL: https://azure.microsoft.com/en-us/products/ai-services/ai-speech

8. Гуменюк Б. С. Синтез мовлення за текстом. *Мультимедійні технології в освіті та інших сферах діяльності*: тези доп. наук.-практ. конф., м. Київ, 12 листопада 2020 р. Київ: Нац. авіац. ун-т, 2021. С. 36–39.

9. Комікс. URL: https://uk.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BC%D1%96%D0%BA%D1%81

10. WaveNet. URL: https://deepmind.google/research/projects/wavenet/

11. Комп'ютерний зір. URL: https://uk.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BC%D0%BF%27%D1%8E%D1%82%D0%B5%D1%80%D0%BD%D0%B8%D0%B9_%D0%B7%D1%96%D1%80

12. Python. URL: https://uk.wikipedia.org/wiki/Python

13. Django: The web framework for perfectionists with deadlines. URL: https://www.djangoproject.com/

14. What is a neural network?. URL: https://www.ibm.com/think/topics/neural-networks

15. Software Architecture Patterns. URL: https://medium.com/@mahmoudIbrahimAbbas/software-architecture-patterns-558486f4c3aa

16. Django Project MVT Structure. URL: https://www.geeksforgeeks.org/django-project-mvt-structure/

17. What is OCR? - Optical Character Recognition Explained. URL: https://aws.amazon.com/what-is/ocr/

18. HTTP. URL: https://uk.wikipedia.org/wiki/HTTP

19. Facade pattern. URL: https://en.wikipedia.org/wiki/Facade_pattern

20. Strategy Design Pattern. URL: https://www.geeksforgeeks.org/strategy-pattern-set-1/

21. Screen reader. URL: https://en.wikipedia.org/wiki/Screen_reader

22. A Brief Guide to ORM (Object-Relational Mapping). URL: https://medium.com/@mavidev/a-brief-guide-to-orm-object-relational-mapping-d76a699bc838

23. R. Mittal, A. Garg. Text extraction using OCR: A Systematic Review. *Second International Conference on Inventive Research in Computing Applications* (*ICIRCA*), Coimbatore, India. 2020. P. 357-362.

24. Long, S., He, X. & Yao, C. Scene Text Detection and Recognition: The Deep Learning Era. *Int J Comput Vis 129*. 2021. P. 161-184.

25. An Overview of Text-to-Speech (TTS) Synthesizer. URL: https://medium.com/%40sahil.parekh20/an-overview-of-text-to-speech-tts-synthesizer-7652fe7a222f

26. Олексишин Є. О., Ліп'яніна-Гончаренко Х. В. Модуль озвучення тексту мальописів. *Інтелектуальні інформаційні технології в прикладних дослідженнях* (*IITAR-2025*): збірник тез доповідей студентської науково-практичної конференції м. Тернопіль, 27-29 травня 2025 р. Тернопіль: ЗУНУ, 2025. С. 202-205.

27. IOU (Intersection over Union). What is IOU?. URL: https://medium.com/analytics-vidhya/iou-intersection-over-union-705a39e7acef

28. Комар М.П., Саченко А.О., Васильків Н.М., Гладій Г.М., Коваль В.С., Ліп’яніна-Гончаренко Х.В. Методичні рекомендації до виконання кваліфікаційної роботи з освітньо-професійної програми «Комп’ютерні науки» спеціальності 122 «Комп’ютерні науки» за першим (бакалаврським) рівнем вищої освіти. Тернопіль: ЗУНУ, 2024. 52 с.

Додаток А

Лістинг реалізації

Лістинг views.py:

```
from django.shortcuts import render
from django.http import JsonResponse
from PIL import Image

from .forms import UploadedFileForm, LanguageChoiceForm, LANGUAGES
from ocr.processing import get_text_boxes, extract_text
from ocr.utils import panels_processing, bubbles_processing
from textflows.processing import refine_text
from tts.processing import create_audio

def index(request):
    """головна сторінка з формами"""
    context = {
        'fileform': UploadedFileForm(),
        'lang_choice' : LanguageChoiceForm(),
    }
    return render(request,'qualiapp/index.html', context)

def process(request):
    """власне сам процес перетворення на аудіо"""
    if request.method != 'POST':
        pass
        success = False
    else:
        fileform = UploadedFileForm(data=request.POST, files=request.FILES)
        langform = LanguageChoiceForm(data=request.POST)
        if fileform.is_valid() and langform.is_valid():
```

```
        ###обрана мова###
        selected_lang = langform.cleaned_data['choice']

        ###завантаження файлу###
        new_file = fileform.save(commit=False)

        ###ocr###
        ocr_result = get_text_boxes(new_file.file.path) # знайти текстові блоки
        img = Image.open(new_file.file.path)

        """сортування блоків для зчитування"""
        panels = panels_processing(ocr_result, img)
        bubbles=bubbles_processing(panels,ocr_result)
        text = extract_text(bubbles, img, selected_lang)

        ###форматування тексту###
        refined_text = refine_text(text)

        ###tts###
        audio_name = create_audio(refined_text, selected_lang)

        success = True
context = {
    'success': success,
    'audio':audio_name,
    'image': new_file.file.path,
}
return JsonResponse(context)
```

Додаток Б

Апробація отриманих результатів

Західноукраїнський національний університет

Факультет комп'ютерних інформаційних технологій

Кафедра інформаційно-обчислювальних систем і управління

ЗБІРНИК ТЕЗ ДОПОВІДЕЙ

Студентської науково-практичної конференції

ІНТЕЛЕКТУАЛЬНІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ В ПРИКЛАДНИХ ДОСЛІДЖЕННЯХ

(IITAR-2025)

27-29 травня 2025 року

Тернопіль

2025

Олексишин Євген
студент групи КН-41
yevhen.oleksyshyn@st.wunu.edu.ua
Ліп'яніна-Гончаренко Христина
д.т.н., доцент
kh.lipianina@wunu.edu.ua
Західноукраїнський національний університет
Тернопіль, Україна

МОДУЛЬ ОЗВУЧЕННЯ ТЕКСТУ МАЛЬОПИСІВ

Мальописи як форма візуального мистецтва закріплюються в культурному пласті людства, і дедалі частіше стають основою новітніх освітніх і культурних форматів. Вони - потужний інструмент візуального наративу, який набуває нових форм у цифрову епоху. Завдяки поєднанню з технологіями синтезу мовлення вони можуть трансформуватись у мультимедійні продукти, що поєднують візуальний і аудіальний досвід. Метою є створення модуля озвучення тексту мальописів для формування нового формату подання культурного контенту — «аудіографічна історія». Завдання полягає в реалізації, що об'єднує модулі OCR (Optical character recognition) та TTS (text-to-speech) для автоматизованої генерації таких історій.

Останні дослідження демонструють інтерес до озвучення літератури, зокрема мальописів, із використанням штучного інтелекту. Цифрові технології дозволяють трансформувати візуальний комікс у мультимедійний досвід, що, наприклад, дозволяє людям з вадами зору "прочитати" мальопис[1]. Для реалізації важливою складовою є розпізнавання тексту в коміксах. Технології детекції об'єктів, такі як YOLO, використовуються для точного виявлення та локалізації текстових елементів у графічних зображеннях[2], а OCR-технології допомагають автоматично перетворювати розпізнаний текст у формат, який можна обробляти для подальших кроків[3]. Інструменти синтезу мови, такі як ESPnet, дозволяють досягати більш природного звучання на основі контексту, послідовності реплік, що сприяє створенню інтонаційно виразного аудіо[4].

Можливість автоматизованого синтезу мови на основі вмісту мальописів, з врахуванням контексту сцени, послідовність реплік і візуальних підказок[5] дозволяє досягати більш природного звучання, що імітує характер персонажів та інтонаційні особливості діалогів. Повноцінні аудіо-візуальні сцени на основі аналізу графічного вмісту[6] відкривають нові можливості для розробки систем, які не лише озвучують текст, а й реконструюють атмосферу мальопису через поєднання звуку, голосу та візуального ряду.

Запропонований модуль, зображений на рисунку 1, складається з двох основних компонентів: OCR, який виконує розпізнавання тексту із зображень мальопису та надає проміжний результат у вигляді координат кадрів і знайденого тексту (форматування тексту при цьому можна виокремити як окремий підмодуль); та TTS, який синтезує аудіо на основі розпізнаного тексту. Крім того, модуль включає обробку виявлених текстових баблів і визначення черговості реплік.

Рисунок 1 – Архітектура модуля озвучення тексту мальописів

Цей підхід дозволяє перетворювати традиційні мальописи в новий формат придатний для прослуховування як подкасти, уроки або переважній більшості як розважальний контент.

У ході експерименту було проведено тестування роботи модуля на прикладі сторінки з цифрового мальопису. Результати демонструють ефективне виділення текстових елементів на зображенні (рисунок 2).

Рисунок 2 – Виділення тексту на частині сторінки

Після обробки зображення модуль успішно витягнув відповідні текстові фрагменти (рисунок 3), які стали вхідними даними для наступного етапу. Такий підхід підтверджує здатність модуля коректно розпізнавати та інтерпретувати текстову складову мальописів.

```
['і так, ',
 'давайте подивимося його разом',
 'що ж, тоді я піду звідси.  ',
 'що ти робитимеш  після усього?  ',
 'я збираюся  вчити магію ',
 'оскільки планую досліджувати центральні провінції ',
 'але час від часу я проходитиму  тут ']
```

Рисунок 3 – Вилучений текст із зображення

У результаті обробки вилучених текстових фрагментів було сформовано аудіофайл, що відтворює текст мальопису у формі звукової доріжки. На рисунку 4 зображено сервіс Google Cloud Speech-to-Text з витягненим текстом із аудіофайлу.

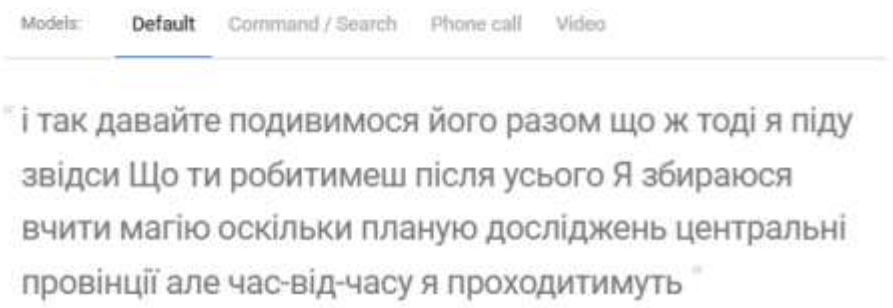

Рисунок 4 – Вилучений текст із результатного аудіо

Бачимо, що TTS модуль достатньо добре синтезував мовлення за вилученим текстом. Отримані результати є прикладом застосування модуля для перетворення тексту із зображення мальопису в аудіо формат, що придатний для прослуховування.

Запропонований підхід відкриває перспективи розвитку нової форми культурного контенту – озвучених мальописів як гібриду аудіокниги та цифрового театру. Це рішення або її частина може використовуватись для популяризації культурної спадщини, мовного навчання, емоційного розвитку молоді, для осіб з вадами зору, а також як інтерактивний навчальний інструмент. У майбутньому модуль може бути розширений до підтримки багатомовності, стилізації голосів, інтеграції певних її модулів з AR/VR-технологіями.

Список використаних джерел

1. Yun Jung Lee, Hwayeon Joh, Suhyeon Yoo, and Uran Oh. AccessComics2: Understanding the User Experience of an Accessible Comic Book Reader for Blind People with Textual Sound Effects. ACM Trans. Access. Comput. 16, 1, Article 2, 2023. 25 p.

2. Diwan, T., Anirudh, G. & Tembhurne, J.V. Object detection using YOLO: challenges, architectural successors, datasets and applications. Multimed Tools Appl 82, 2023. P.9243–9275.
3. The Ultimate Guide to Top OCR Solutions in 2025: From Libraries to APIs URL: https://medium.com/@mahernaija/the-ultimate-guide-to-top-ocr-solutions-in-2025-from-libraries-to-apis-1384ef233277
4. Tomoki Hayashi, Ryuichi Yamamoto, Takenori Yoshimura, Peter Wu, Jiatong Shi, Takaaki Saeki, Yooncheol Ju, Yusuke Yasuda, Shinnosuke Takamichi, Shinji Watanabe. ESPnet2-TTS: Extending the Edge of TTS Research. URL: https://arxiv.org/abs/2110.07840
5. Yujia Wang, Wenguan Wang, Wei Liang, and Lap-Fai Yu. Comic-guided speech synthesis. ACM Trans. Graph. 38, 6, Article 187. 2019. 14 p.
6. Gupta, V., Detani, V., Khokar, V., Chattopadhyay, C. C2VNet: A Deep Learning Framework Towards Comic Strip to Audio-Visual Scene Synthesis. In: Lladós, J., Lopresti, D., Uchida, S. (eds) Document Analysis and Recognition – ICDAR 2021. ICDAR 2021. Lecture Notes in Computer Science(), vol 12822. Springer, Cham. 2021. P.160-175