DLHLP HW2
Voice Conversion

2020.08.23

負責助教：姜成翰

What is Voice Conversion (VC) ?

https://motoneta.fandom.com/zh/wiki/%E8%9D%B4%E8%9D%B6%E7%B5%90%E5%9E%8B%E8%AE%8A%E8%81%B2%E5%99%A8

What is Voice Conversion

Source speaker

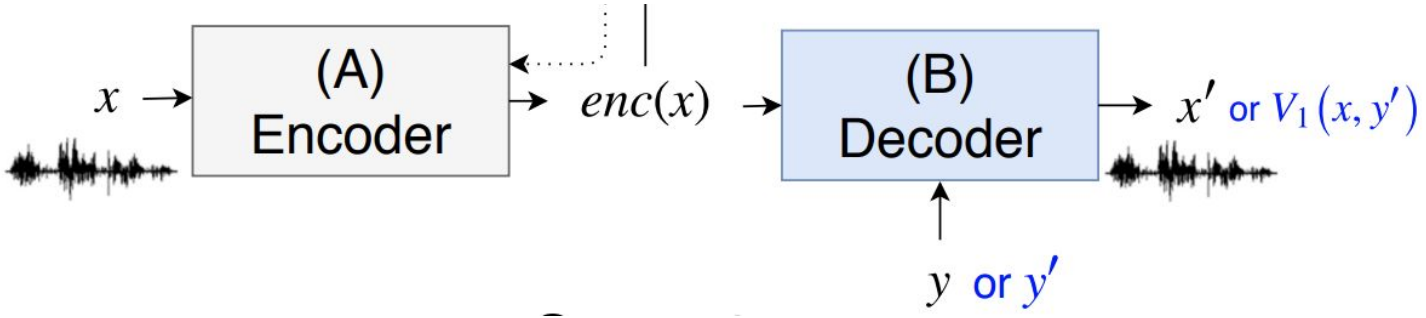
Target speaker

Utterance of speaker A’s sound

VC model

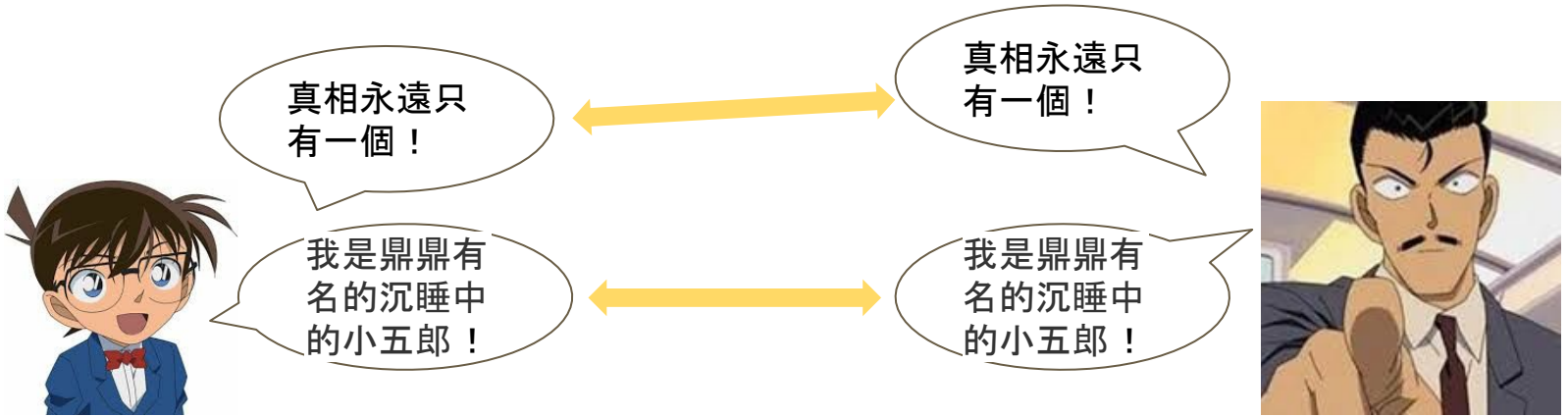
Utterance of speaker B’s sound

VC in Real World: Auto-Encoder



VC in Real World: Data

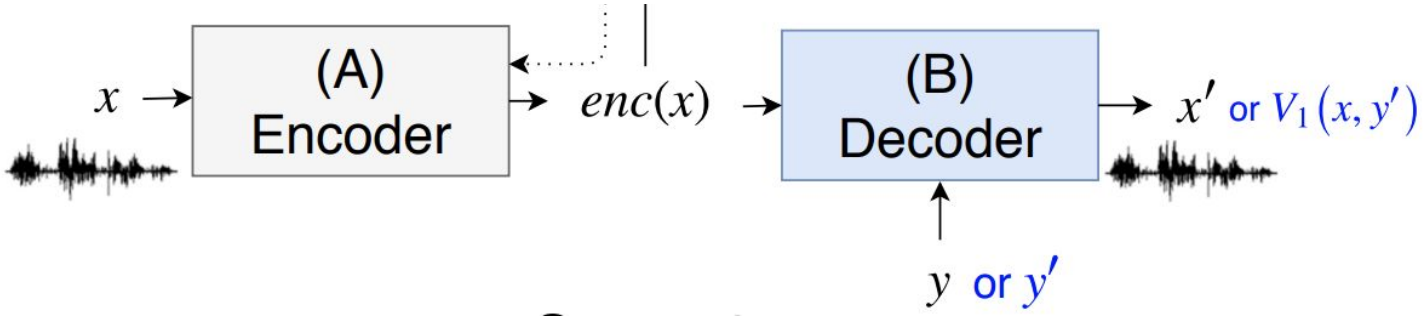
1. Parallel data: VCTK



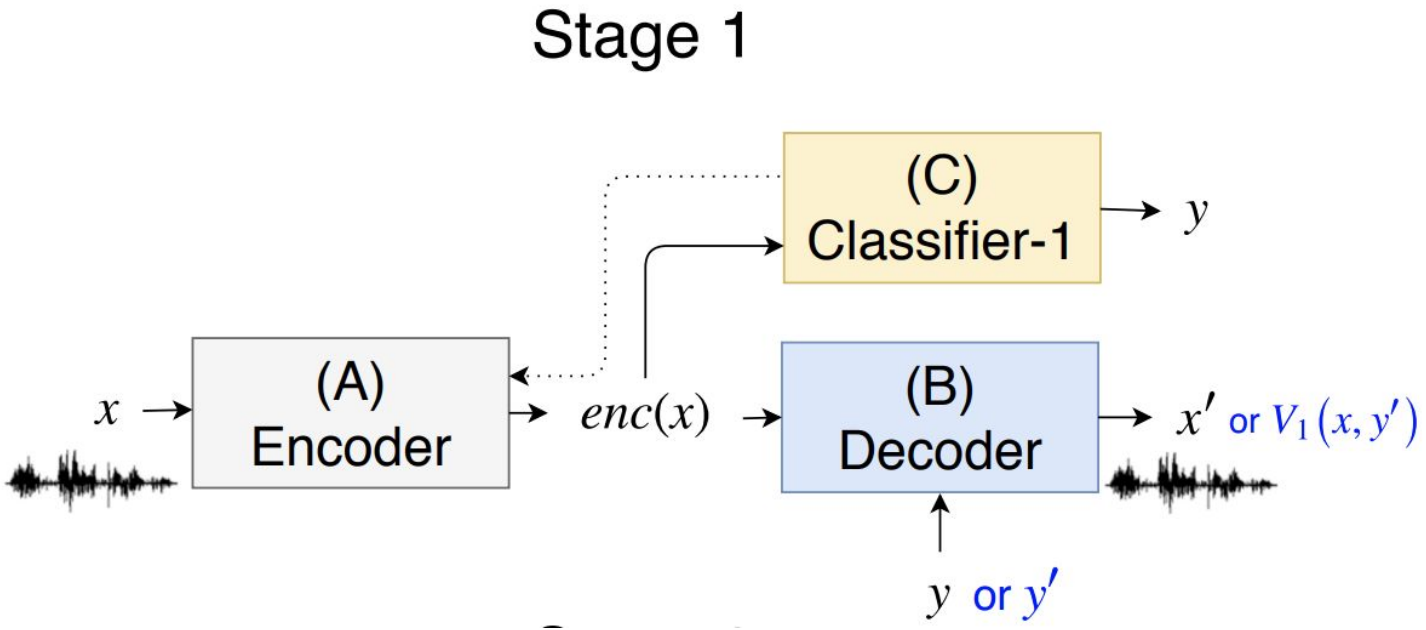
VC in Real World: Data

2. Non-Parallel data: Can be obtained everywhere

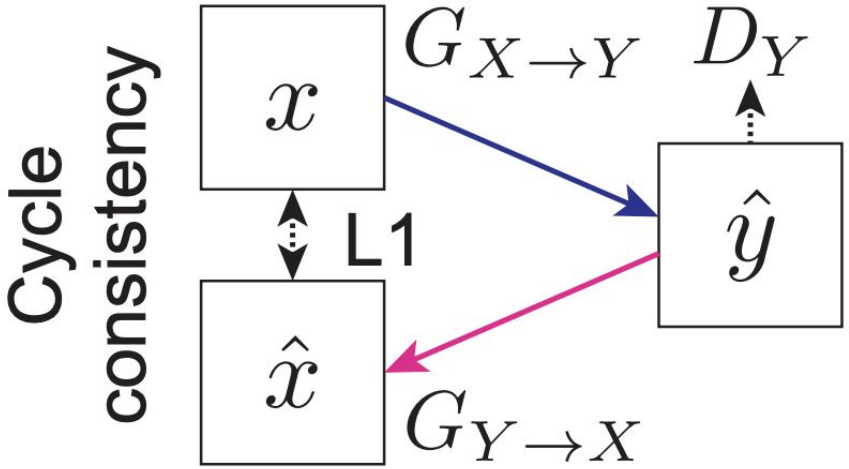
VC in Real World: Auto-Encoder

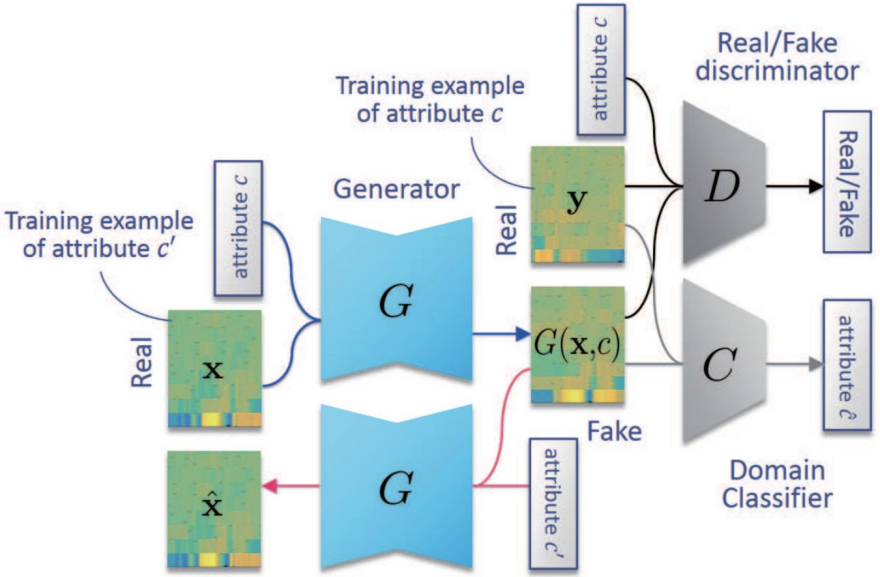


VC in Real World: Adversarial Feature Disentanglement



VC in Real World: CycleGAN and StarGAN-VC





https://www.eurasip.org/Proceedings/Eusipco/Eusipco2018/papers/1570438014.pdf

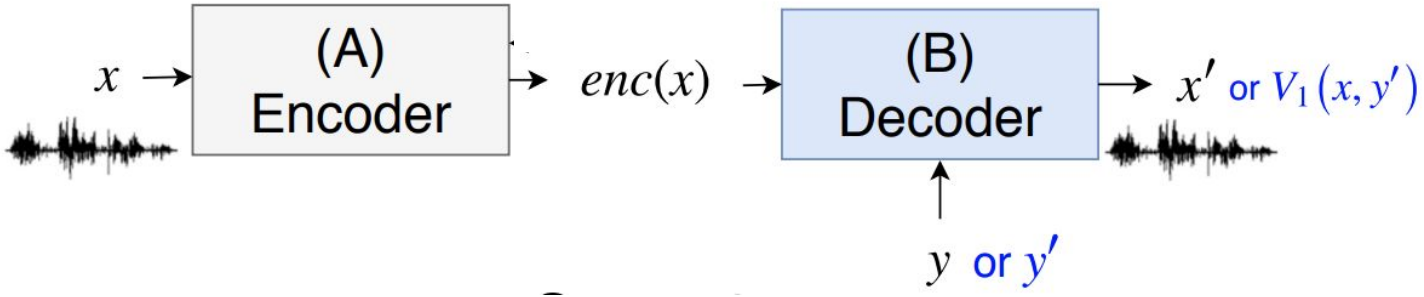
https://arxiv.org/pdf/1806.02169.pdf

VC: Evaluation

Human evaluation

Report 連結

HW 2-1: Auto-Encoder (4.5%)



Multi-target Voice Conversion without Parallel Data by Adversarially Learning Disentangled Audio Representations

HW 2-1: Auto-Encoder (4.5%)

(2%) :助教 human evaluation

(2.5%) :Report
(1) 請以 Auto-Encoder 之方法實做 Voice conversion。如果同學不想重新刻一個 auto-encoder，可以試著利用這個repo的部分程式碼，達到實現出 auto-encoder。如果你是修改助教提供的 repo，請在 report 當中敘述你是如何更改原本程式碼，建議可以附上修改部分的截圖以利助教批閱；同時，如果各位有更動原本模型參數也請一併列出。如果你的 auto-encoder是自己刻的，那也請你簡單敘述你的實作方法，並附上對應程式碼的截圖。(1%)
。hint: 大約跑 100000 個 steps 即可有不錯的結果

HW 2-1: Auto-Encoder (4.5%)

(2.5%) :Report
(2) 在訓練完成後，試著將助教要求轉換的音檔轉成 source speaker 和 target speaker 的 interpolation，也就是在 testing 的時候，除了將指定的音檔轉成 p1 和 p2 的聲音之外，請嘗試轉成p1 和 p2 interpolation 的聲音。並比較分析 interpolated 的聲音和 p1 以及 p2 的關係。你可以從聲音頻率的高低、口音、語調等面向進行觀察。只要有合理分析助教就會給分。請同時將這題的音檔放在github 的 hw2-1資料夾中，檔名格式請參考投影片。(1.5%)

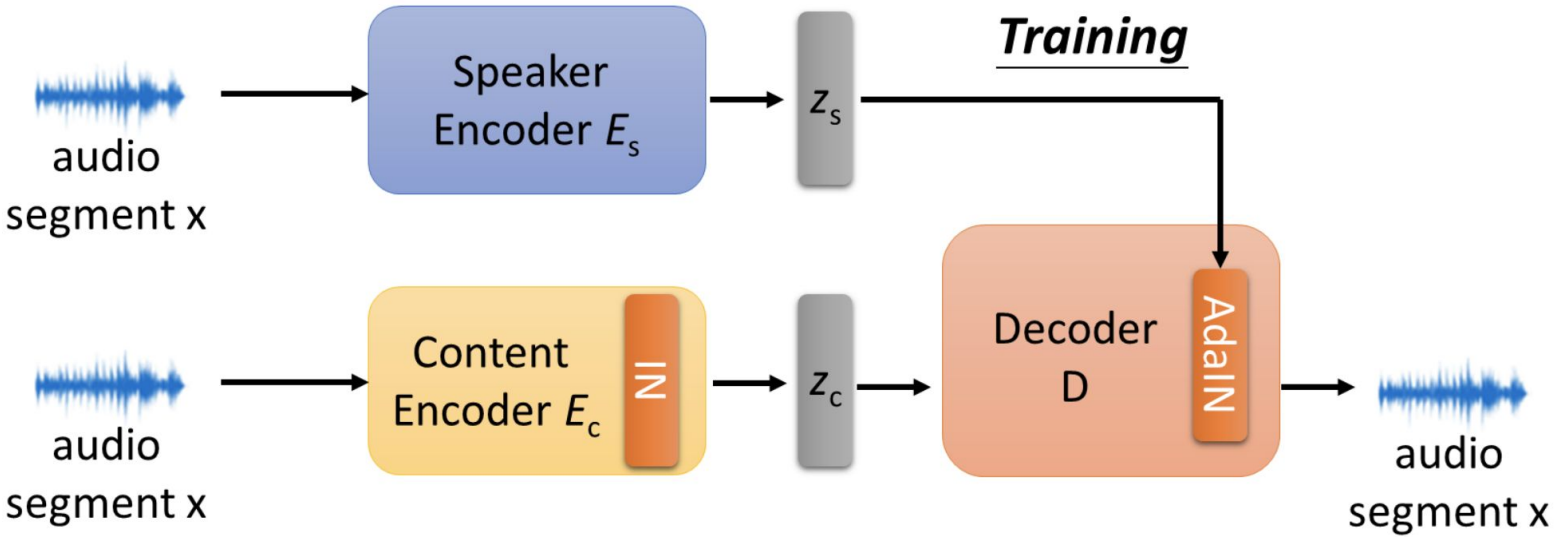
https://arxiv.org/pdf/1904.05742.pdf

https://github.com/jjery2243542/adaptive_voice_conversion

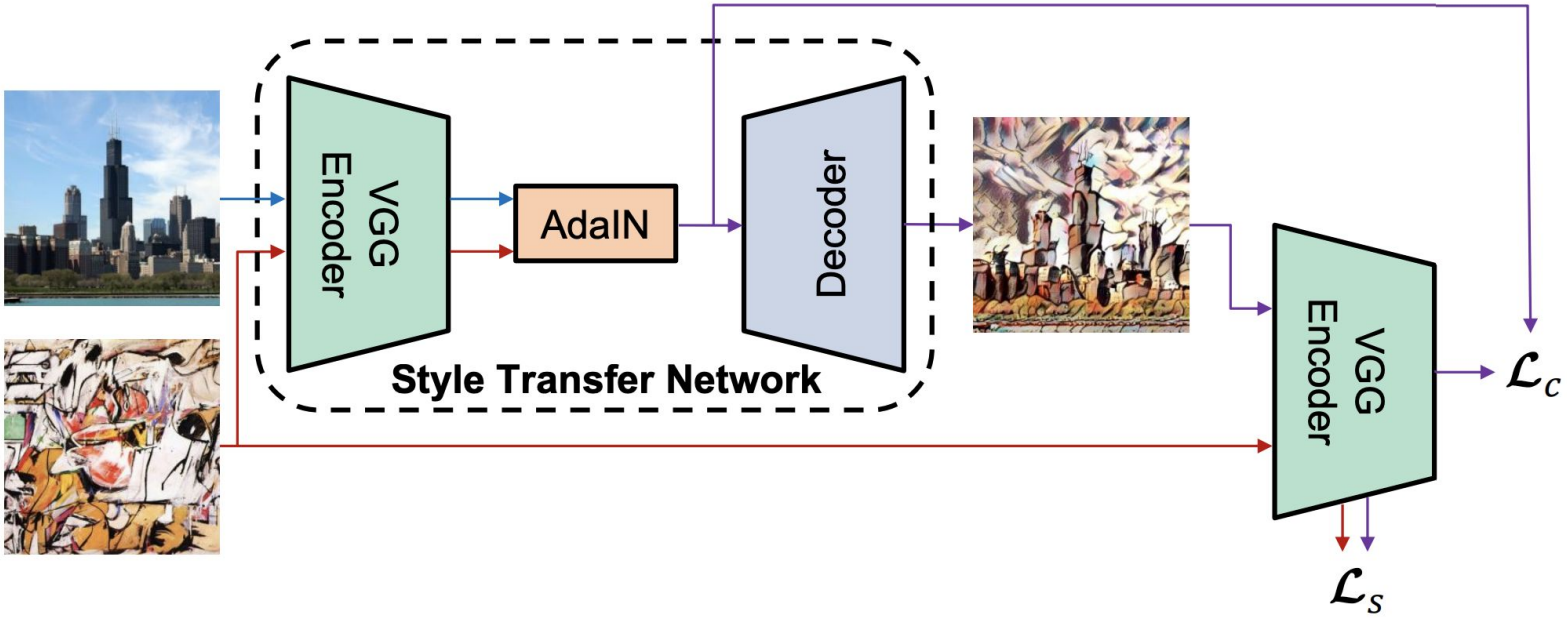
preprocessed data:
https://drive.google.com/file/d/1m98NCKvM9u5D_IJNnW_mopJRWvnl2lhf/view?usp=sharing

HW2-1: One-shot VC

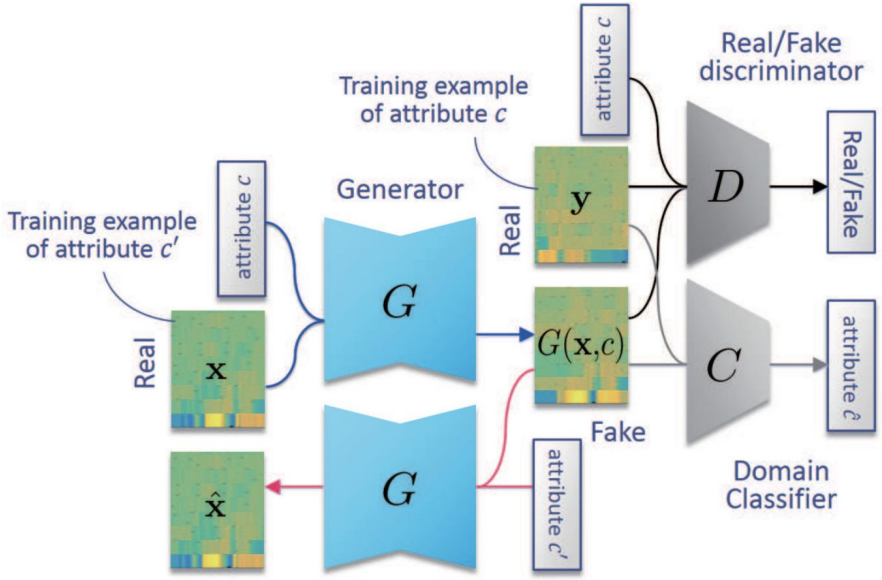
Use instance normalization to disentangle content from speaker information



HW2-1: AdaIN for style transfer



HW2-2: GAN (4.5%)


Training example of attribute c
attribute c
Real/Fake discriminator
Generator
Training example of attribute c'
attribute c
Real
$\mathbf{y}$
D
Real/Fake
G
Real
$\mathbf{x}$
$G(\mathbf{x},c)$
C
attribute $\hat{c}$
Fake
G
$\hat{\mathbf{x}}$
attribute c'
Domain Classifier

https://github.com/hujinsen/pytorch-StarGAN-VC/issues/8

HW2-2: GAN (4.5%)

(2%) :助教 human evaluation

(2.5%) :Report

Report 問題：請描述在助教提供的這個 repo (pytorch-StarGAN-VC) 裡面，speaker embedding 是怎麼放到 generator 裡面的呢？請問這樣的做法會有什麼優缺點？(像是會不會影響 model 參數量？假設我今天要增加 speaker，那我的 embedding table 會不會有什麼變化？)另外請注意這個 repo 裡面是 4 個 speaker 互轉的 code，但是老師跟助教擔心這樣同學要花很多時間才會 train 到收斂，所以**只要轉 2 個 speaker 就好**，所以請把程式改成只轉兩個 speaker 的 code，並在 report 中描述你怎麼改的。最後，請問這個 model 使用的 input acoustic feature 是什麼？generator 的 output 又是什麼？

HW2-3 (6%)

Option 1. 自己找一個不是 StarGAN-VC，也不是 HW2-1 的 model，實際 train 看看。

Hint: 用 cVAE 做 VC、用 Flow 做 VC、用 auto-encoder 做 VC (useful link)

Option 2. 想辦法 improve HW2-1或是 HW2-2 的 model (或是改一些有趣的東西)。

Hint: 各位可以想想看 speaker embedding 有沒有什麼其他方式？如果今天我在 testing 的時候想要讓他有 unseen speaker 也可以成功轉過去的話，用什麼 embedding會比較好？(hint: d-vector, i-vector) 又或者要怎麼把這個 speaker embedding 餵進 model 裡面呢？有什麼不同的方法？

HW2-3 (6%)

(2%) :助教 human evaluation

(4%) :Report

Report 請詳述實作的內容，若選擇 Option 1 請分析比較 model 的差別，選擇 Option 2 請說明實作的 improvement 及效果

你如果只是改 batch size 或是調調 learning rate，我們是不會給你任何分數的。

分享大會 (Bonus, up to 2%)

1. 要做簡單的投影片，投影片請於分享大會一週前交給助教 (1%)
2. 請描述自己在 HW2-3 做的內容，並且現場 demo 音檔
3. 其他組的同學會在這時候做 human evaluation (1%)

Bonus部分：只要有做投影片介紹自己在 hw2-3做的事情，就會拿到 (1%) 的 bonus，另外 human evaluation 的 bonus 部分，會透過其他組同學的評分，排名前三的組別會另外拿到 1% bounis

時間：暫定 5/6

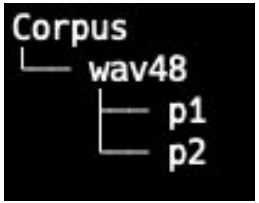
注意事項

- 為了讓各位可以輕鬆做出 Voice Conversion 的結果，我們用的 dataset 只有兩個語者，一男一女
- HW2-1~3 每題都要交指定的音檔，音檔基本上我們只要聽到男的有轉成女的，女的轉成男的，就可以拿到助教 human evaluation 的 6%
- 因為基本上衡量 improvement 的結果就是助教自己聽，但是 VC 的結果基本上聽起來都差不多，所以各位報告的 HW2-3 要寫清楚自己做了哪些 improvement，要是你在報告分沒有寫 improvement 是什麼，那就算你有繳交音檔，助教也不會給你分數

Dataset:

HW2-1: data連結

解壓縮之後會是如右圖這樣的檔案



使用 p1 及 p2 兩個 speakers 做 training

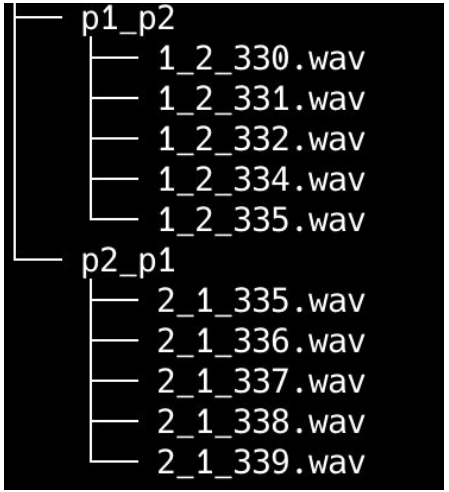
Follow README instructions to finish preprocessing data and training !!!

You do NOT need to split the data into testing sets since the provided preprocessing code will do it for you.

You might need to modify some specific path in the provided code.

音檔繳交

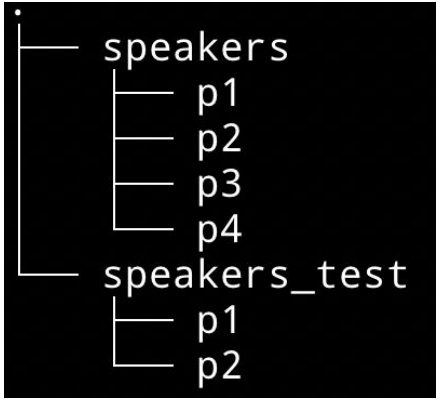
HW2-1
請使用提供的 code 轉換音檔 hint: python convert.py
產生的結果應為 p1 轉 p2 音檔及 p2 轉 p1 音檔，如右圖(數量可調整)
同學僅需繳交 1_2_334.wav 及 2_1_338.wav
同時必須繳交 interpolation 的結果，同樣為上述兩個音檔
請註明清楚何者為 interpolation 的結果
(ex. 1_2_334_inter.wav)
共四個音檔



Dataset

HW2-2: data連結

解壓縮之後會是如右圖這樣的檔案



speakers 是 training data，**只要用 p1 跟 p2** train 就好，請不要用 p3 p4 訓練。只是為了讓同學可以玩玩看 2 個以上 speakers 的 VC 所以才給另外兩個語者。

speakers_test 是 testing data，請在 p1 和 p2 中**選擇指定的檔案**轉成另一個語者的聲音。

Dataset:

HW2-2

右圖是 speaker_test 裡面的檔案。每組到時候會繳交指定的檔案到 github上面。指定檔案的方式如下：

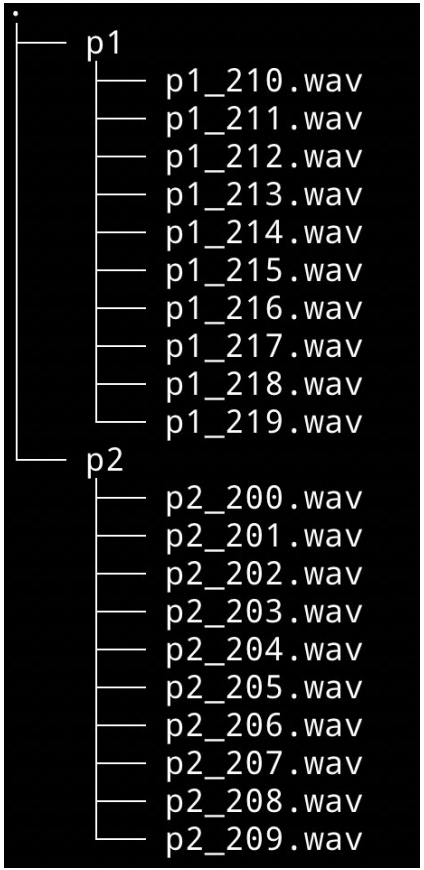
if (組長學號 % 10) == 0

then 把 p1_210.wav 轉成 p2 的聲音，把 p2_200.wav轉成 p1的聲音

elif (組長學號 % 10) == 1

then 把 p1_211.wav 轉成 p2 的聲音，把 p2_201.wav轉成 p1的聲音

依此類推



作業繳交

你的 github 上面應該有這樣的東西

1. report.pdf: 就是你 hw2-1~hw2-3 的 report 回答
2. wav 這個資料夾，裡面會有 3 個資料夾，分別命名為 hw2-1, hw2-2, hw2-3。每個資料夾裡面都要有助教指定的 voice conversion 結果。HW2-1 音檔格式請參照投影片 20 頁；HW2-2 音檔請命名成 p1_to_p2.wav 和 p2_to_p1.wav

Deadline

2020.08.30, 2020.09.06