Enhancing Voice Authenticity in Speech-to-Speech Translation System

任務

本系統旨在開發一套「保有使用者聲色的語音翻譯系統」，此系統具備兩項核心功能。首先，將簡體中文文本轉換為單一角色的人聲語音（T2ST），確保語音輸出自然流暢且富有表現力。其次，實現從英語語音到同一角色人聲的轉換（S2ST），以準確重現原始語音信息。

任務	輸入	輸出
T2ST	輸入簡體中文文字	單一角色人聲語音輸出
S2ST	錄製一段英文語音	單一角色人聲語音輸出

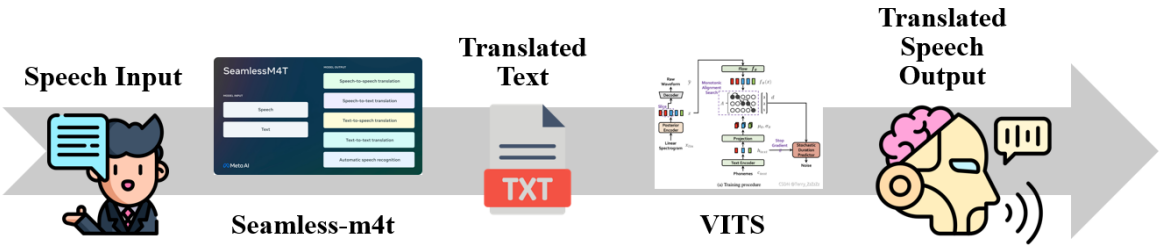
資料集

在本研究項目中，我們採用的數據完全源於自行蒐集的資料集。此過程涵蓋了對錄製者的精心挑選以及對錄製內容的嚴格控制，以確保數據的真實性和多樣性。詳細的錄製者資訊和錄製內容將在後續的表格中提供，以便於更深入地理解數據背景和使用情境。

錄製者	錄製內容	錄製語言	資料數量	資料格式	音檔長
Ivan (https://github.com/ivanc2k3)	笑話大全 幹話大全	中文	106筆	mp3	5~12秒
Neodoggy (https://github.com/neodoggy)	名言佳句 笑話全集 網路新聞	中文	121筆	mp3	5~10秒

系統架構

本系統採用了先進的模型架構，整合了SeamlessM4T（S2TT）和VITS（T2S）兩大核心技術。此外，我們還運用了VITS-fast-fine-tuning發布的介面，以提供使用者直觀的介面去使用。



SeamlessM4T

SeamlessM4T在2023年八月由Meta釋出，它在自動語音識別、語音到文本、語音到語音、文本到語音和文本到文本等多項任務上取得了最先進的成果。

參考資料	連結
Seamless Comunication Github	link (https://github.com/facebookresearch/seamless_communication)
SeamlessM4T Meta Blog	link (https://ai.meta.com/blog/seamless-m4t)
Seamless Comunication Github	link (https://github.com/facebookresearch/seamless_communication)
SeamlessM4T Demo	link (https://huggingface.co/spaces/facebook/seamless_m4t)
SeamlessM4T—Massively Multilingual & Multimodal Machine Translation	link (https://ai.meta.com/research/publications/seamlessm4t-massively-multilingual-multimodal-machine-translation/)
seamless-m4t-v2-large	link (https://huggingface.co/facebook/seamless-m4t-v2-large)
hf-seamless-m4t-large	link (https://huggingface.co/facebook/hf-seamless-m4t-large)
hf-seamless-m4t-medium	link (https://huggingface.co/facebook/hf-seamless-m4t-medium)

VITS-fast-fine-tuning 簡介

VITS可以在現有的VITS TTS模型中添加新的角色聲音，或是自己的聲音，使得模型能夠進行不同角色間聲音轉換，其中包括中文、英文和日文。

環境

Python Version: Python 3.11.6

```
pip install requirements.txt
```

VITS模型訓練

1. 訓練模型

- VITS-FAST_FINETUNE：可以直接在Colab運行此程式碼 (https://colab.research.google.com/drive/1pn1xnFfdLK63gVXDwV4zCXfVeo8c-I-0?usp=sharing)
- 使用教學：可以參考Youtube的教學影片 (https://www.youtube.com/watch?v=riYOD_EFKDE)

2. 將下列路徑中的*G_latest.pth*和*finetune_speaker.json*儲存到本基端

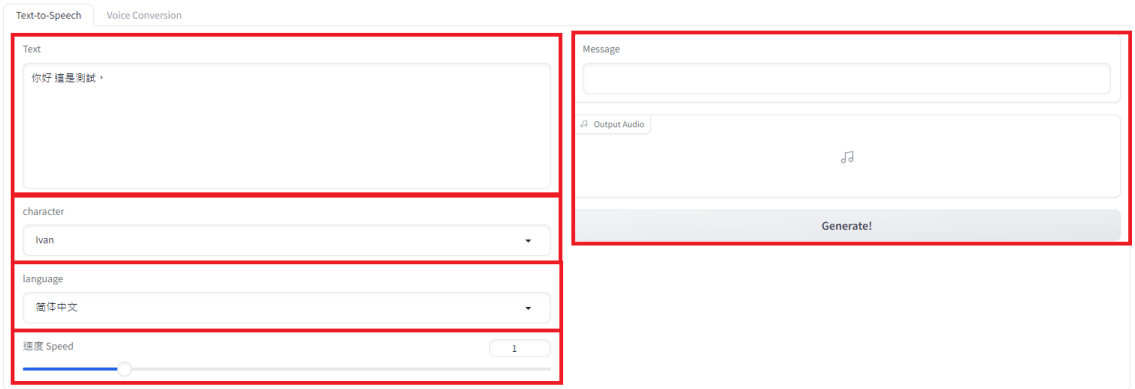
```
content/drive/MyDrive
└──VITS-fast-finetuning
    ├──G_latest.pth
    └──finetune_speaker.json
```

使用者介面

1. 將VITS模型訓練出的*G_latest.pth*和*finetune_speaker.json*兩個檔案複製到下圖的資料夾中

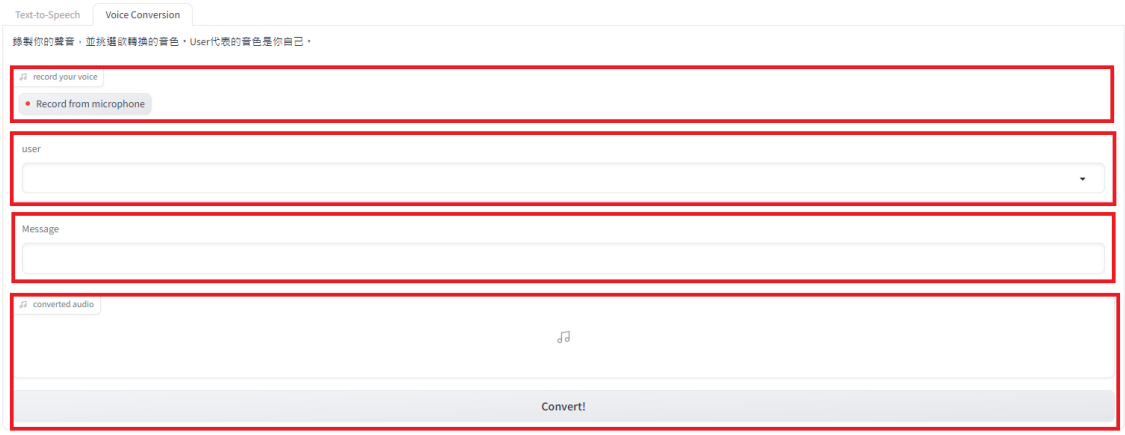
```
VITS-fast-fine-tuning
└──fine_tune_models
    ├──G_latest.pth
    └──finetune_speaker.json
```

2. 開啟 *run.ipynb* 並執行所有儲存格
3. 介面展示
    - T2S Interface



欄位	使用方法
Text	輸入欲翻譯文句
character	選擇欲切換之角色
language	輸入文句之語言
speed	說話速度
Message	按下Generate生成音訊，並可直接播放

- S2ST Interface



欄位	使用方法
record your voice	錄製欲翻譯之語音
user	選擇欲切換之角色
Message	是否轉換成功
converted audio	按下Generate生成音訊，並可直接播放

建議

- 不要用Google Colab運行此程式碼，會出現版本Colab==1.0.0和套件的版本衝突
- 轉換時記得按下欲轉換角色之選項！

Reference

- https://github.com/Plachtaa/VITS-fast-fine-tuning (https://github.com/Plachtaa/VITS-fast-fine-tuning)