

**ДОСЛІДЖЕННЯ ТА РОЗРОБКА
СИСТЕМИ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ МОВЛЕННЯ
ТА ГЕНЕРАЦІЇ СУБТИТРІВ ДЛЯ ВЕБ-ПЛАТФОРМ**

Торосян А.А.

e-mail: arzman.torosian@nure.ua

Науковий керівник – к.т.н. Яцик М.В.

Харківський національний університет радіоелектроніки, каф. КМІТ
м. Харків, Україна

The paper presents the results of a study on the automation of subtitle generation for short-form video content using automatic speech recognition (ASR). The relevance of the problem is substantiated by the growing role of subtitled video in social media marketing and the limitations of existing manual approaches. An asynchronous microservice architecture for a web-based subtitle generation platform is proposed, enabling a complete processing pipeline from video upload to the production of a ready-to-publish file with embedded subtitles. The feasibility and efficiency of the proposed solution are confirmed through the development and experimental evaluation of a functional prototype.

Короткі відео стали домінуючим засобом просування контенту в соціальних мережах. Дослідження поведінки користувачів свідчать, що переважна більшість переглядів відеоконтенту відбувається без звуку, внаслідок чого субтитри перетворюються з допоміжного елемента на обов'язкову складову ефективної комунікації [1]. Водночас існуючі інструменти автоматичної генерації субтитрів мають суттєві обмеження: або потребують локального встановлення та технічних знань, або не забезпечують готового результату у форматі, придатному для безпосередньої публікації на соціальних платформах. Це формує нерозв'язану прикладну задачу – створення доступного веб-рішення повного циклу, орієнтованого на потреби SMM-спеціалістів.

Порівняльний аналіз моделей автоматичного розпізнавання мовлення (ASR) дозволив встановити, що найбільшу точність розпізнавання на різноманітних мовних даних демонструють моделі, побудовані на архітектурі трансформер та навчені на великих обсягах слабо анотованих даних [2]. Зокрема, модель Whisper від OpenAI, яка навчена на 680 000 годин мультимовних даних, виявляє стійкість до фонового шуму та не потребує додаткового налаштування під конкретну мову, що є критично важливим для платформи загального призначення. Для веб-орієнтованої системи найбільш доцільним визнано використання Whisper через зовнішній API (OpenAI Speech-to-Text API), що виключає необхідність розгортання власної ML-інфраструктури та суттєво знижує поріг входу для розгортання системи.

Задача генерації субтитрів для веб-платформи потребує асинхронної мікросервісної архітектури, оскільки операції транскрибування та обробки відео є ресурсомісткими та тривалими за часом. Ефективність подібного підходу до декомпозиції веб-платформи на спеціалізовані мікросервіси для обробки різномірних даних підтверджено, зокрема, у роботі [3]. Запропоноване рішення розділяє систему на три незалежні компоненти: клієнтський інтерфейс (Web App), серверний API (Web Service) та сервіс фонові обробки (Worker Service). Компоненти взаємодіють через чергу повідомлень, що забезпечує слабку зв'язність та дозволяє масштабувати сервіс обробки незалежно від інших частин системи. Така архітектура гарантує, що тривалі операції не блокують основний сервіс і не погіршують досвід користувача. Архітектурна діаграма зображена на рисунку 1.

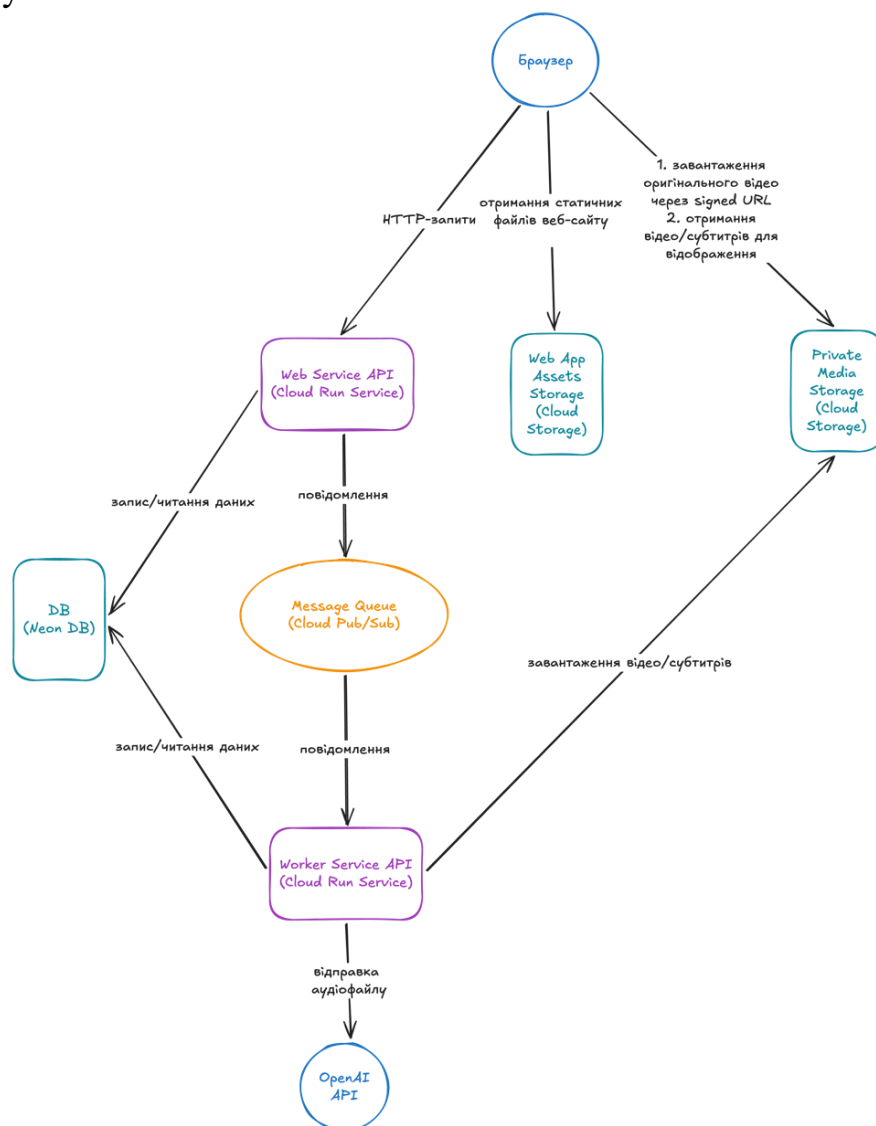


Рисунок 1 – Архітектурна діаграма системи

Також, доцільно використовувати двоетапний підхід до обробки відео, що зумовлено технічними обмеженнями цільових платформ.

На першому етапі система виконує транскрибування мовлення та генерує файл субтитрів у форматі WebVTT (.vtt), що дозволяє попередньо переглянути результат у браузері за допомогою HTML5-відеоплеєра. На другому етапі, ініційованому за запитом користувача, виконується вбудовування субтитрів безпосередньо у відеопотік, оскільки соціальні платформи не підтримують зовнішніх файлів субтитрів. Розділення на два етапи є принциповим, адже дозволяє користувачу оцінити якість розпізнавання перед генерацією фінального файлу та уникнути зайвого використання обчислювальних ресурсів.

Обрано хмарну модель розгортання у Google Cloud Platform з використанням контейнеризованих сервісів Cloud Run, конфігурація яких описана декларативно за допомогою Terraform (принцип «інфраструктура як код»). Це рішення обумовлене потребою у відтворюваності середовища, автоматизації розгортання через конвеєр CI/CD та можливості горизонтального масштабування при зростанні навантаження.

За результатами дослідження спроектовано та реалізовано функціональний прототип веб-платформи, який підтвердив працездатність запропонованого архітектурного рішення. Експериментальна перевірка засвідчила, що система успішно виконує повний цикл обробки – від завантаження вихідного відео до отримання файлу з вбудованими субтитрами – в автоматичному режимі без втручання користувача у процес обробки. Асинхронна архітектура забезпечила ефективне використання ресурсів та масштабування. Використання попередньо навченої моделі OpenAI Whisper через зовнішній API дозволило досягти прийнятної якості транскрибування без розгортання власної ML-інфраструктури, що значно спрощує супроводження системи. Перспективами подальшого розвитку визначено: інтеграцію засобів ручного редагування згенерованих субтитрів, розширення можливостей стилізації тексту та реалізацію прямої публікації контенту через програмні інтерфейси цільових платформ.

Список використаних джерел:

1. Patel S. 85 percent of Facebook video is watched without sound. Digiday. URL: <https://digiday.com/media/silent-world-facebook-video/> (дата звернення: 01.03.2026).
2. Robust speech recognition via large-scale weak supervision / A. Radford та ін. arXiv. URL: <https://arxiv.org/abs/2212.04356> (дата звернення: 01.03.2026).
3. Кузьменко Д., Гребеннік І., Урняєва І. Інформаційно-пошукова система генеалогічних досліджень. Цифрова платформа: інформаційні технології в соціокультурній сфері. 2025. Т. 8, № 1. С. 153–171.