Voice Cloning — The Future of Personalized Speech AI

Заняття 13

Розробка моделей Generative AI

Лектор
Олександр Крижанівський

WHAT IS VOICE CLONING?

- Клонування голосу — це створення власної копії голосу з використанням нейронної мережі
- Цього досягають шляхом аналізу звукових зразків їхнього мовлення, фіксації таких шаблонів, як-от акцент і дихання, а потім застосування штучного інтелекту для створення синтетичного мовлення, яке імітує цільову особу. Після цього штучний інтелект може перетворити текст на мову, яка звучить надзвичайно схоже на клонований голос.
- Після клонування голосу для його виведення здебільшого задіюють процес синтезу тексту в мовлення
- Клонування голосу є буквально опцією деяких TTS-архітектур та систем

DIFFERENCE BETWEEN TTS AND VOICE CLONING AND VOICE CONVERSION

Основною відмінністю цих всіх технологій є те, що:

TTS	VOICE CLONING	VOICE CONVERSION
зазвичай використовує окремі натюнені моделі з різними голосами для синтезу аудіо або одну модель, яка містить декількох спікерів (але є основою для них)	дозволяє TTS-моделям відтворити голос, отримавши його невеликий шматочок, та буквально моментально відтворити його, перетворивши текст на аудіо	дає змогу натренувати модель без будь-якого тексту, застосовуючи тільки голос спікера, а під час інференсу прийняти вхідне аудіо з голосом A і змінити його на голос B, який був у тренувальному сеті. Немає ніякого залученого тексту — тільки вхідний голос перетворюється на вихідний.

DATASET REQUIREMENTS

Types of Required Data:

Для того щоб підготувати датасет для тренування TTS-моделі, яка в майбутньому зможе клонувати вхідні аудіо та відтворювати їх, потрібно дотримуватися декількох вимог:

- Чисті та високоякісні аудіозаписи з достатніми паузами на початку аудіо та в кінці (preferably 16kHz+ WAV)
- Відповідні транскрипції зі шляхами до аудіофайлів для того, щоб відтворити supervised learning
- Один спікер в датасеті або дуже багато спікерів, щоб узагальнити модель, навчатися нових голосів і пізніше за прикладом вхідного аудіозапису клонувати його вхідний голос на інференсі (datasets like LJspeech VCTK, LibriTTS)

РІЗНОВИДИ КЛОНУВАННЯ ГОЛОСУ

FOR VOICE CLONING:

- **Few-shot cloning:** це моделі, які здатні після 1–5 хвилин тренування на голосі спікера склонувати його
- **Zero-shot cloning:** використовує 1 приклад голосу від 5 до 30 секунд для повного клонування і виконує це моментально під час синтезу аудіо (або із вхідного тексту чи вхідного аудіо)
- **Many-shot cloning:** ці моделі потребують декількох годин аудіо для повноцінного клонування голосу та вивчення всіх ознак для якісного кінцевого синтезу

IMPORTANT QUALITIES:

- якісне середовище голосу (всі артефакти, звуки та шуми поза голосом також скопіюються в результаті)
- різноманітна вибірка слів, висловлень емоційних, забарвлень у голосі для найкращого та натурального результату під час синтезації

ЯК ПРАЦЮЄ КЛОНУВАННЯ ГОЛОСУ?

- Target Audio → Speaker Embedding
- Text → Acoustic Model (зумовлений ембедингами (внутрішні унікальні ознаки голосу)) → Mel
- Mel + Vocoder (Синтез мел-ознак в аудіоформу з клонованим голосом) → Cloned Speech

Основні елементи архітектур:

Speaker Encoder: learns fixed-size vector that captures speaker identity (вивчає збудовані ознаки голосу, щоб витягнути його ідентичність).

Acoustic Model: (e.g., Tacotron2, FastSpeech2, VITS) generates mel spectrogram (ця модель генерує мел-спектрограму, зумовлену клонованим голосом і вхідним текстом, який буде озвучено).

Vocoder: (e.g., HiFi-GAN, Vocos) converts mel to audio (конвертує мел-спектрограму, яка містить у собі всі цифрові визначення та ознаки голосу і заданого тексту, в живу мову, тобто в аудіозапис).

КОМЕРЦІЙНІ РІШЕННЯ ДЛЯ КЛОНУВАННЯ ГОЛОСУ

Commercial Voice Clonings

- Elevenlabs
- Respeacher
- PlayHT
- Google, Microsoft TTS
- Cartesia TTS

RETRIEVAL-BASED VOICE CONVERSION (RVC)

What is RVC?

- Перетворює промову одного мовця на іншу, використовуючи тональність і поділ вмісту
- На основі вилучення вмісту (зазвичай HuBERT) і вбудовування голосу
- Потокове перетворення голосу

How It Works

- Отримайте вміст (лінгвістичну інформацію) і презентацію з вихідного звуку
- Тренування / точне налаштування моделі для створення цільового голосу відповідно до цього вмісту
- Форма вихідного сигналу з клонованими голосовими характеристиками

Strengths

- Нема потреби вводити текст
- Також можна використовувати для співу

Оригінальний та клонований голос дівчини

Оригінальний та клонований голос чоловіка

ЯК НАВЧАЮТЬСЯ RVC-МОДЕЛІ?

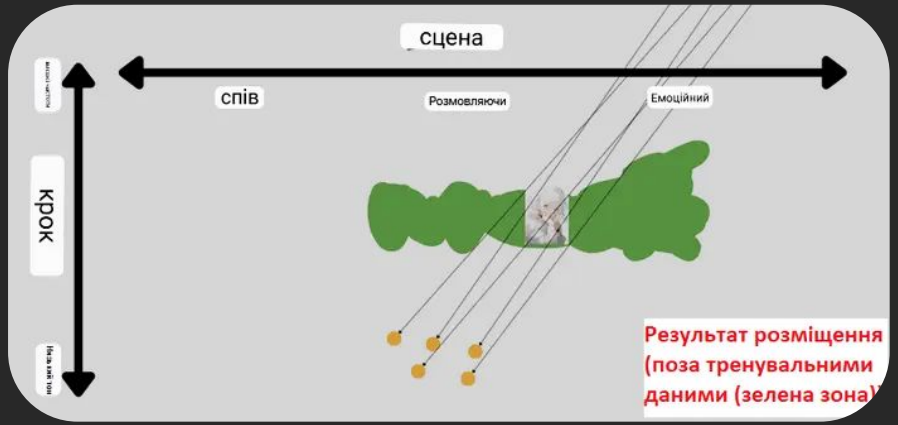
Всі дані, які ми передаємо для навчання, розташовуються в просторі. Простір, у якому вони розміщуються, називають «латентним простором». Нижче ви можете побачити, який вигляд він може мати в RVC. Тут ми припускаємо, що латентний простір RVC є двовимірним. На двовимірній осі вертикальна вісь — висота звуку, а горизонтальна — місце дії (читання, емоційне вираження, спів тощо).

ЯК ПРАЦЮЄ RVC

Коли в навчену модель RVC подається аудіо з голосом людини, який потрібно перетворити, то вона змінює мовлення на основі заданих функцій.

Спочатку ми аналізуємо голос, який вимовляє користувач. У цей час він розбивається на невеликі частини й обробляється, так само, як і під час навчання. Як приклад, ми аналізуємо фразу Hello та наносимо фонетичні особливості на латентний простір. Припустимо, що тут говорить людина з низьким голосом.

У цьому випадку відтворена позиція також буде розміщена в діапазоні низьких частот. Однак є одна проблема. Вхідний (переданий) звук може не існувати ні в одній ділянці латентного простору, отриманого під час навчання. Іншими словами, коли надається вхід із характеристиками мовлення, які модель не вивчила, важко конвертувати мовлення належним чином.



ПЕРЕСПІВУВАННЯ ГОЛОСУ В ПІСНЯХ

- Процес створення та як це відтворити
- Які дані використовували для навчання моделей голосами інших співаків
- Наявні приклади та порівняння

Артур Мізорян -> Кузьма

KOLA -> Христина Соловій

ВІД ЧОГО ЗАЛЕЖИТЬ ЯКІСТЬ МОДЕЛІ?

Думаю, багато хто вважає, що якість моделі залежить передусім від кількості вхідних даних. Насправді це не зовсім так, але я теж раніше так гадав. RVC розроблено на основі наявної моделі, тому для створення своєї відносно невеликої кількості високоякісного голосу з низьким рівнем шуму.

Тобто якщо є купа аудіо, наприклад, на 500 годин і в ньому є багато шуму й інших завад, якість моделі буде на незадовільному рівні + для навчання потрібно буде багато часу.

Рекомендовано створити невеликі відрізки аудіо (десь по 10 секунд) без сторонніх шумів та з різним емоційним спектром (звичайний голос, переляканий тощо). Поширена думка, що мінімальна кількість високоякісного аудіо для навчання — це приблизно 10 хвилин.

Підсумовуючи, важливо, щоб дані навчання для RVC були «гарними», а не «що більше, то краще». Іншими словами, середовище запису голосових даних і різноманітність вмісту, що читають уголос, сильно впливають на продуктивність моделі.

ОПЕНСОРС-РІШЕННЯ

TTS/VOICE CLONING:

- **Coqui TTS** (багатоспікерність та опційне клонування)
- **StyleTTS2 (моментальне клонування на базі малого вхідного прикладу)**
- **Tortoise TTS (моментальне клонування)**
- **Parler TTS (моментальне клонування)**

VOICE CONVERSION:

- **Retrieval-Based VC (RVC)**
- **Seed VC**
- **so-vits-svc**

MOS SURVEY AND COSINE DISTANCE

Для оцінювання успішності клонування чи конвертації голосу існують два методи порівняння:

- косинусна схожість (Pyannote та Titannet)
- людське опитування MOS (Mean Opinion Score)

Автоматичні алгоритми косинусної дистанції оцінюють аудіо за допомогою векторного представлення аудіозаписів.

$$\text{Косинусна схожість} = \frac{x \cdot y}{||x||y|}$$

В той час як учасники проходять опитування, оцінивши схожість клонованого голосу до оригіналу за шкалою від 1 до 5 балів. Це дає змогу порівняти об’єктивні результати алгоритмів з оцінками людей та визначити досконаліший спосіб оцінювання та порівняння.

CHALLENGES & LIMITATIONS

- Якість і кількість даних (zero shot example as input speech prompt)
- Вловлення емоцій мовця, просодії та інтонації
- Акценти й багатомовна складність
- Етичні питання: неправильне використання deepfake, згода, безпека

ETHICS & REGULATION

- Важливість згоди користувача та водяних знаків
- Можливість неправомірного застосування в шахрайстві, дезінформації тощо
- Нові закони та інструменти виявлення аудіофейків

ДЗ

Реалізуйте просте клонування голосу за допомогою доступного інструменту або API

- За допомогою open-source інструментів для клонування голосу створіть скрипт, який клонує голос. Застосуйте короткий аудіосемпл свого або іншого голосу та озвучте параграф тексту, перетворивши його на живу мову, та збережіть у текстовий файл.
- Опційно можна використати платні TTS API.

https://github.com/search?q=voice%20cloning&type=repositories

Q&A

???

r_d

ЗАВЖДИ Є КУДИ ЗРОСТАТИ