柯南的秘密工具

Voice Conversion

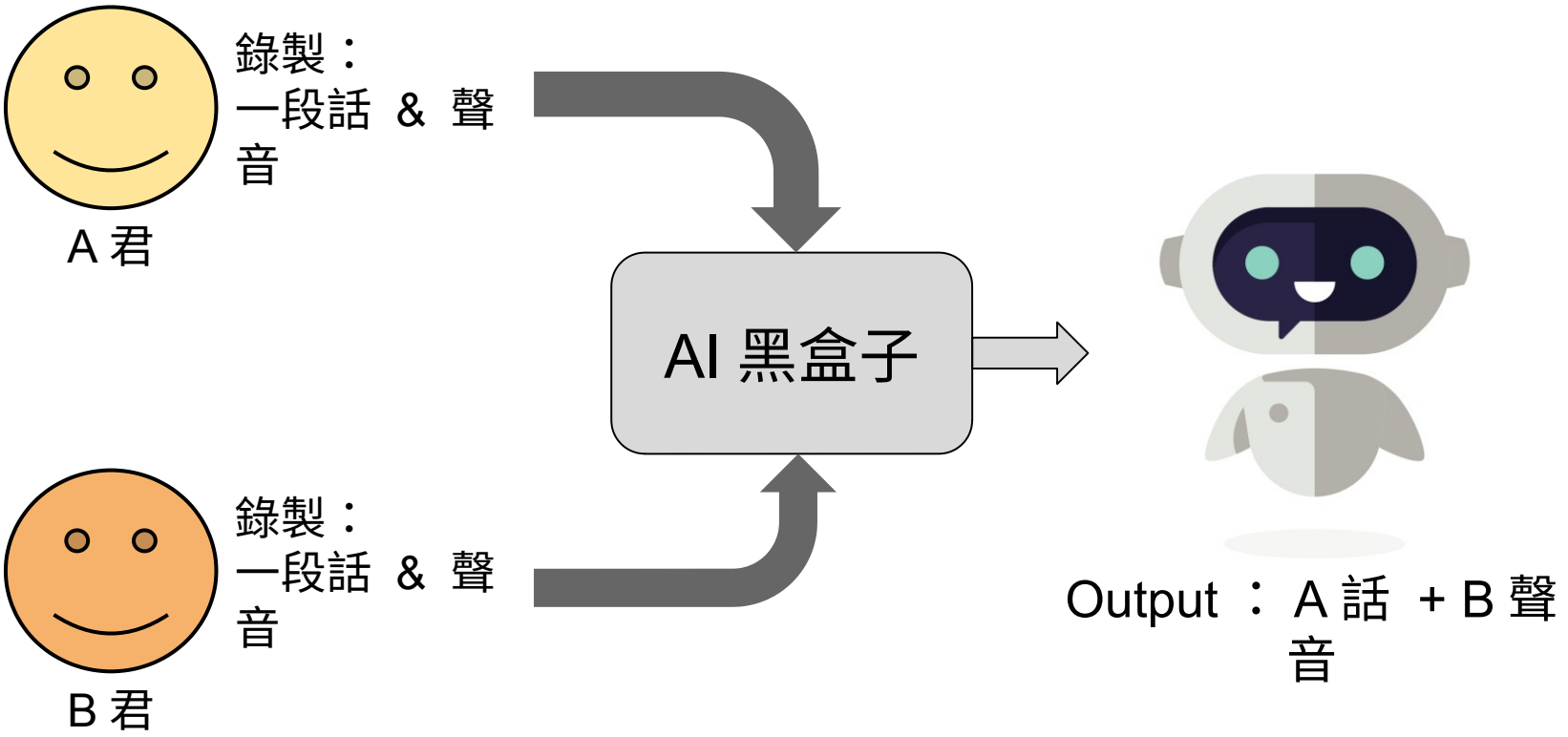
專題組別： ET083-08

組長：ET083023　陳義翔
組員：ET083021　陳威佐
　　ET083035　韓懷恩

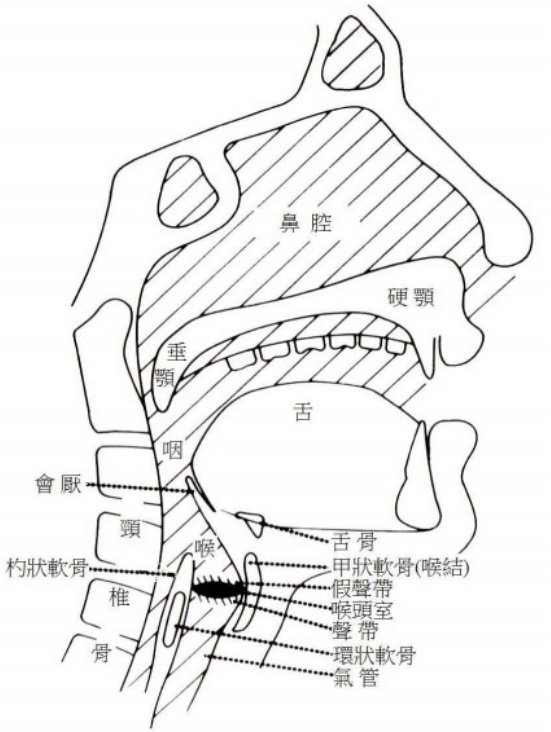
專案目的

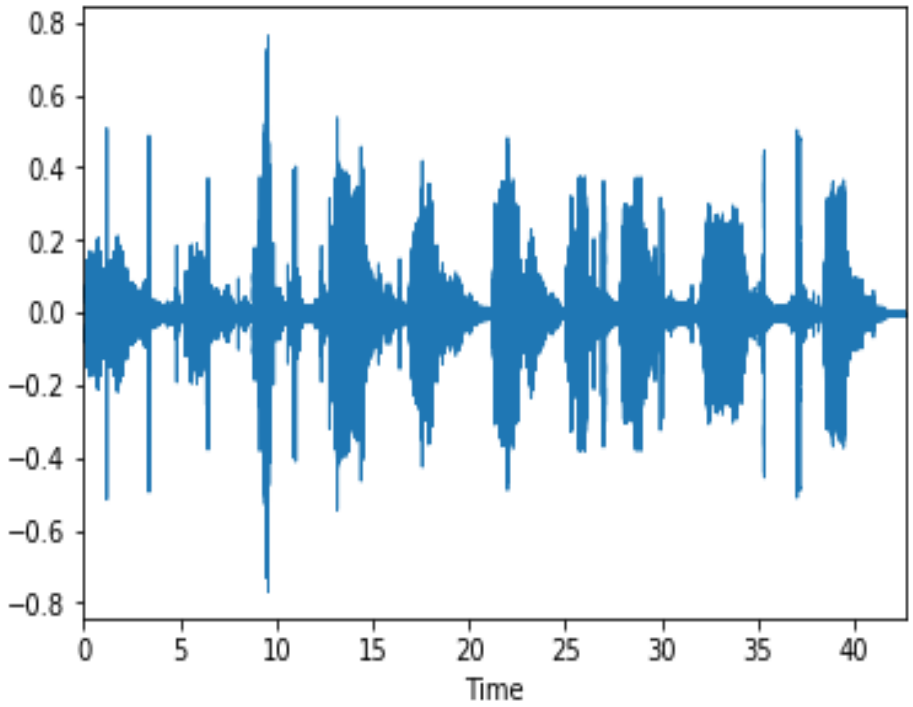
專題目標

柯南領結變聲器模式	即時聲音轉換
專案目標模式	錄音後轉換



語音取樣與前處理原理研析 (1/8)
聲音三特徵 (音調 響度 音色)

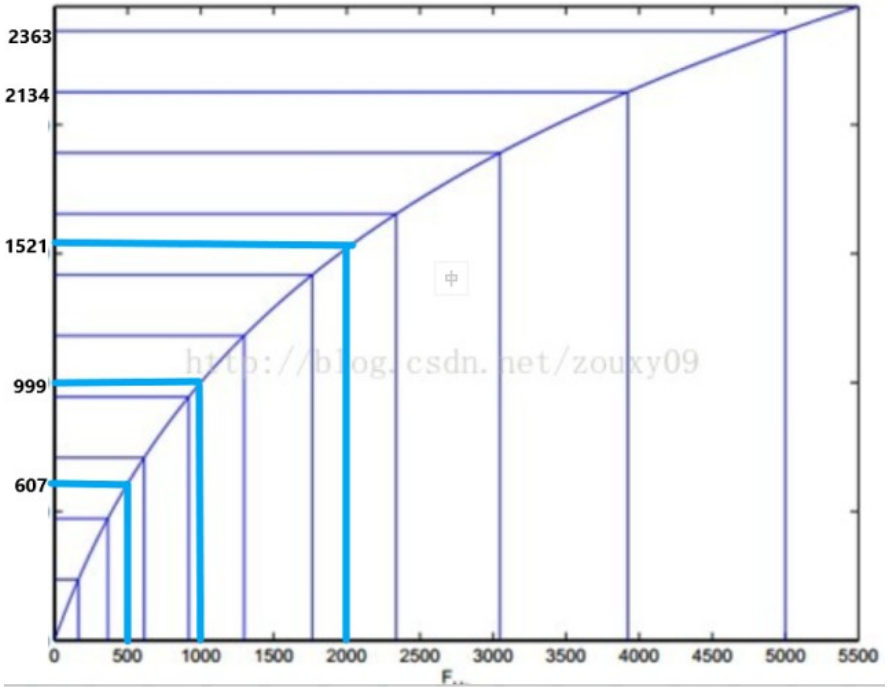




音調：聲音的高低（頻率）
響度：聲音的大小（振幅）
音色：不同高低聲音所產生的組合（多種強弱不同的頻率組成）

語音取樣與前處理原理研析 (2/8)
人類聽覺的生理限制

- 人耳能聽到的頻率範圍是 20-20KHz
- 人耳對低頻音調的感知較靈敏，在高頻時人耳是很遲鈍的。
- 人耳對 Hz 這種標度單位並不是線性感知關係



例如：如果我們適應了 1000Hz 的音調，如果把音調頻率提高到 2000Hz ，我們的聽覺只能覺察到頻率提高了一點點，根本察覺不到頻率提高了一倍。

Mel Scale:

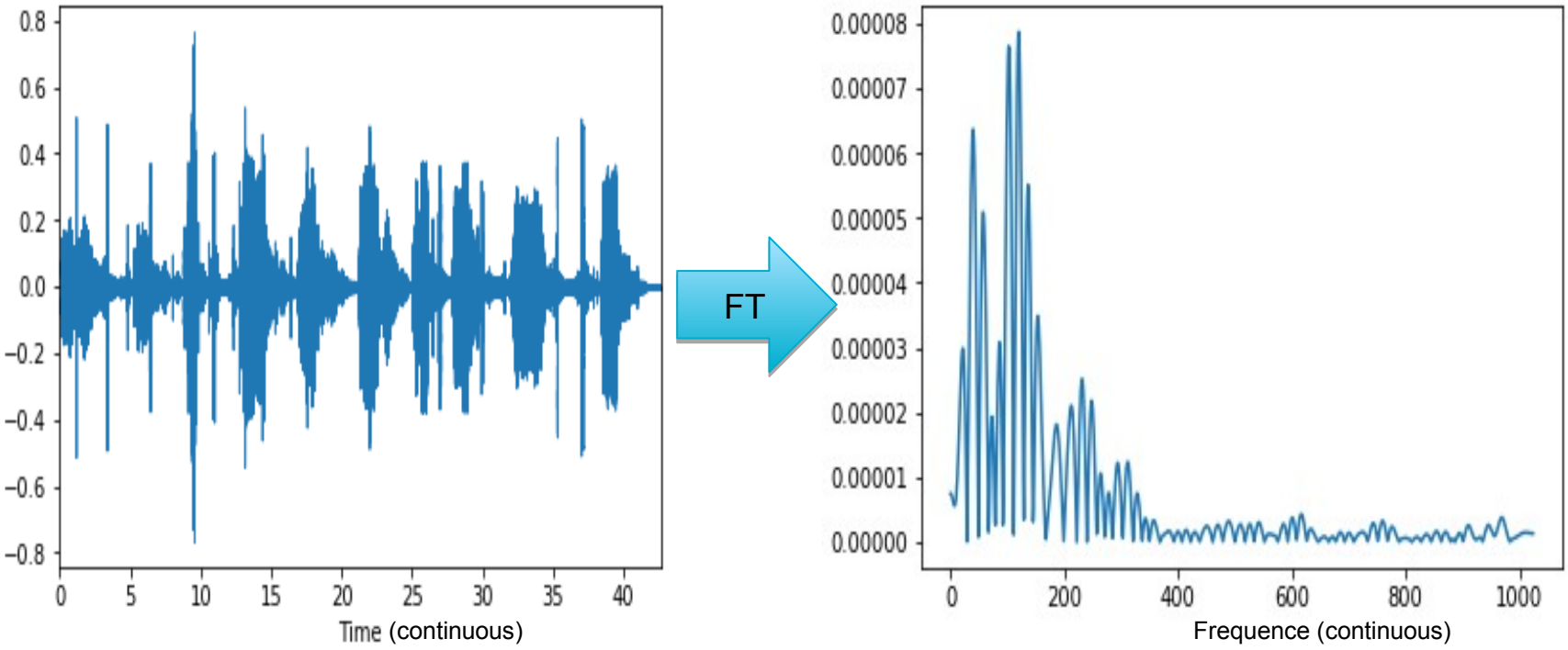
$$mel(f) = 2595 * \log_{10}(1 + f/700)$$

梅爾標度可以反應出人類聽覺感知：
在頻率較低的狀況下，如果兩段語音的梅爾頻率相差兩倍，則人耳可以感知到的音調大概也相差兩倍。在頻率較高的狀況下，人耳就難分辨出聲音差異。

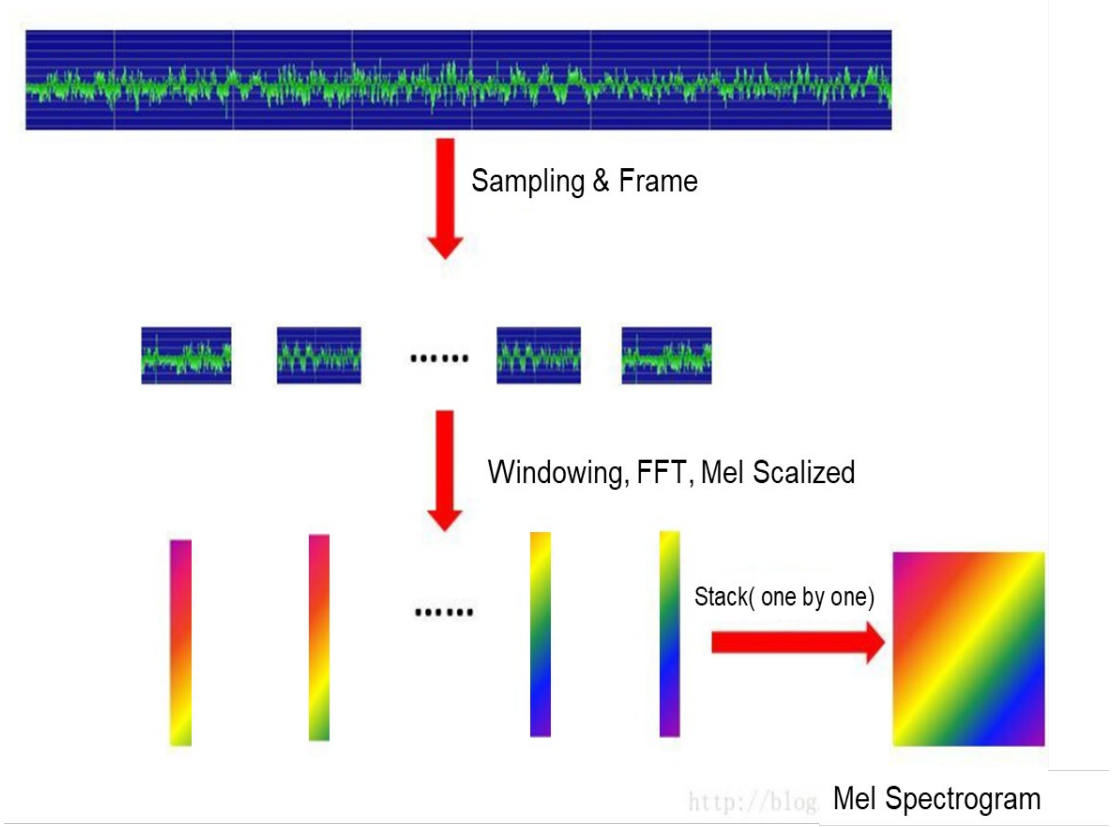
語音取樣與前處理原理研析 (3/8)
類神經網路 (NN) 的運算需求

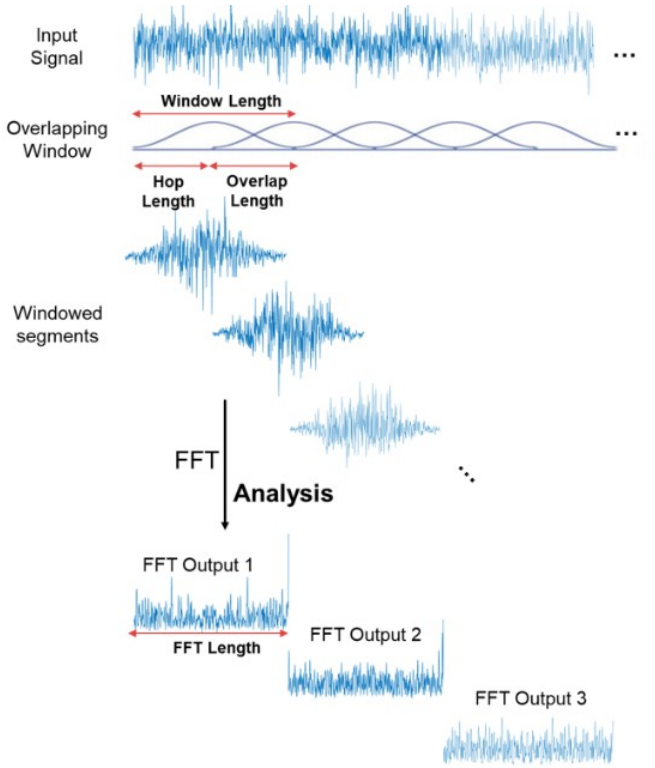
- 對於 NN 而言，是處理離散與有限長度的資料。
- NN 必須以大量的離散數據訓練取得特徵權重。
- 語音特徵隨時間變化具前後關係性。
- Fourier Transform 是連續函數，必須以 Discrete Fourier Transform(DFT) 取得語音的離散數據。
- 語音的離散數據依據樣本取樣定理，對一個有限頻寬之訊號，採樣頻率大於最高頻率的 2 倍，便可重建原來的連續信號，即此取樣率可獲得原訊號之特徵。
- Fast Fourier Transform(FFT) 為 DFT 之特例，可將 DFT 矩陣分解為稀疏矩陣，加快運算速度。

語音取樣與前處理原理研析 (4/8)
語音頻譜轉換 (Fourier Transform)

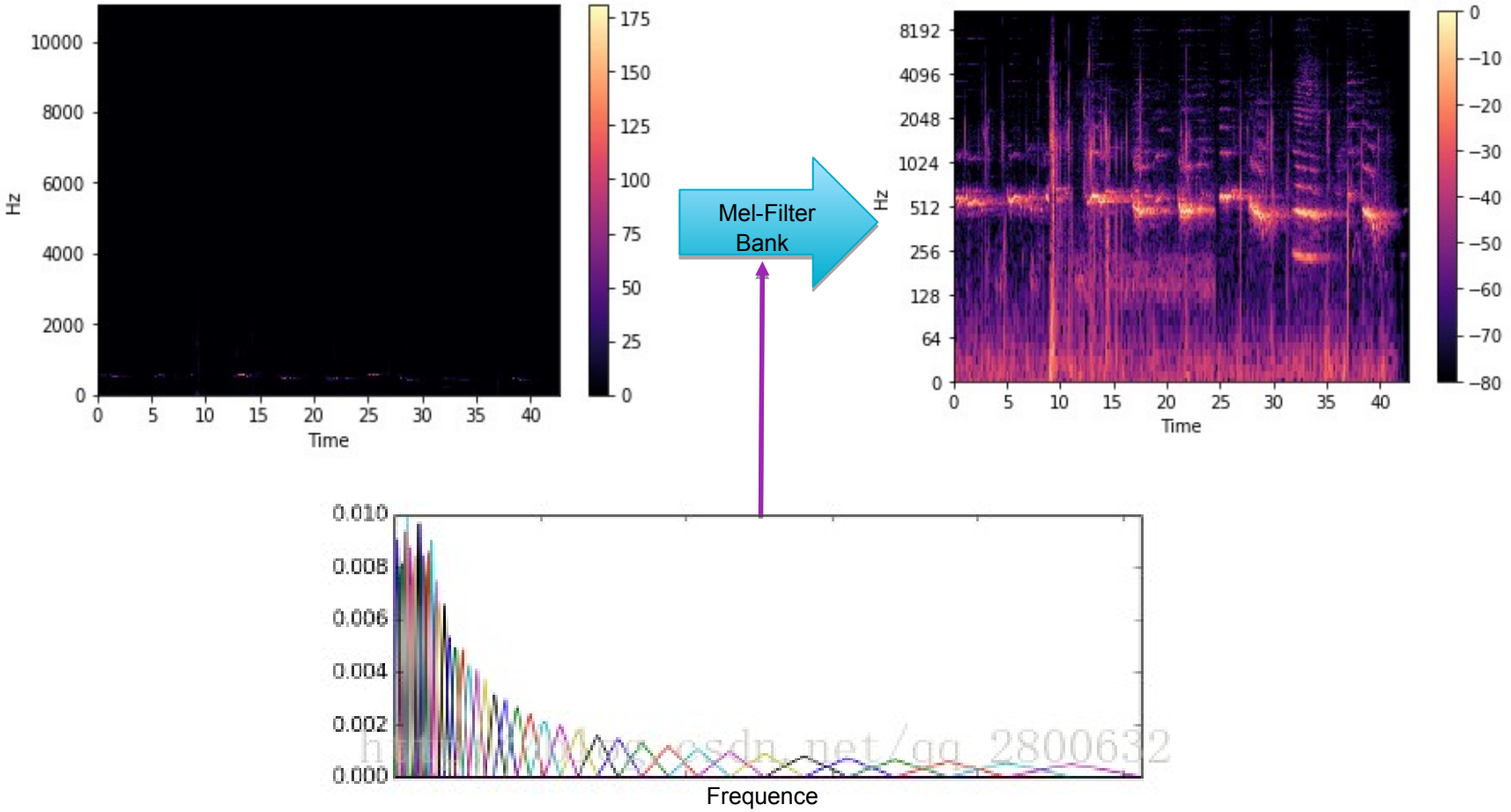


語音取樣與前處理原理研析 (5/8)
語音的特徵抽取步驟 (STFT)



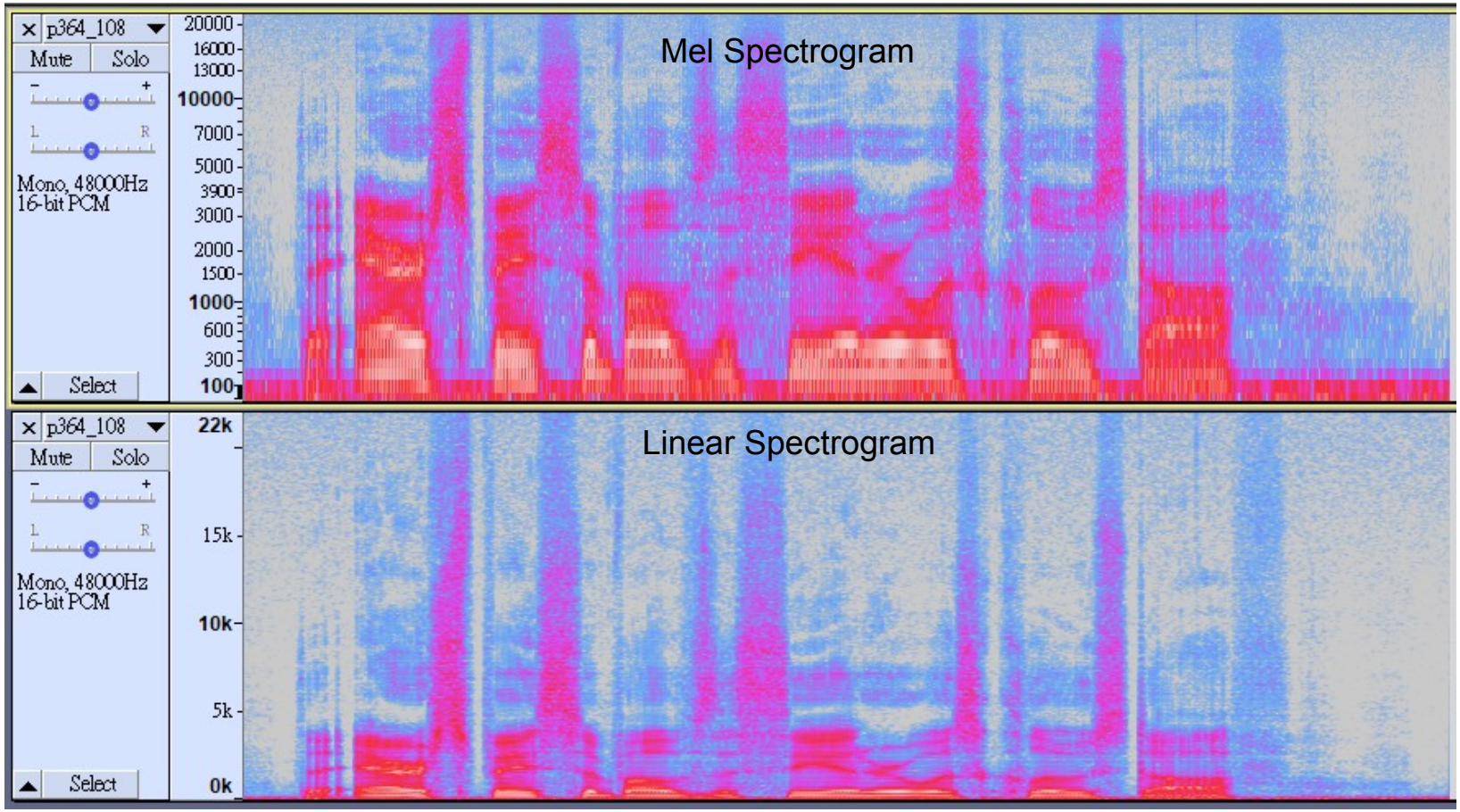


語音取樣與前處理原理研析 (6/8)
Mel Spectrogram with Mel-Filter Bank



語音取樣與前處理原理研析 (6/8)
Mel Spectrogram with Mel-Filter Bank (範例)

VCTK p364_108.wav (Male voice)



語音取樣與前處理原理研析 (7/8)
訓練語音資料來源 (VCTK)

- All speech data was recorded using an identical recording setup: an omni-directional head-mounted microphone (DPA 4035), 96kHz sampling frequency at 24 bits and in a hemi-anechoic chamber of the University of Edinburgh.
- **All recordings were converted into 16 bits, were**

語音取樣與前處理原理研析 (8/8)
語音訊號處理程式庫 - LibROSA

librosa 是一個強大的 python 語音訊號處理的第三方程式庫

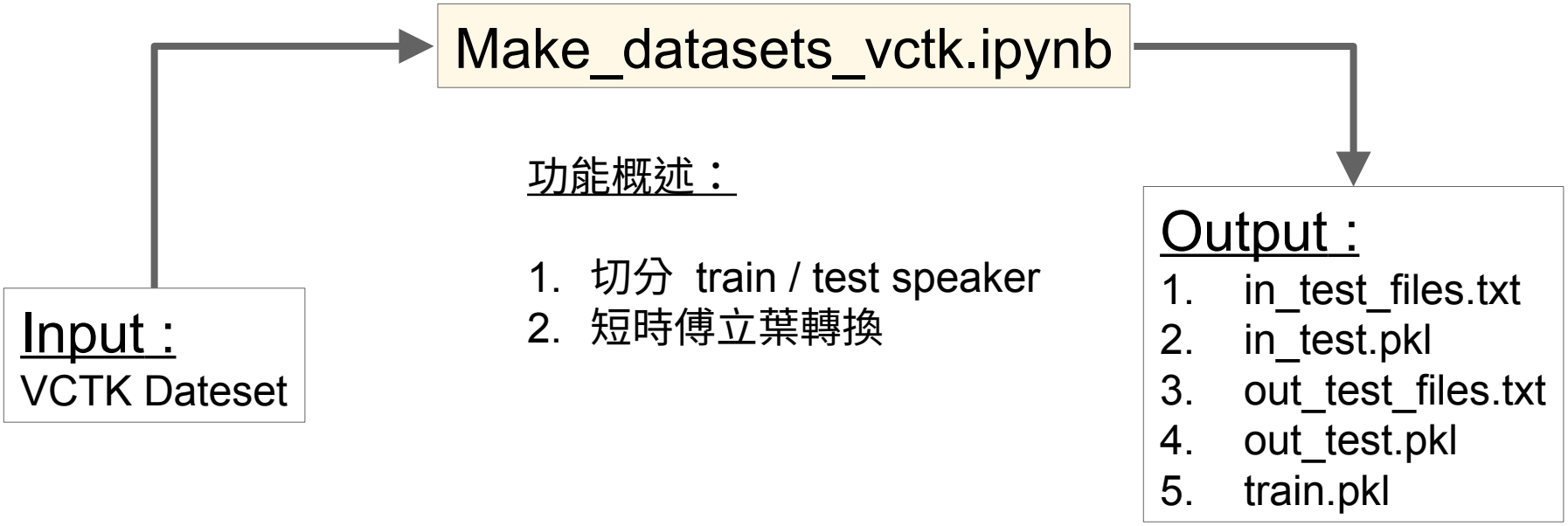
常用的專有參數 :

[y ：音訊時間序列] 、 [sr ：取樣率] 、 [hop_length ：幀移] 、 [overlapping ：連續幀之間的重疊部分] 、 [n_fft ：視窗大小] 、 [spectrum ：頻譜] 、 [spectrogram ：頻譜圖或叫做語譜圖]

- **librosa.load**(path, sr=22050, mono=True, offset=0.0, duration=None)
- **librosa.resample**(y, orig_sr, target_sr, fix=True, scale=False)
- **librosa.stft**(whale_song[:n_fft], n_fft=n_fft, hop_length=n_fft+1)
- **librosa.filters.mel**(sr=sr, n_fft=n_fft, n_mels=n_mels)
- **librosa.output.write_wav**(path, y, sr, norm=False)

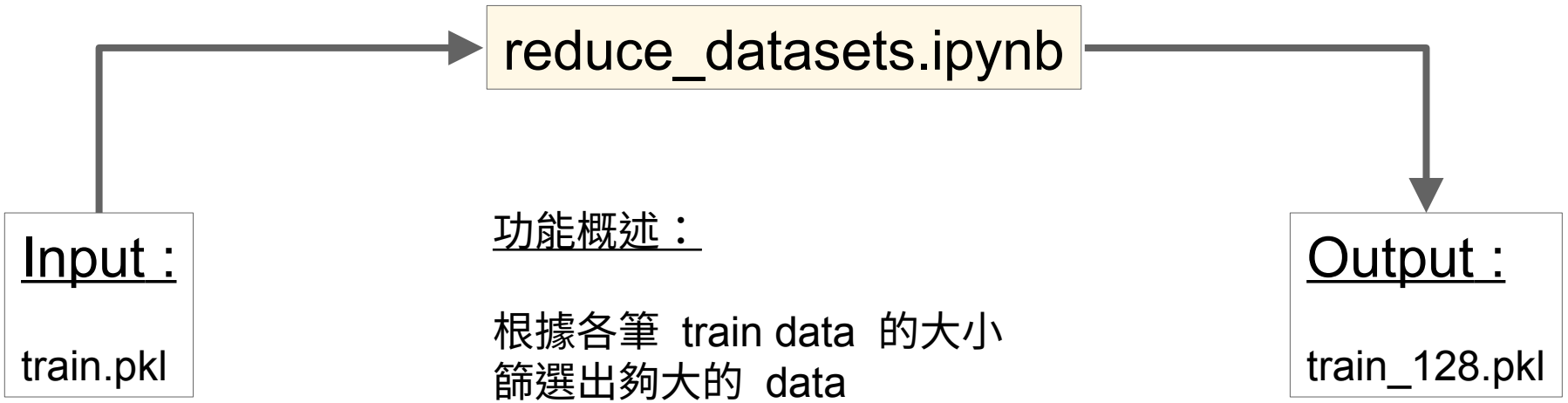
VC 系統之資料處理、模式建立、訓練與驗證 (1/5) Preprocess

STEP 1



VC 系統之資料處理、模式建立、訓練與驗證 (1/5) Preprocess

STEP 2



VC 系統之資料處理、模式建立、訓練與驗證 (1/5) Preprocess

STEP 3

sample_single_segments.ipynb

Input :

1. train_128.pkl
2. in_test.pkl
3. out_test.pkl

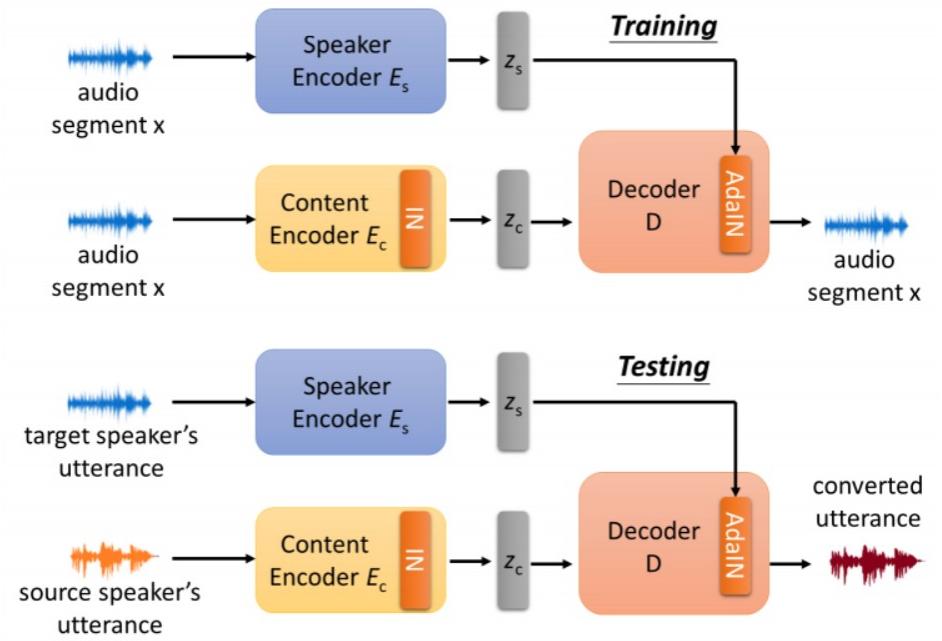
功能概述：

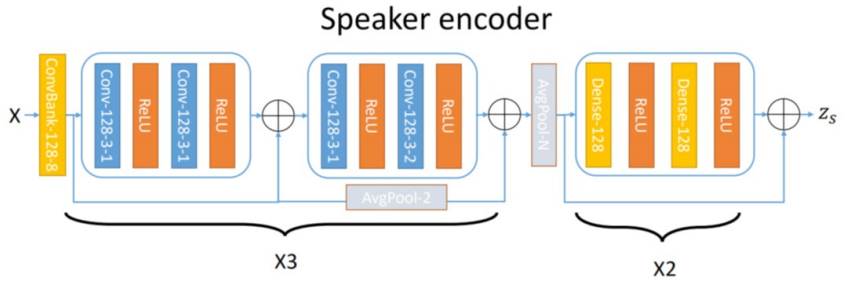
1. Indexing data
2. Data save as .json

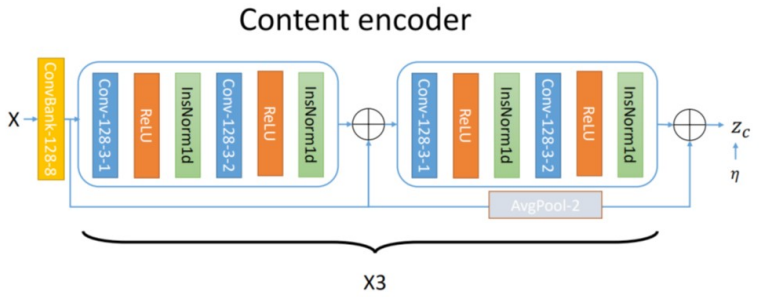
Output :

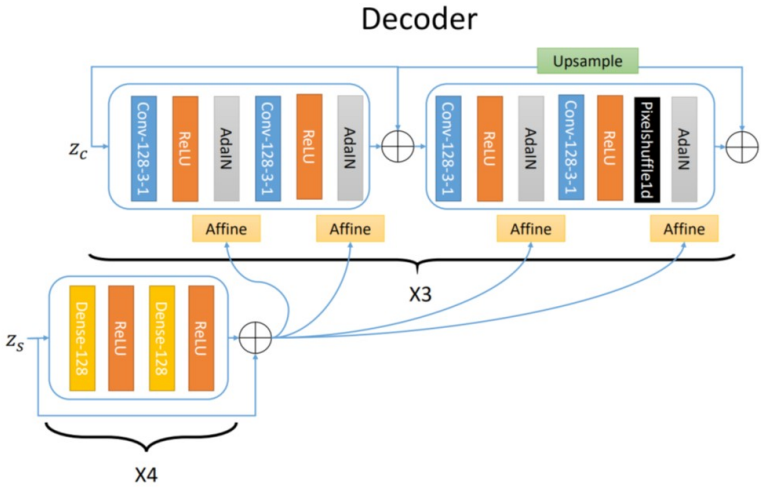
train_samples_128.json
in_test_samples_128.json
out_test_samples_128.json

VC 系統之資料處理、模式建立、訓練與驗證 (2/5) Modeling

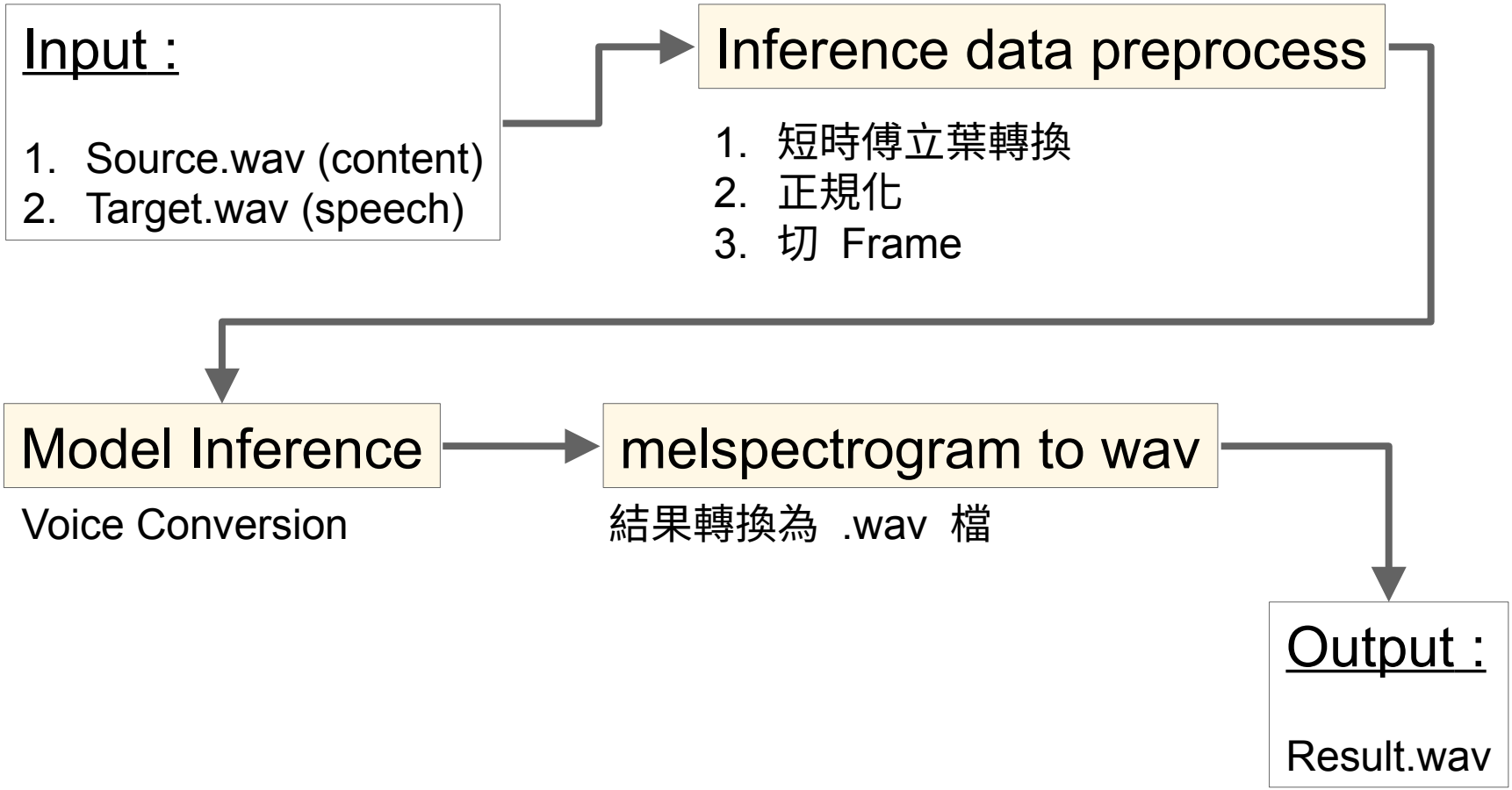






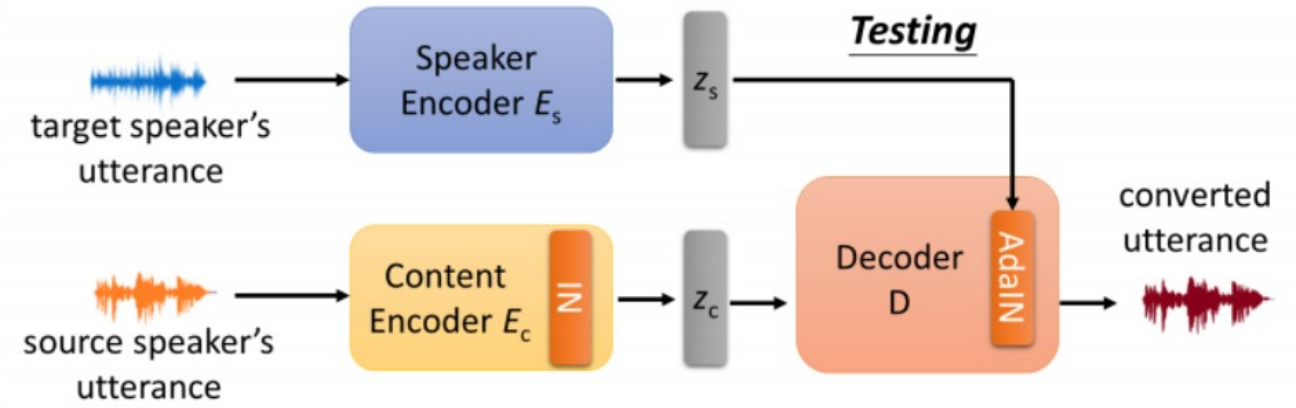


VC 系統之資料處理、模式建立、訓練與驗證 (4/5) Inference



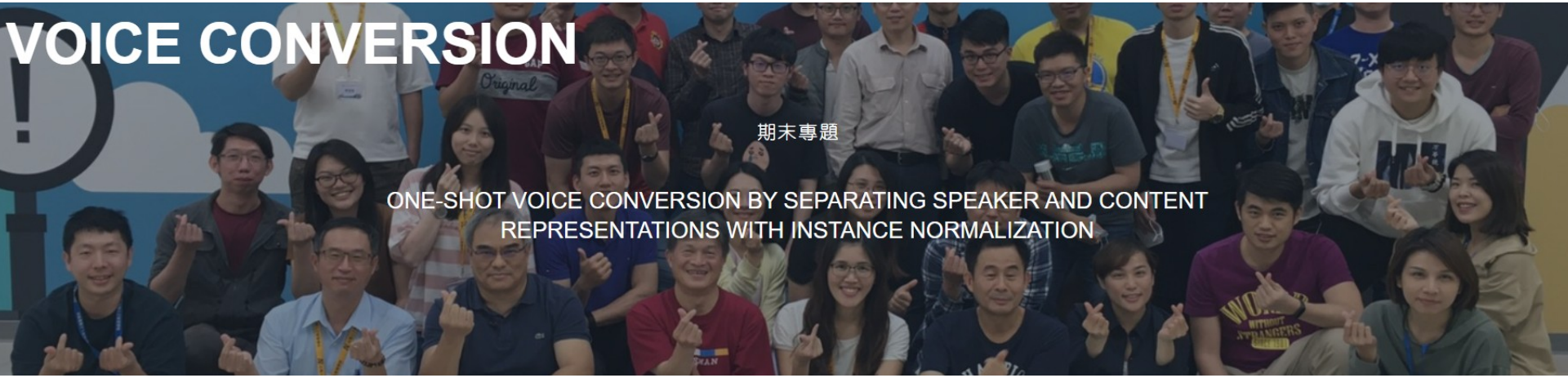
VC 系統之資料處理、模式建立、訓練與驗證 (5/5) Valification

Target Voice
By ET083023
陳義翔



Source Voice
from VCTK
P227_013

VC 系統網頁展示界面


VOICE CONVERSION
期末專題
ONE-SHOT VOICE CONVERSION BY SEPARATING SPEAKER AND CONTENT
REPRESENTATIONS WITH INSTANCE NORMALIZATION

建議後續研究議題

1. Multi langurage problem (口音)
2. VC 後語音問題 (平滑度、噪音)