運用基於生成預訓練轉換器架構的 OpenAI Whisper 多語言語音辨識引擎之台語及華語語音辨識之實作

Taiwanese/Mandarin Speech Recognition using OpenAI's Whisper Multilingual Speech Recognition Engine Based on Generative Pretrained Transformer Architecture

長庚大學 資工所 謝岳哲
指導教授 呂仁園 博士
@Chang Gung University, Taoyuan, Taiwan
Computer Science and Information Engineering

MSP Lab
Chang Gung University

背景與動機

- Whisper 發表前著重於即時的聲音處理系統
- 實際使用語音辨識技術
- 2022年9月 OpenAI 發表了 whisper 可對超過90種語言進行語音辨識的模型
- 2022年10月 Meta 使用台語連續劇語料建立台語AI 翻譯系統
- 2022年12月 Huggingface 舉辦 whisper Fine-Tuning Event

目的

- 使用台語資料庫與我們從電視公司 Youtube 網站收集的台語連續劇對 OpenAI-Whisper 進行微調
- 使用 CNN 與語音辨識基底的即時語音辨識遊戲
- 使用CNN、LSTM 合併的 LRCN 和 U-net 進行音樂的歌聲偵測與分離所進行的卡拉OK系統

章節

- 多語言語音辨識引擎： OpenAI Whisper 台語語音辨識
- 詞彙辨識：即時語音辨識遊戲
- 歌聲辨識：卡拉OK系統

多語言語音辨識引擎

OpenAI Whisper 台語語音辨識

章節內容

- 研究背景與動機
- 研究目的：微調whisper
- 模型架構：transformer、whisper
- 語料集介紹： Common Voice 、民視
- 實驗結果：台語、華語
- 未來方向：語料集、微調小型模型

研究背景與動機

- 根據台灣109年人口及住宅普查
  - 台灣人有6,897,535人使用台語為主要使用語言，約31.7%
  - 有18,728,839人會說台語，約86.0%
- 台語為非書寫語言，在目前的語音辨識中難以進行建置
- 2022/9/21 OpenAI whisper 可對 97 種語言進行語音辨識的模型
- 2022/10/20 Meta 使用台語連續劇語料建立閩南語 AI 翻譯系統
- 2022/12/5 Huggingface 舉辦 whisper Fine-Tuning Event 並提供了 Whisper checkpoints 讓所有人能使用不同的語言微調模型

研究目的

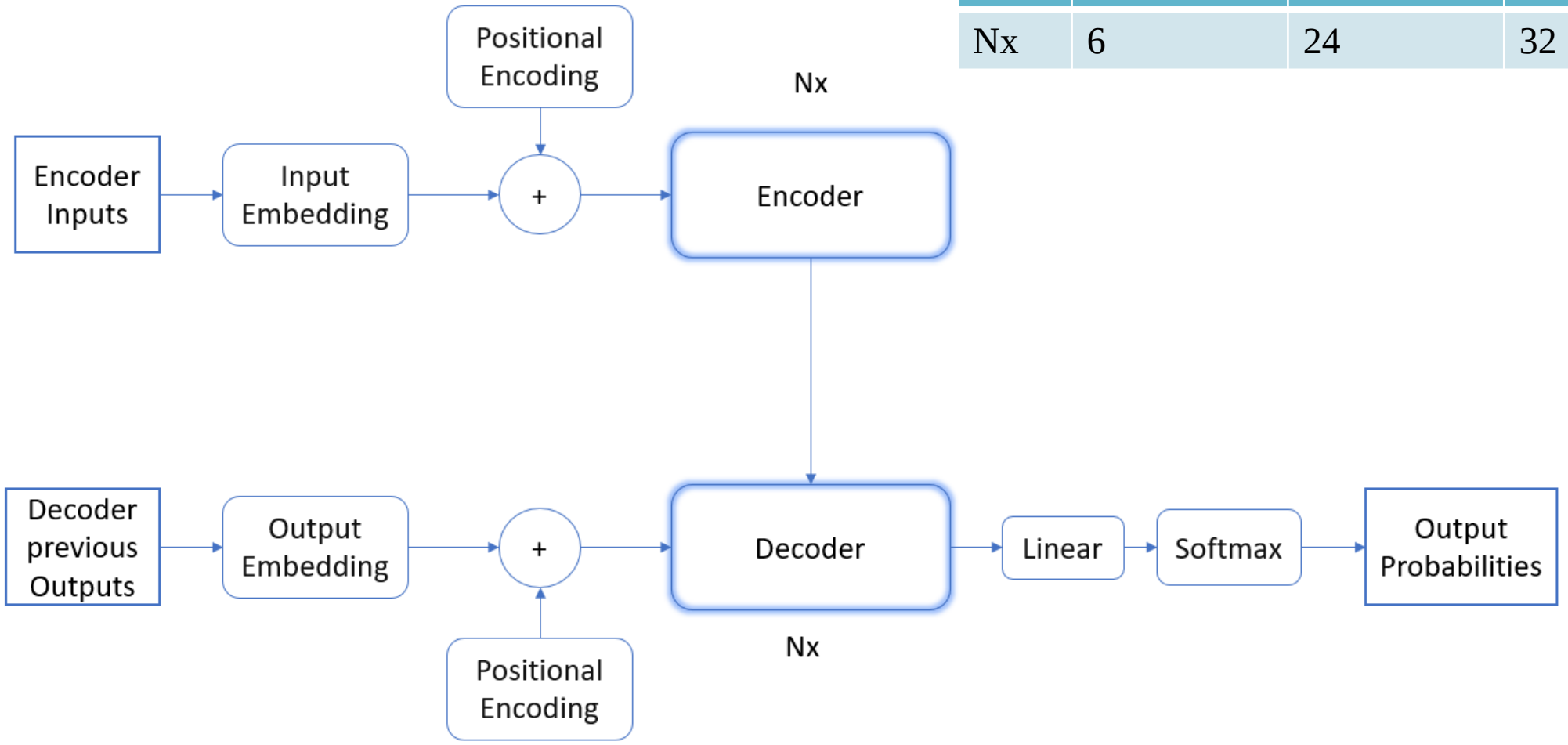
- 對 whisper 的微調進行研究
- 嘗試使用 whisper 對台語進行語音辨識
  - 輸出台語文字
  - 輸出華語文字

Whisper

- OpenAI 所建立的弱監督預訓練語音辨識模型
- 使用了 680,000 小時的標記音訊數據，其中有 117,000 小時是英文以外的其他 96 種語言
- 使用”Attention Is All You Need“ 編碼器-解碼器Transformer架構
- 英文的WER最佳為2.7%，在未使用於訓練的語料集的WER為12.8%，比wav2vec 2.0下降了55.2%
- 華語的WER最佳為12.1%，多語言語音辨識的語音辨識語料使用23446小時

Transformer網路架構

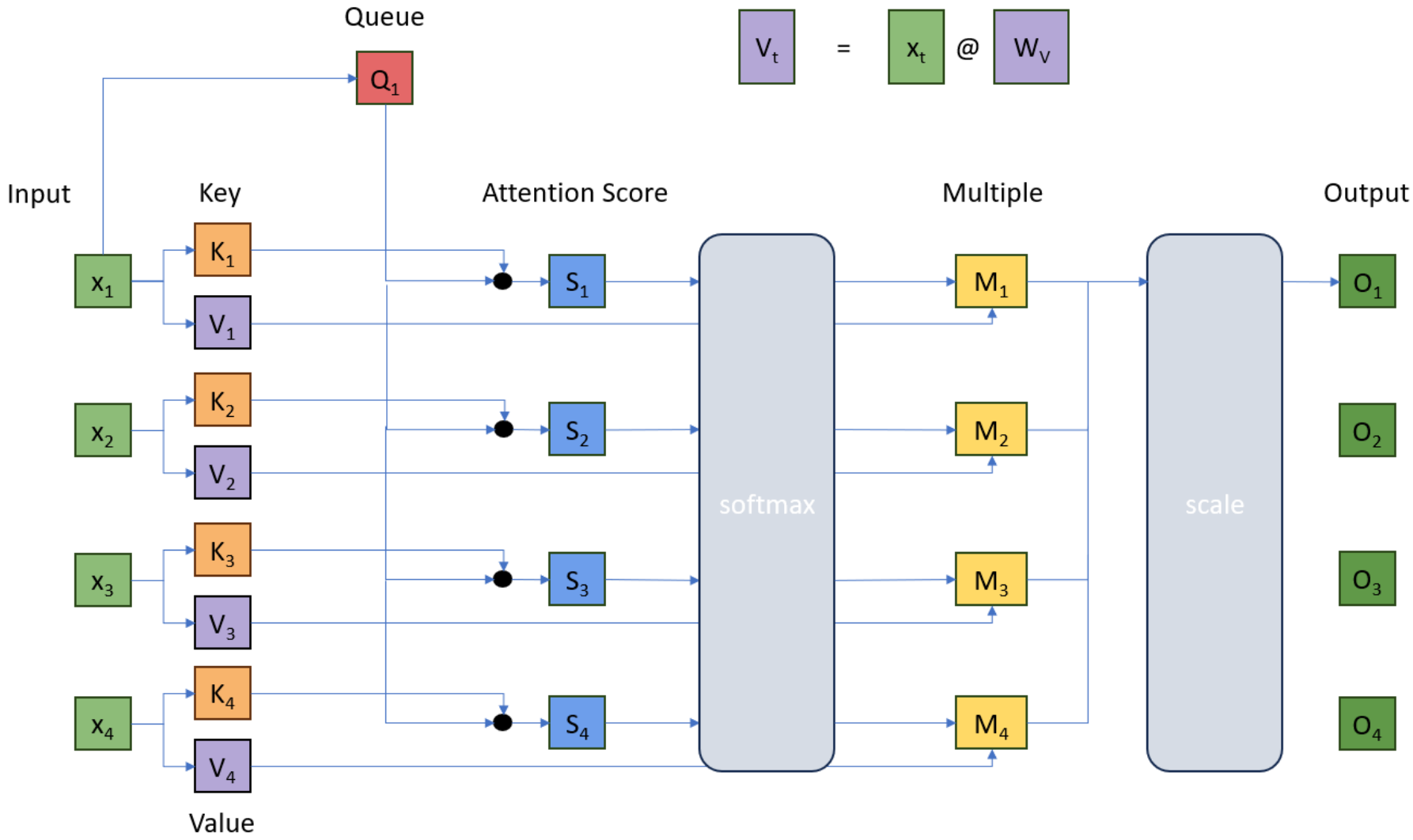
層數	Attention	Medium	Large-v2
Nx	6	24	32



自注意力機制

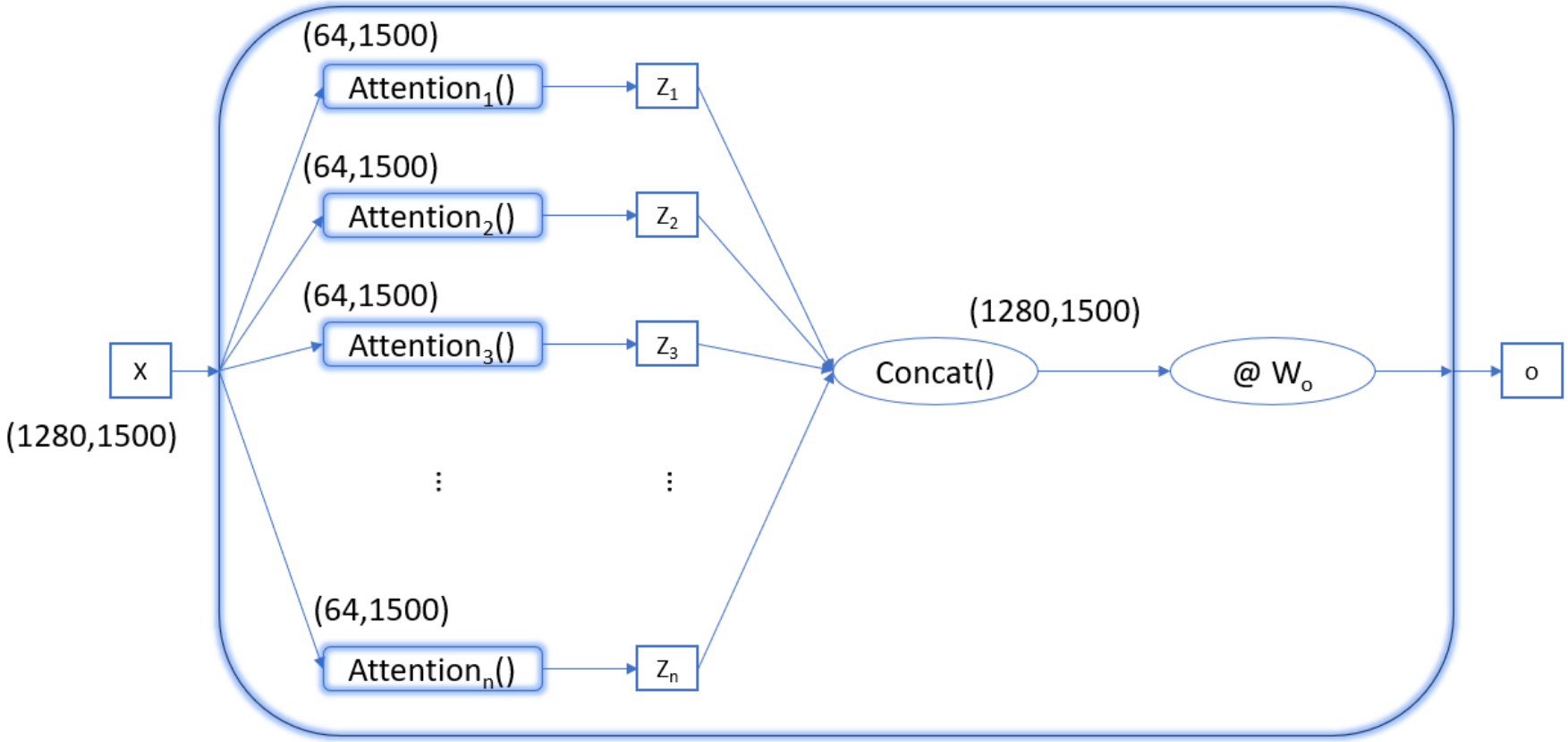
$$Q_t = x_t @ W_Q$$
$$K_t = x_t @ W_K$$
$$V_t = x_t @ W_V$$

Q,K,V維度
64



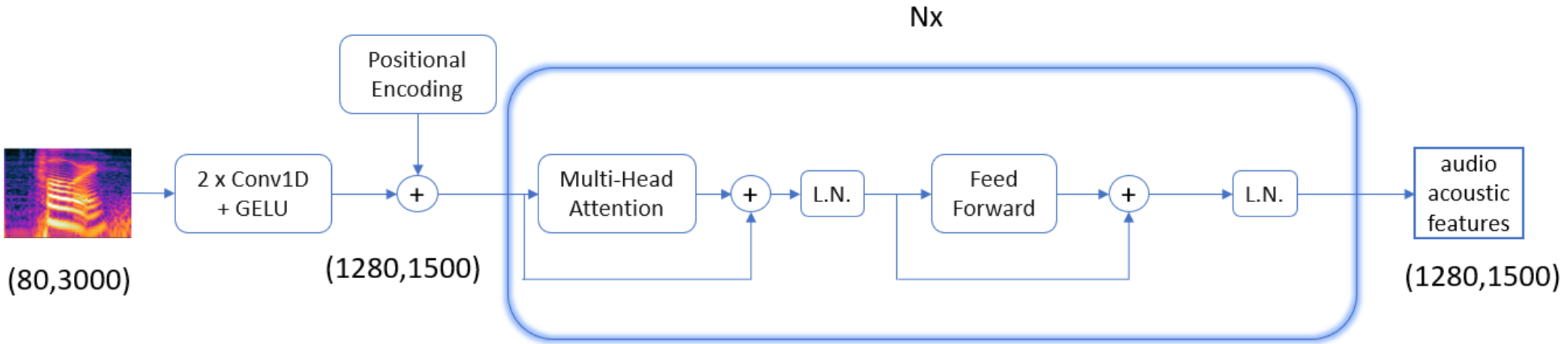
多頭注意力

Head數	Medium	Large-v2
n	16	20

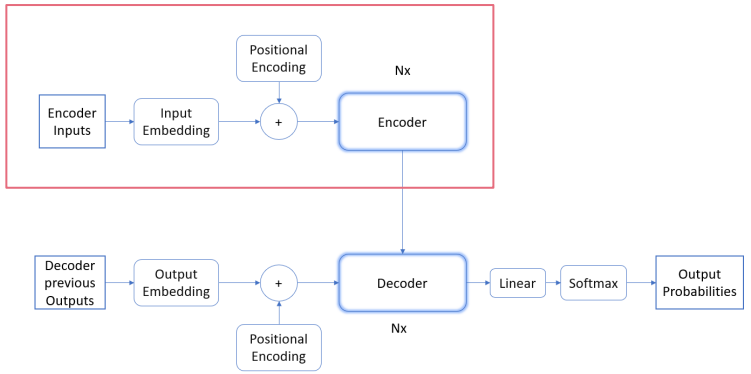


Whisper Encoder

	Medium	Large-v2
Head	16	20
width	1024	1280
Nx	24	32

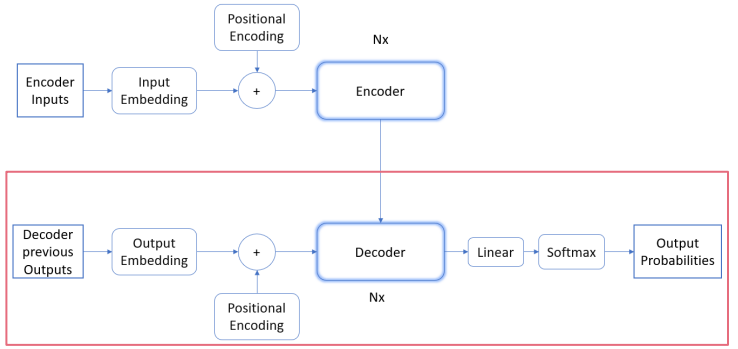


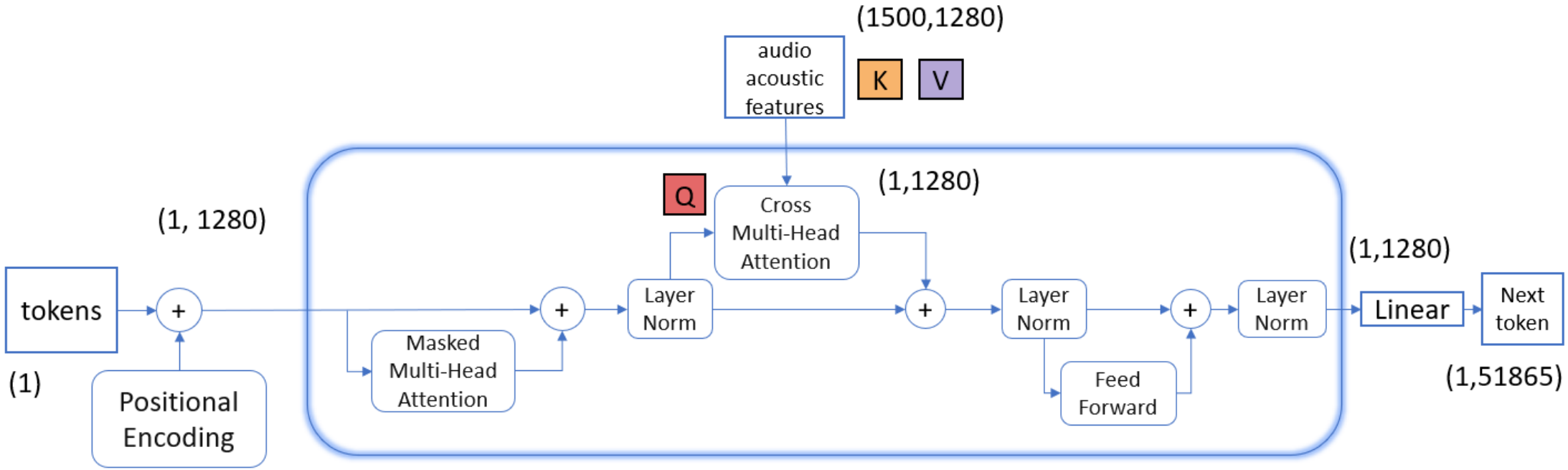
inputs
16,000 Hz
30s 長
80 channels的頻譜圖
25 ms的窗框
每步長10 ms



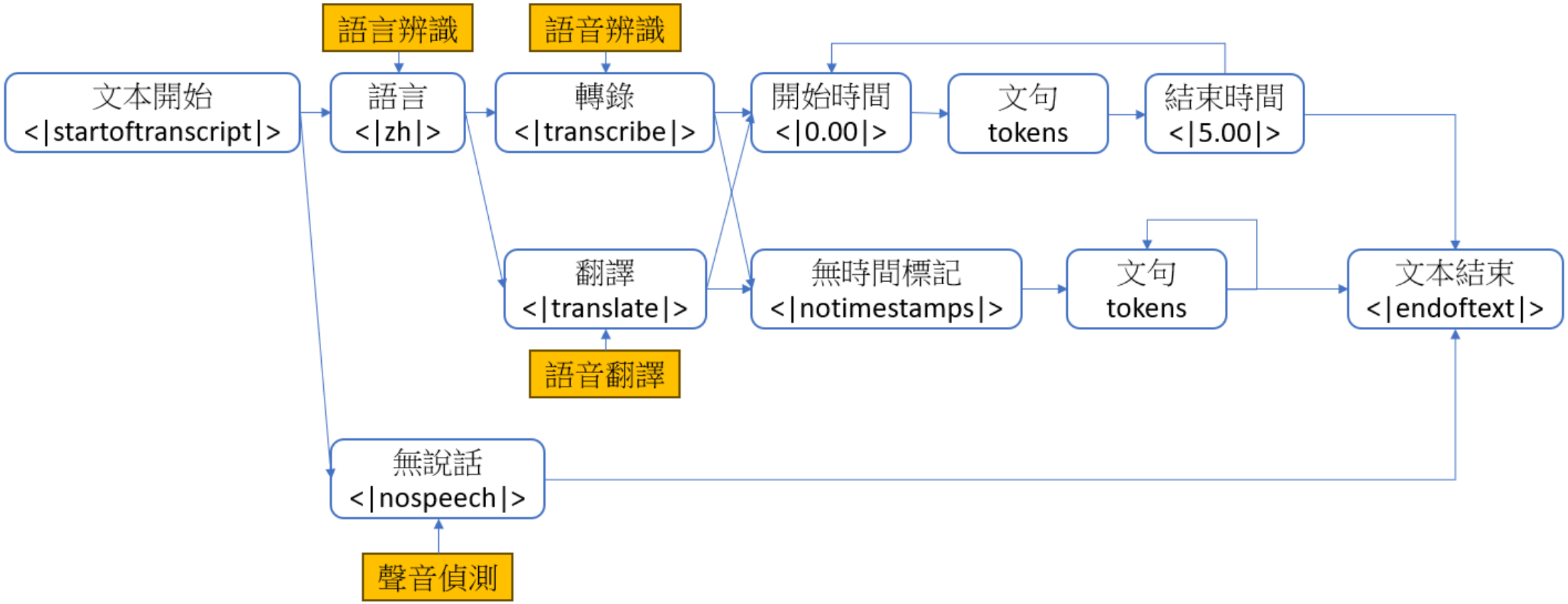
MSP Lab
Chang Gung University

Whisper Decoder





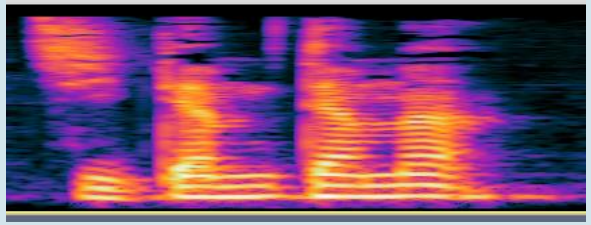
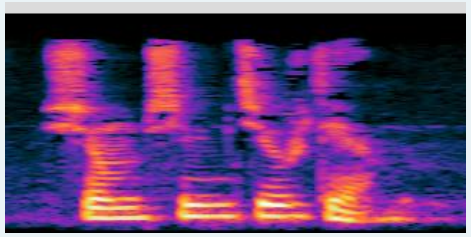
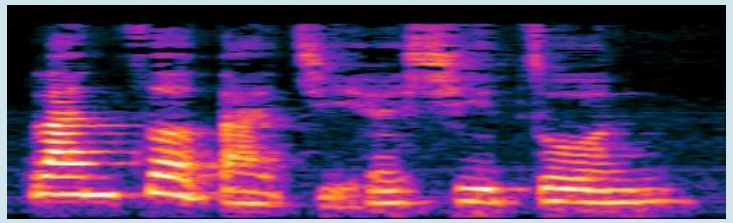
Multitask training format of tokens for Whisper decoder



台語語料集/台語文字

- Common Voice nan-tw
  - 120 人錄音
  - 11 小時錄製完成的片段，其中包含 3 小時的已驗證資料
  - MP3格式
  - 提供台語漢字(漢文)和台語羅馬字(台羅)

Common Voice nan-tw

台語漢字(台語羅馬字)	聲音	頻譜圖
一點點仔 (tsi̍t-tiám-tiám-á)		
傷心酒店(siong-sim tsiú-tiàm)		
咱做代誌的時陣 (Lán tsò tāi-tsì ê sî-tsūn)		

台語語料集/華語文字

- 台語連續劇
  - 民視戲劇館 YouTube 收集約920小時的影片
  - 市井豪門74小時，阿不拉的三個女人46小時，風水世家800小時
  - 民視戲劇館可下載的共有19821部影片約 7445小時

台語連續劇

檔名	文本	聲音	實際翻譯
市井_001_0094.mp3	世明 春梅 欠錢要還 我還有孩子的補習費要繳		世明 春梅 欠錢要還阿夠我還有孩子的補習費要納
阿不_001_0188.mp3	你會說台灣話啊 我是台灣人當然會說台灣話		你會說台灣話啊 我台灣人當然要說台灣話
風水_001_0299.mp3	媽，先喝杯熱開水祛寒		阿母阿 先喝一杯熱開水較不會冷

微調流程

- 使用Whisper Fine-Tuning Event 提供的微調方式進行微調
  - Hugging face上提供之 openai/whisper-large-v2 和 openai/whisper-medium
  - whisper-fine-tuning-event/run_speech_recognition_seq2seq_streaming.py
  - Transformer Seq2SeqTrainer
- Common Voice nan-tw 事先去除羅馬拼音部分
- 台語連續劇使用Hugging face的dataset格式
  - 依照台詞時間點切割聲音檔
  - 不超過10秒或30秒

微調結果：台語輸出

台語文字	聲音	辨識結果
我攏有看著		我攏有看著
大海毋驚大水		大海毋驚大水
大鑼大鼓		大路大股

辨識模型	CER(%)
Huggingface Medium	96.6
Huggingface Large-v2	96.7
Medium Fine-tune	50.9
Large-v2 Fine-tune	52.8

微調結果：華語輸出

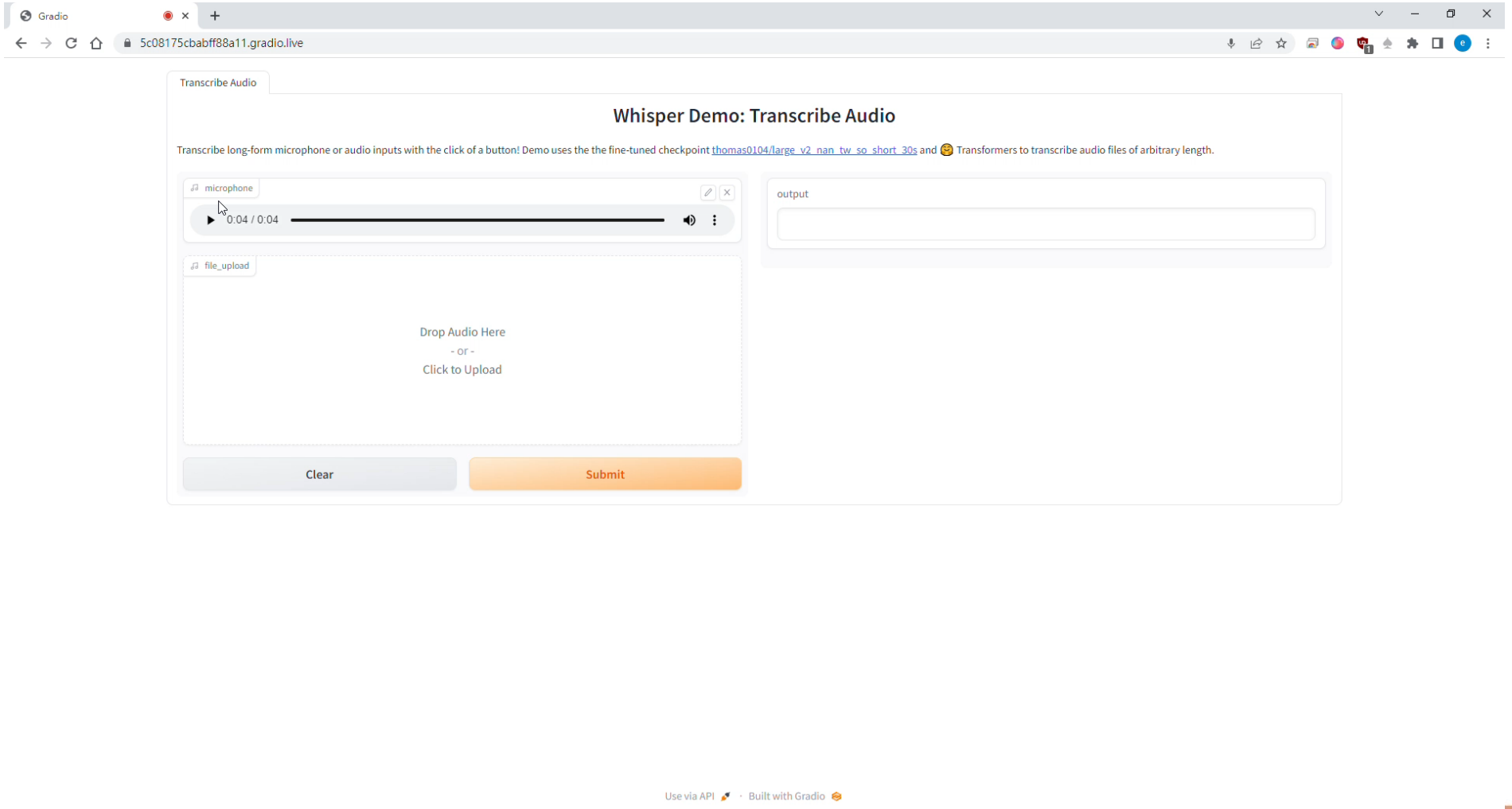
字幕	聲音	辨識結果
媽媽沒有準備生日禮物 但是... 從今天開始		媽媽沒有準備生日禮物不過從現在開始
謝謝你送阿不拉回來		謝謝你送阿不拉回來
如果大海能夠喚回曾經的愛		如果大海能夠換回曾經的愛情

辨識模型	每句台詞長度(秒)	CER(%)
Medium Fine-tune	10	82.6
Medium Fine-tune	30	71.5
Large-v2 Fine-tune	10	53.8
Large-v2 Fine-tune	30	50.7

未來方向

- 語料集
  - 台語：增加資料量
  - 連續劇：增加資料量，台詞處理，聲音處理
- 微調小型模型
  - 使用或微調 medium 與 large-v2 需要大量的VRAM
  - 嘗試微調 base 或 small 與後處理

Live demo



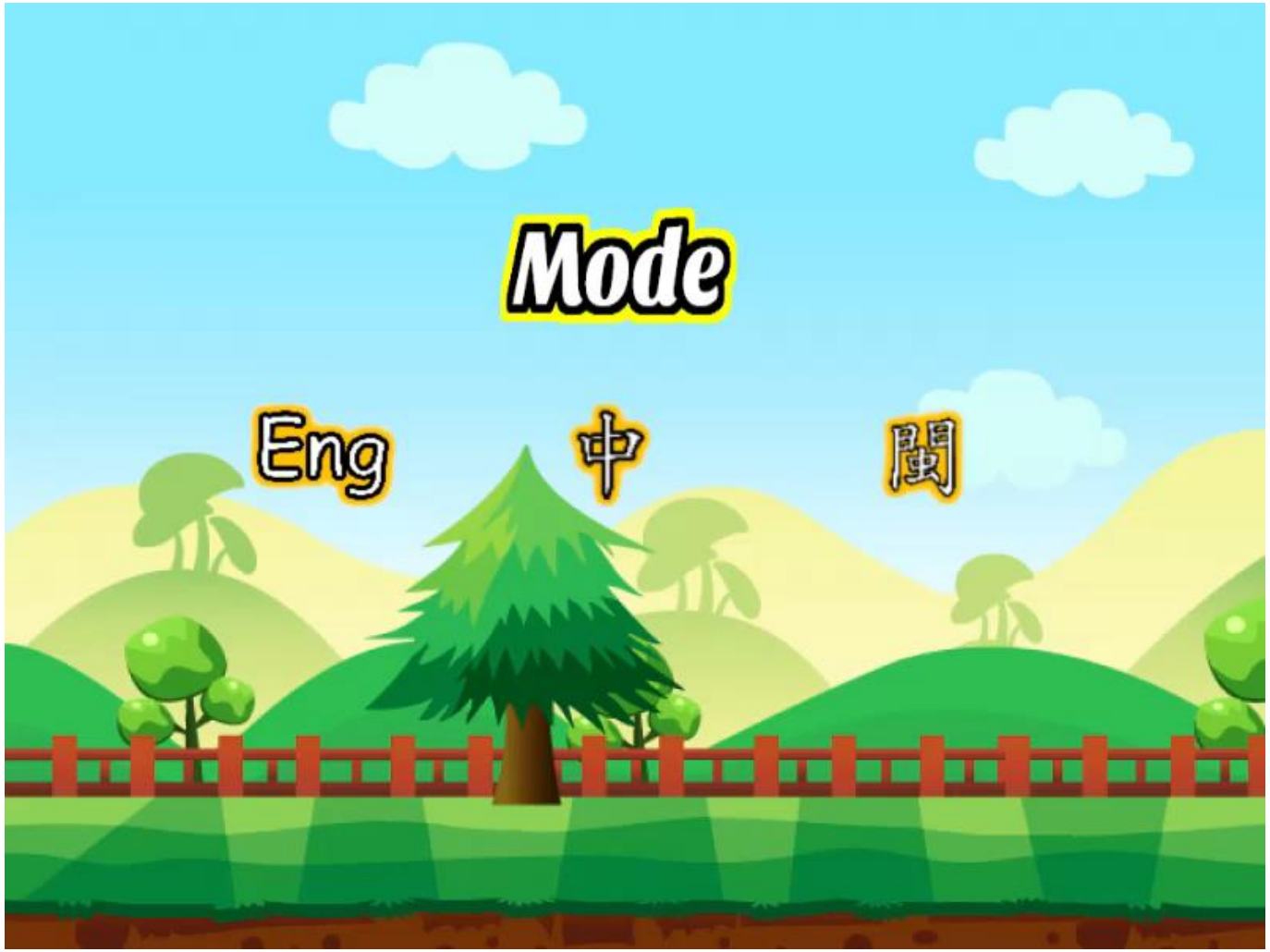
詞彙辨識

即時語音辨識遊戲

章節內容

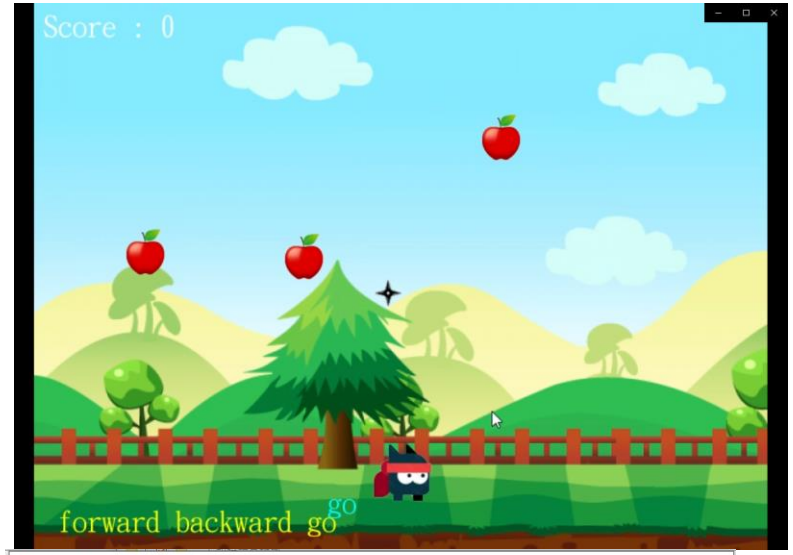
- 目前成果展示：遊戲畫面
- 研究背景與動機
- 研究目的
- 程式架構：平行處理、語音控制
- 語料集介紹： Google speech command V2
- 模型架構：2D CNN
- 實驗結果
- 未來方向：個人化、增加詞彙與啟動詞、端到端模型訓練

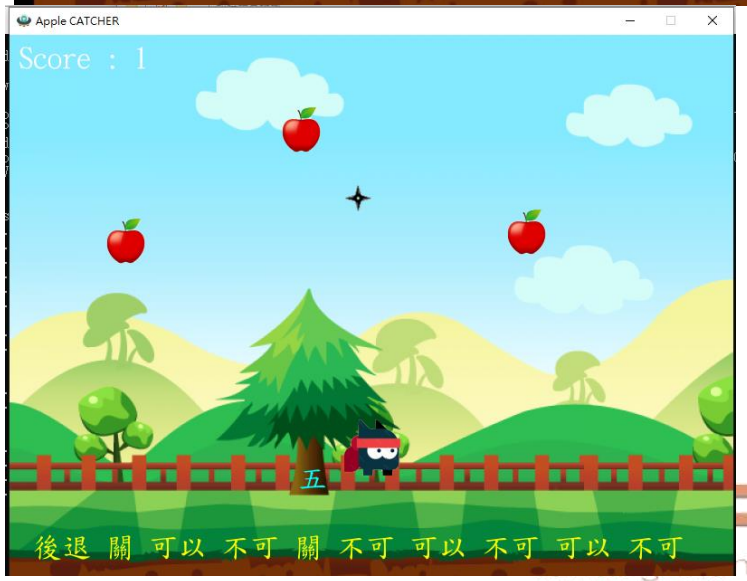
目前成果展示



目前成果展示

- 下方的黃色文字為辨識到的命令詞彙
- 角色左邊的藍色文字為即時辨識到的詞彙
- 角色在聽到命令詞彙會執行動作
  - 前進後退為畫面上下方移動
  - 左右為畫面左右方移動
  - 可以/go 的時候會發射手裡劍
  - 不可/stop 為停止移動並裝填手裡劍





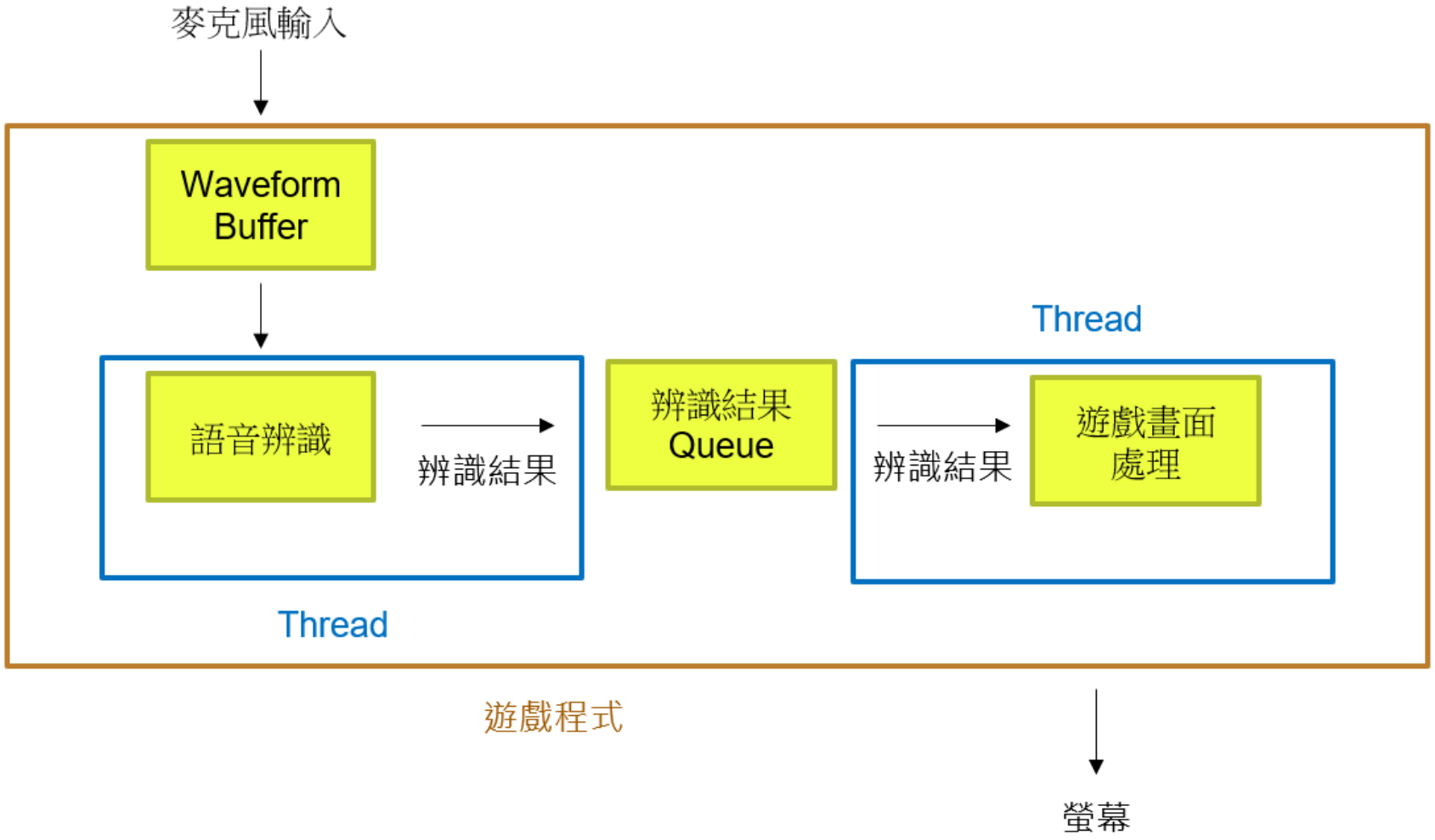
研究背景與動機

- 在語音辨識技術發展逐漸成熟的現在，只要有語音資料與網路模型，人們即可建構出簡單的語音辨識模型。
- 在對機器下達語音指令時，並不需要大量的詞彙進行控制。

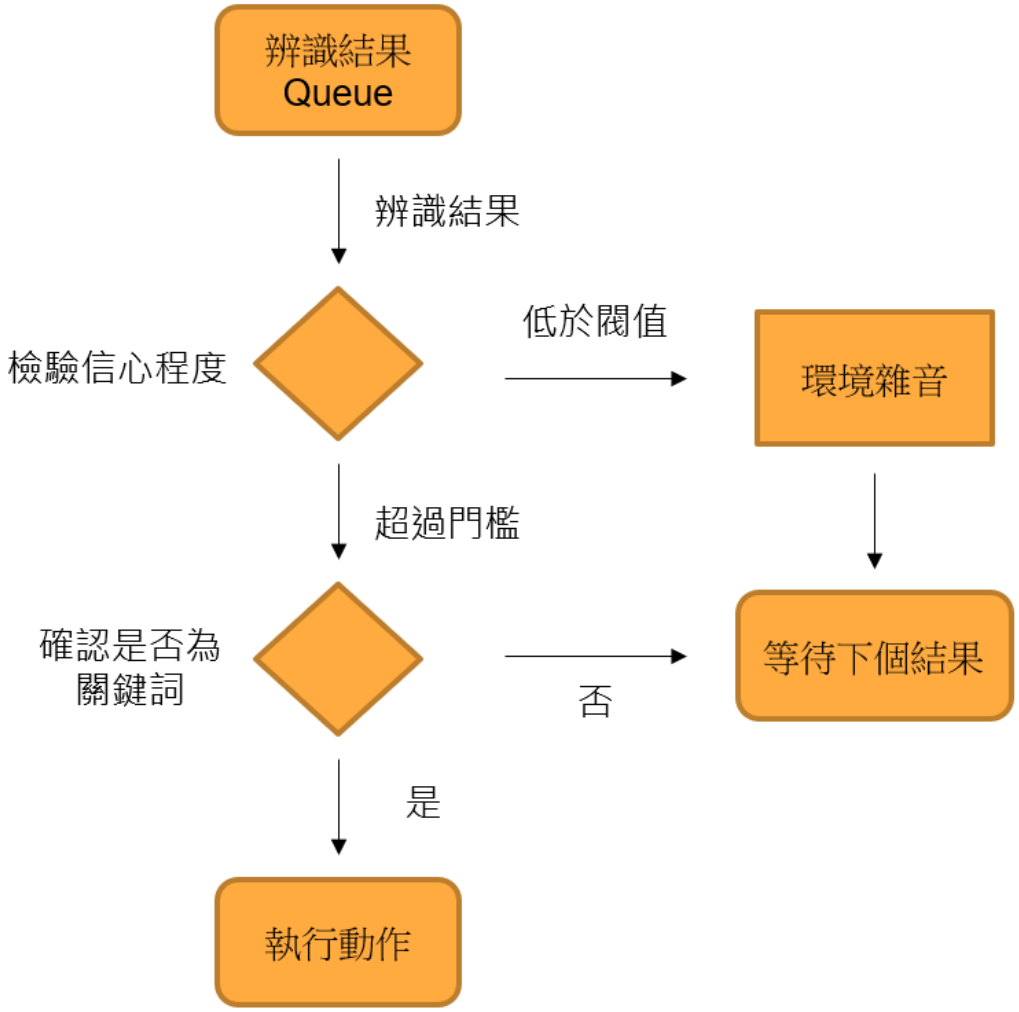
研究目的

- 製作即時詞彙辨識的聲控遊戲系統
- 模型小巧且精確
- 使用2D CNN進行模型訓練
- 可以英文、華語、台語之間進行切換

程式架構：平行處理語音輸入和遊戲



程式架構：語音控制

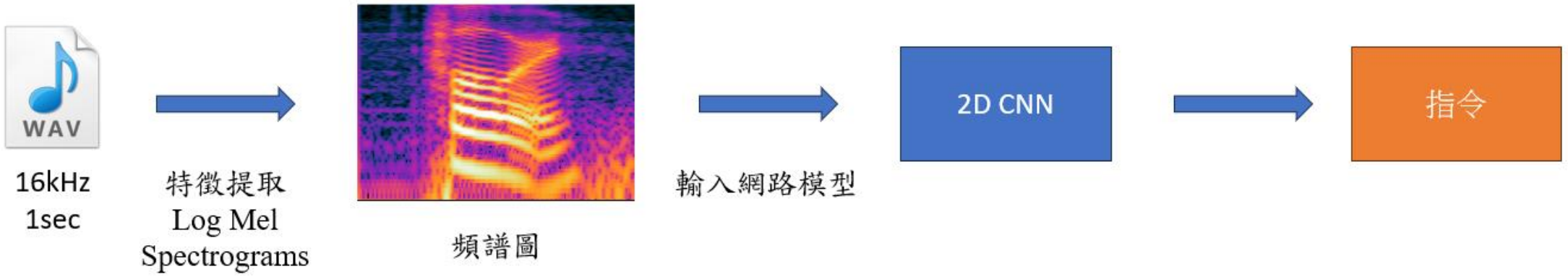


語音資料

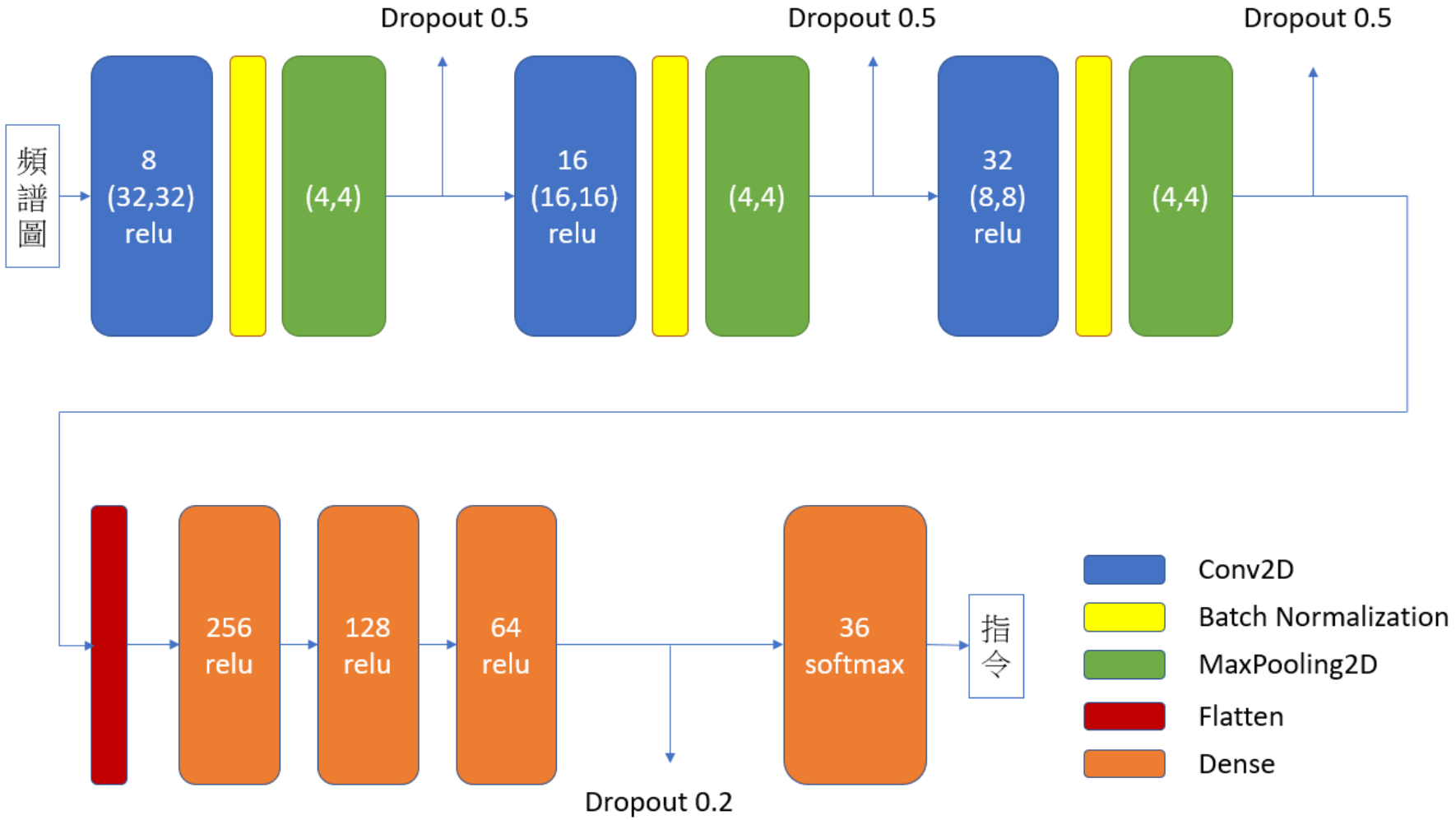
- 英文使用 Google speech command V2 語料集
- 2618人錄音，約30小時
- 每個錄製的語音檔案固定為一秒、採樣率(Sampling Rate)為 16000hz、 32bit 的浮點數(float)、單聲道的wav檔
- 35種的指令，分為數字、方向、開關、是否、名詞

語言	來源	時數(小時)	人數
英文	Google speech command V2	29.4	2618
中文	實驗室收集	1.3	3
台語 (閩南語)	實驗室收集	5	15

語音辨識流程



模型架構：2D CNN



辨識結果

- 三個語言的辨識率皆達到90%
- 因中文資料集所包含的人數較少，模型訓練為使用訓練集二人驗證集和測試集是另一人。
- 在英文辨識率相近的情況下，我們的模型.h5檔大小只有3MB，參數為348K，而Whisper tiny有73MB ，參數為39M

訓練語言	辨識率	loss	Whisper tiny辨識
英文	0.8998	0.4086	**0.9**
中文	0.9015	0.3109	**0.12**
台語	0.8969	0.5240	**x**

未來方向

- 個人化
  - 提取使用者語音特徵
  - 少量或單一語者
- 增加詞彙與啟動詞
  - 增加能辨識的指令，如：開始、結束
  - 使用啟動詞確認要辨識的詞
- 端到端模型訓練
  - 訓練語音辨識時，直接輸入原始的音訊波型
  - 3分鐘內即可有70%辨識率

歌聲辨識

卡拉OK系統

章節內容

- 目前成果展示：卡拉OK畫面
- 研究背景與動機
- 研究目的
- 程式流程：音樂前處理、主程式
- 資料集介紹： Jamendo Corpus、Electrobyte
- 模型架構：U-Net、LRCN
- 實驗結果
- 未來方向：歌詞顯示

目前成果展示

目前成果展示

- 白色圓圈為使用者所唱的音高
- 黃色圓圈為歌手所唱的音高
- 左下顯示資訊：
  - 音高相同次數
  - 歌手所唱的音高音名、頻率
  - 使用者所唱的音高音名、頻率
- 右下顯示當前時間/音樂長度

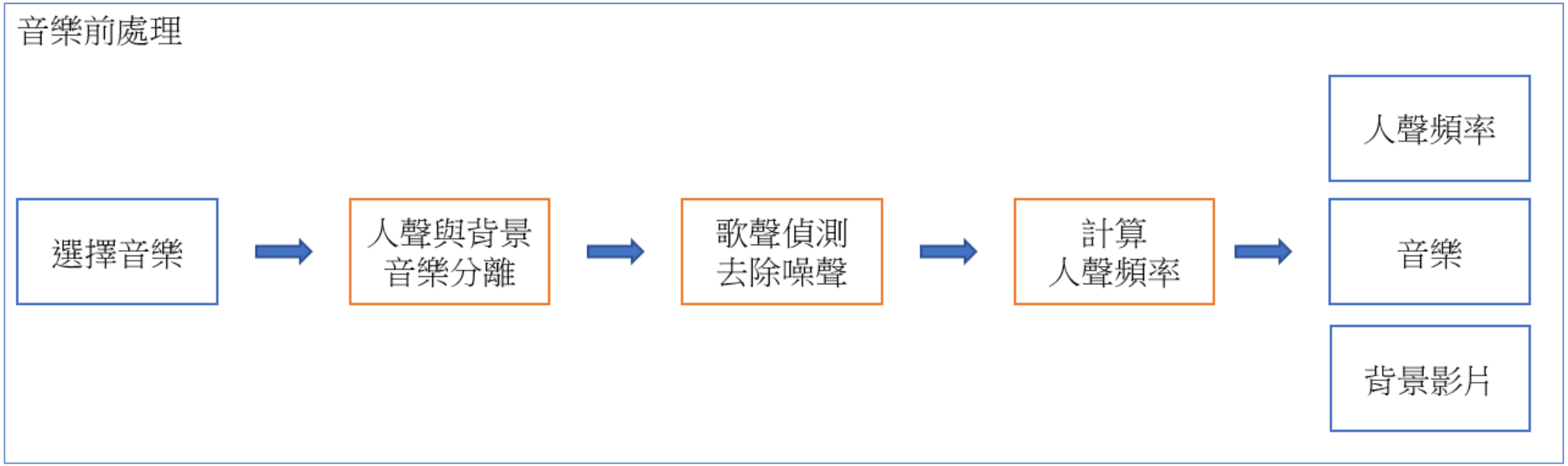
研究背景與動機

- 從音樂中直接進行語音處理如歌詞辨識、歌聲偵測、音高計算等，準確率容易受到背景音樂、雜音等因素的影響。
- 音源分離如 Spleeter 能將人聲部分從音樂中分離
- 分離出的人聲部分還是有可能會包含雜音

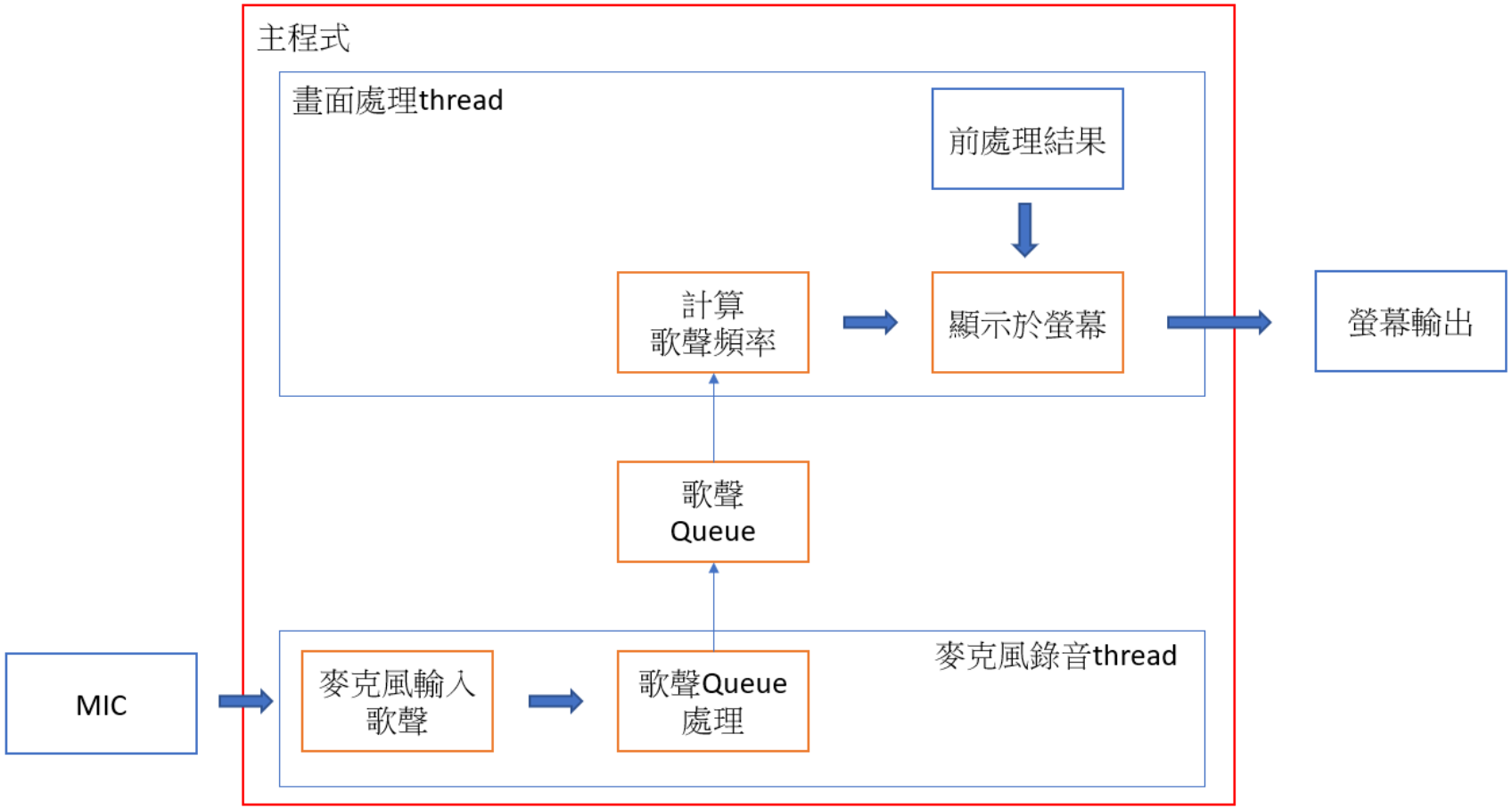
研究目的

- 將音源分離、歌聲偵測、即時計算集成一個卡拉OK系統
- 使用歌聲偵測處理歌聲分離噪聲

程式流程：音樂前處理



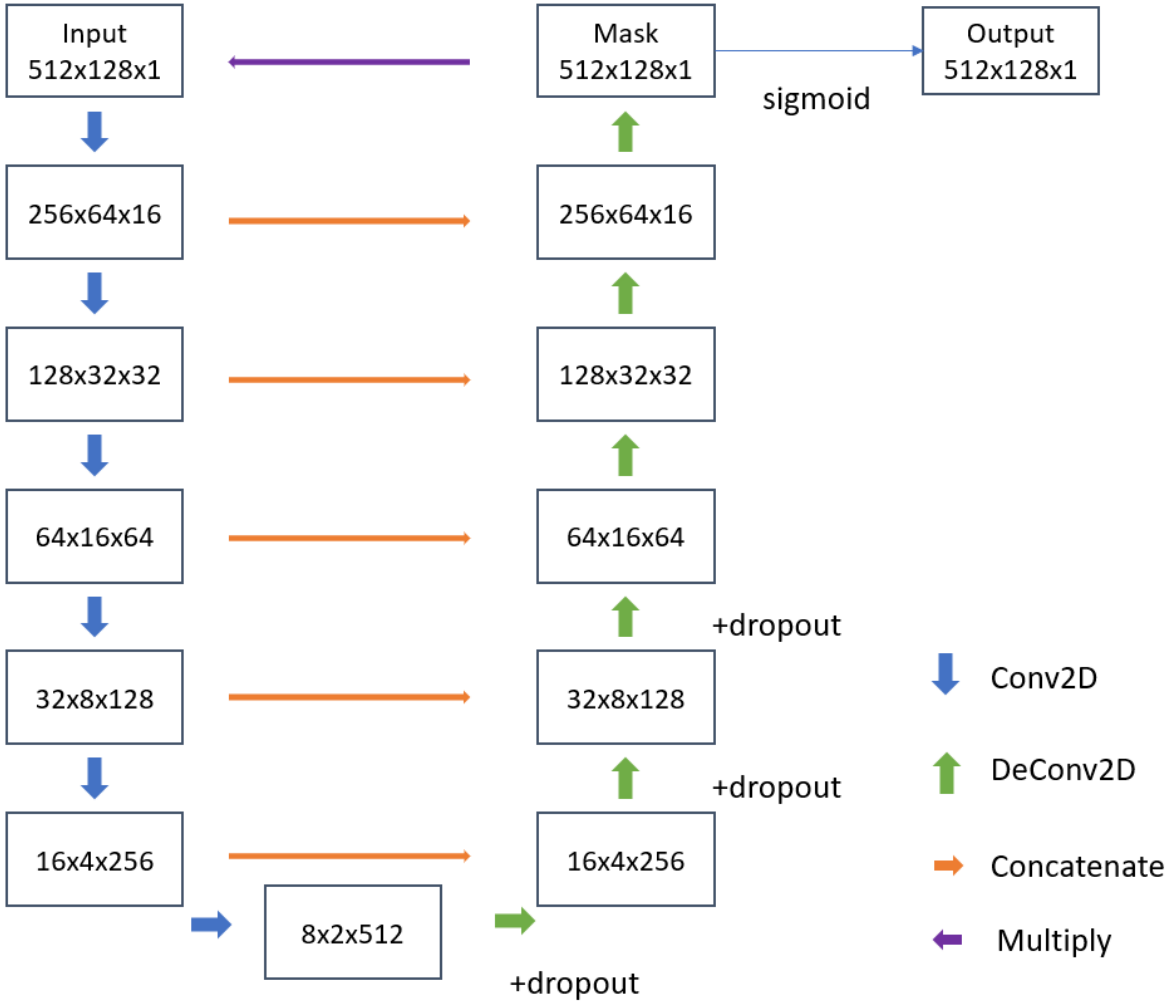
程式流程：主程式



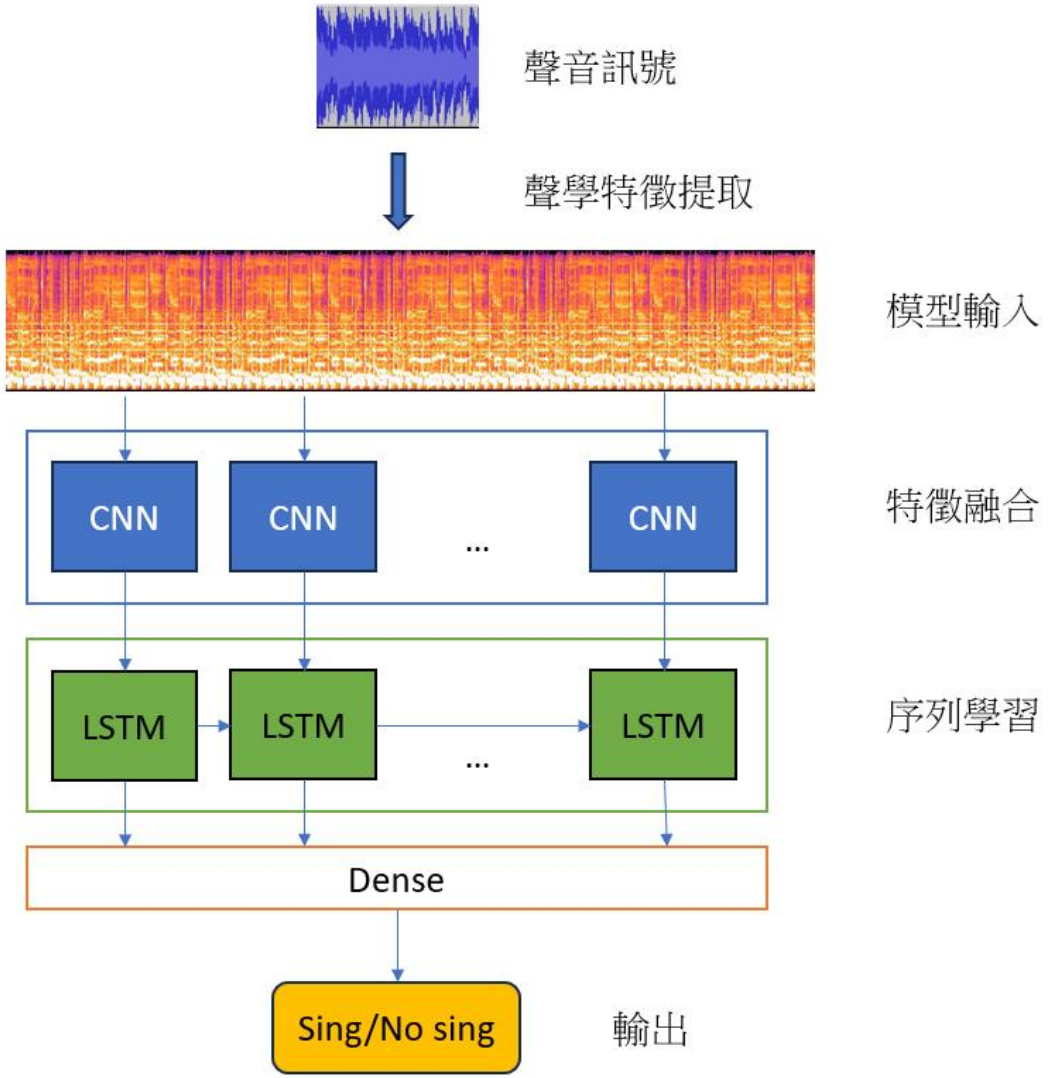
資料集介紹

- Jamendo Corpus
  - 93 首在 Jamendo 免費音樂分享網站上擁有創用CC授權的音樂
  - 標記「唱歌」和「無唱歌」部分
  - 雙聲道 OGG 44.1kHz 與 112KB/s 位元速率
- Electrobyte
  - 90 首電子風格音樂
  - mp3 格式
  - 標記「唱歌」和「無唱歌」部分

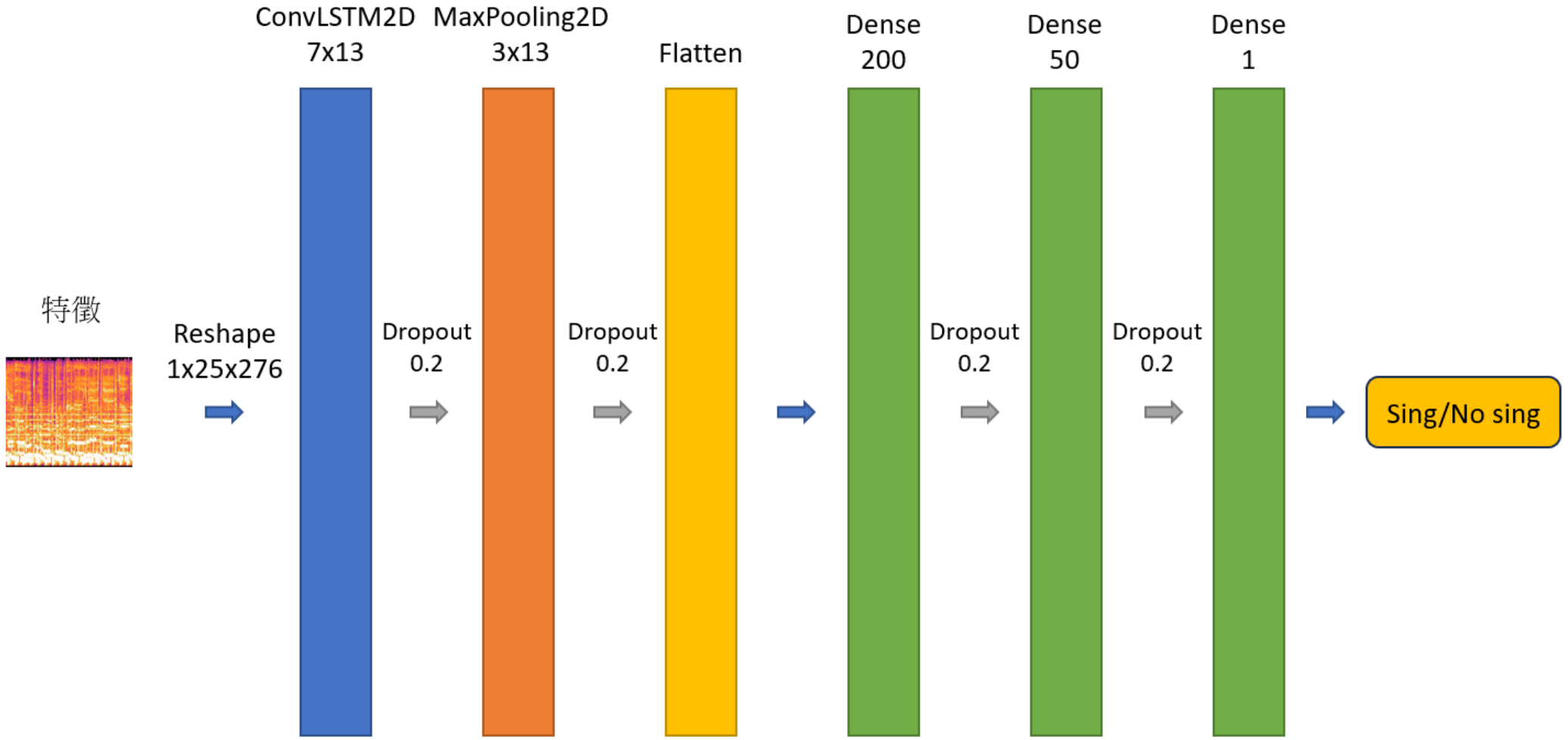
人聲與背景音樂分離：U-net



歌聲偵測：LRCN



LRCN網路架構



辨識輸出

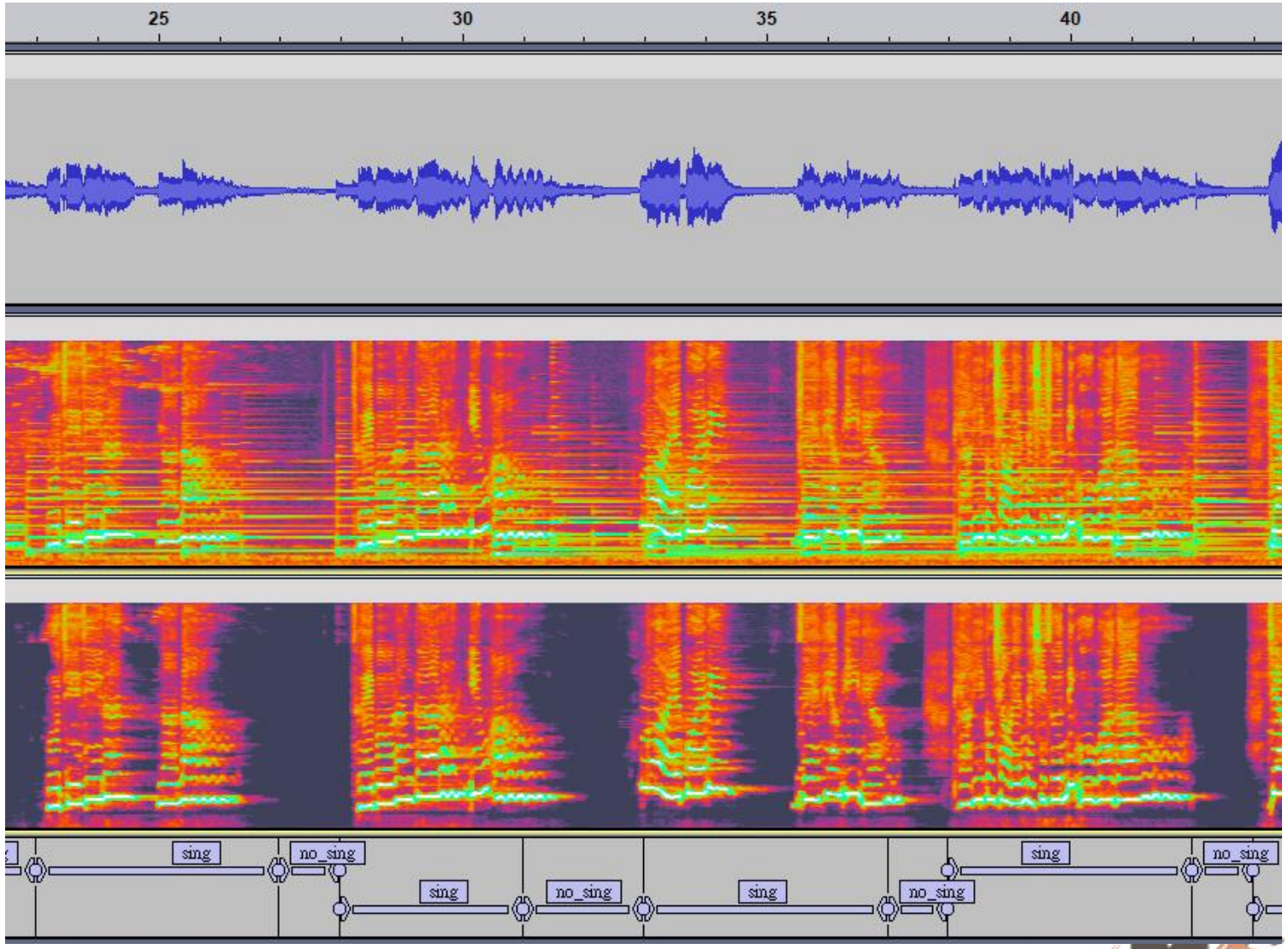
- 以每秒輸出一次是否唱歌

```
1  00000000000000111111001010110000000000
   00000000000000000001111111111100011111
   11111111111111111111111100000000000000
   00000000000001111111101111111111111110
   01111111111111111111111111111111111111
   11111111000000000
2
```

```
 1  0   14  no_sing
 2  14  20  sing
 3  20  22  no_sing
 4  22  23  sing
 5  23  24  no_sing
 6  24  25  sing
 7  25  26  no_sing
 8  26  28  sing
 9  28  57  no_sing
10  57  68  sing
11  68  71  no_sing
12  71  100 sing
13  100 126 no_sing
14  126 135 sing
15  135 136 no_sing
16  136 151 sing
17  151 153 no_sing
18  153 198 sing
19  198 206 no_sing
20
```

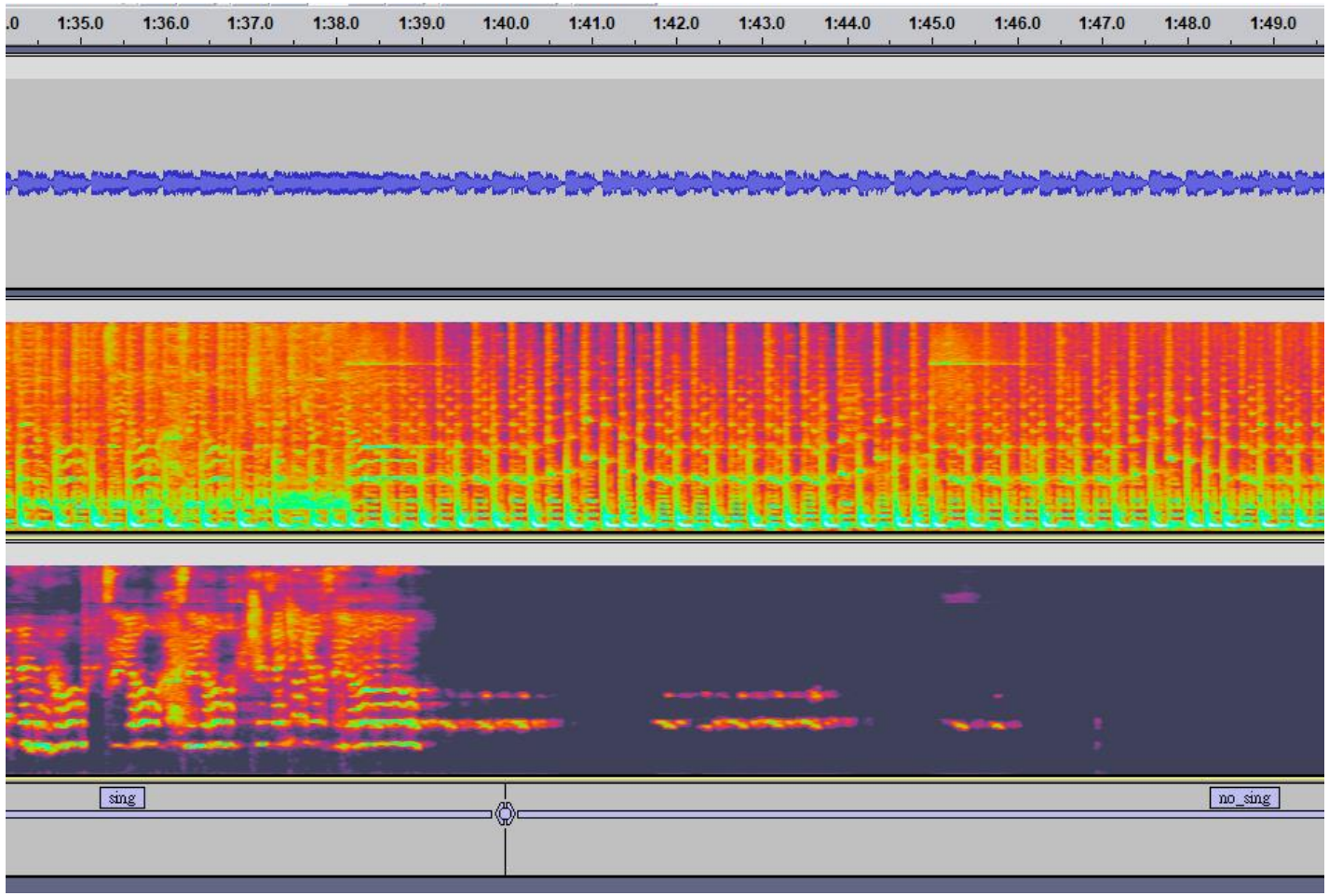
辨識結果：江蕙-家後

- 原始waveform
- 原始頻譜圖
- 歌聲分離後
- 歌聲偵測



辨識結果：鄭秀文-眉飛色舞

- 原始waveform
- 原始頻譜圖
- 歌聲分離後
- 歌聲偵測



辨識結果

- 事先經過人聲分離的辨識結果皆優於未進行人聲分離
- 電子音樂的背景音樂較容易影響人聲分離

Model	accuracy	precision	recall	f1measure
Jamendo vocal	0.9350	0.9251	0.9499	0.9373
Jamendo normal	0.7314	0.7182	0.7821	0.7488
Electrobyte vocal	0.8402	0.8007	0.8685	0.8333
Electrobyte normal	0.8331	0.7957	0.8573	0.8253

未來方向

- 歌詞顯示
  - 網路爬蟲
  - 歌詞進行語音辨識

感謝聆聽