Звіт про проєкт перекладу та клонування голосу у відео

Вступ

Ціль проєкту використати ШІ для перекладу та синтезу голосу у відеоконтенті, зокрема старих мультфільмів. Щоб зробити переклад на доступні мови, зберігаючи інтонацію та динаміку оригіналу.

На разі реалізований спрощений підхід, та переклад лише з української на англійську мову.

Процес роботи та використані технології

- **Виділення звуку з відео**
  за допомогою бібліотеки moviepy, витягується звукова доріжка з оригінального відео та зберігається як WAV-файл для подальшої обробки.
- **Розділення голосу та фонових звуків**
  використовується Spleeter (2-stem модель), яка ділить аудіо на два компоненти: "vocals" і "accompaniment". Це дозволяє зосередити подальше розпізнавання лише на мовленні, що суттєво покращує якість транскрипції.
- **Сегментація за спікерами**
  застосовується pyannote.audio (модель speaker-diarization-3.1), яка визначає, коли і який спікер говорить. Результат поділ голосової доріжки на сегменти за унікальними голосами.
- **Транскрипція**
  використовується Whisper (від OpenAI) для розпізнавання тексту з кожного сегменту. Модель здатна працювати з українською мовою та зберігати часові мітки.
- **Переклад**
  розпізнаний текст перекладається на англійську мову за допомогою GoogleTranslator
- **Синтез голосу**
  на основі перекладеного тексту створюється новий аудіофайл за допомогою TTS-моделі (Tacotron2-DDC або XTTSv2 для голосового клонування).
- **Клонування вокальної інтонації**
  застосовується Seed-VC, яка змінює синтезований голос, наближаючи його до тембру та інтонації оригінального мовця, використовуючи вихідні сегменти.
- **Склеювання звуку з фоновим звуком**
  сегменти озвучки синхронізуються відповідно до часових міток, і поверх фонового треку накладається новий озвучений контент.
- **Додавання нового аудіо до відео**
  за допомогою moviepy, нова звукова доріжка інтегрується у відеофайл, і фінальний результат експортується у форматі MP4

Результати

Візуальний результат: перекладене відео з синхронізованим, емоційно подібним озвученням.
Якість транскрипції не ідеальна приблизно 75% точності. Швидкість обробки орієнтовно~5 хв на 1 хв відео у Colab.

Виклики та рішення

Основним викликом було знайти ефективний спосіб передати інтонацію українського мовлення. Більшість доступних моделей клонування голосу не справлялися з цим завданням. Також складністю стало клонування голосів із мультфільмів, більшість відкритих моделей навчено на звичайному розмовному мовленні, і вони демонстрували низьку якість відтворення мультяшних інтонацій.

Рішенням стало поєднання двох підходів: перша модель (TTS) генерувала англомовний текст у вигляді звукового сигналу, а друга (Seed-VC) — накладала на нього інтонаційні характеристики з оригінального українського мовлення.

Ще одним викликом було налаштування середовища для запуску моделей. Оскільки я не маю власного обладнання з GPU, використовувався Google Colab. Проте він часто переривав сесію, а також виникали труднощі з сумісністю бібліотек (dependency hell).

Висновок

Система демонструє не погану якість результату на коротких і середніх відео. Переклад виконується з урахуванням контексту та зберігаються інтонації мовлення. Прототип показав, що об'єднання TTS і Voice Conversion цікаве та діюче рішення.

Майбутні можливості покращення

- Використати докер контейнер або об’єднати одну програму щоб не викачувати залежності при кожному запуску
- Заміна GoogleTranslator на DeepL API для вищої точності перекладу та кращої адаптації до контексту.
- Підтримка автоматичного визначення мови в аудіо.
- Підтримка більших мов для перекладу
- Покращення логіки синхронізації часу для більш точного звучання
- Інтеграція підтримки батч-перекладу для пришвидшення процесу обробки великої кількості сегментів.
- Підтримка тестування та статистичної оцінки
- Метрики автоматичної оцінки (BLEU, MOS з eval toolkits)