**Cloud Computing and Big Data Analytics**

**Term Project Proposal**

**Detective Conan’s Voice Converter**

**0760222黃昱銘、0760241徐傳恩、0750724王晟凱**

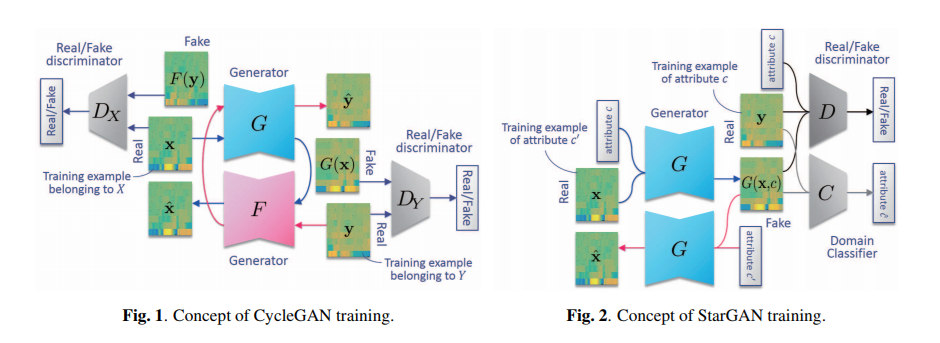


1. **Idea**

In the cartoon “名偵探柯南”, we can always see Conan hide in the Dark and use the voice converter to mimic the voice of the Detective 小五郎 to solve the homicide cases.

We aim to create such an application that enables everyone to change the voice to another person’s tone.

1. **Technique**

****

* 1. **one to one conversion - CycleGAN**

We can translate one person’s sound into another person restricted by the source and target training data.

* 1. **many to many conversion - *StarGAN*** (Optional)

Extension of CycleGAN which is able to simultaneously learn many-to-many mappings across different attribute domains using a single generator network

* 1. **Using non-parallel data**

While training, we aim at using non-parallel data, which means we don't need two persons speaking the same set of sentences. Parallel data are only used to evaluate our model performance.

1. **Uniqueness or the comparisons with state-of-the-art**
   1. Add our favorite actor’s voice or our voice into target data set
   2. (Optional) Combine the model into a real-time system, such that when a person speaks into the microphone, the speaker instantly outputs the target person's voice
2. **Dataset**
   1. Voice Conversion Challenge (VCC) 2018 dataset
   2. Voice Conversion Challenge (VCC) 2016 dataset
   3. (Optional) Our voice, teacher’s voice, celebrities’ voice (Detective Conan, weekly addresses by Trump)
3. **Reference**

# *[1] Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks* - <https://arxiv.org/abs/1711.11293>

# *[2] StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks -* [*https://arxiv.org/pdf/1806.02169.pdf*](https://arxiv.org/pdf/1806.02169.pdf)

*[3] CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion -* <https://arxiv.org/abs/1904.04631>

**Cloud Computing and Big Data Analysis**

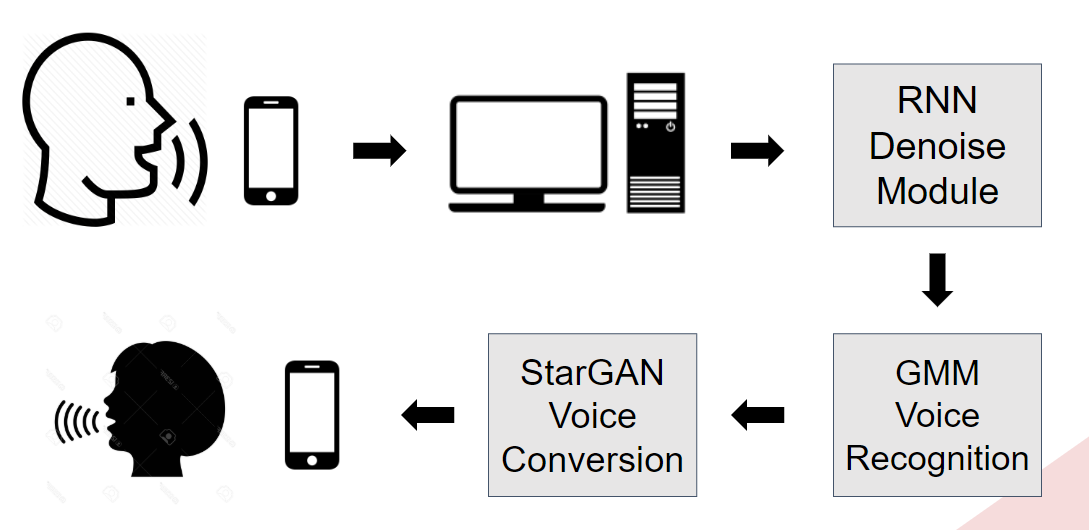
**Final Project Report**

**Detective Conan’s Voice Conversion APP**

1. **Motivation**

本次期末專題受知名動畫”名偵探柯南”啟發，決定製作一變聲器程式，並整合至iOS的App。變聲器目的為在保留語音內容的同時，改變說話者的聲音，語音內容不限語言，並期望達到即時的聲音轉換。我們採用了StarGAN聲音轉換深度學習模型，除了能達成多對多(many-to-many)的聲音轉換之外，訓練時不需要平行(parallel)聲音資料，且使用時聲音轉換時間也壓在1秒左右，達成即時性的要求。除此之外，我們加入了除雜訊功能和聲音辨識功能，讓整個系統更方便使用者使用，效果也更好。聲音轉換的結果雖然還是包含部分失真，但仍能輕易分辨出目標的聲音特徵。

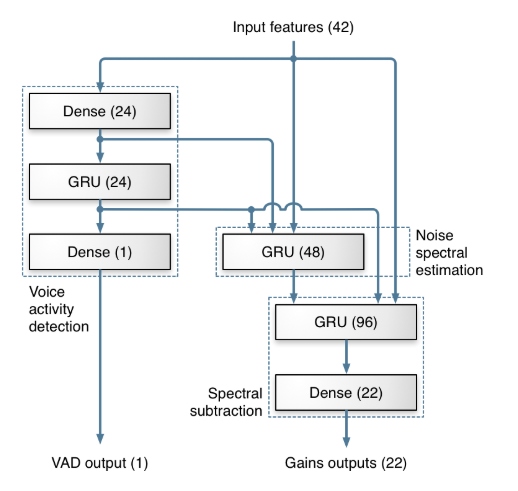
1. **System Architecture**

****

上圖為我們系統的整體架構，使用者使用iPhone下載App後，利用App錄好聲音並選擇目標，因受限於手機的運算能力不足，App將會把資料傳送至負責聲音轉換的電腦，包含除雜訊、聲音辨識、聲音轉換的功能都在電腦中處理，處理完畢後將檔案傳回手機，使用者即可播放並檢查效果。

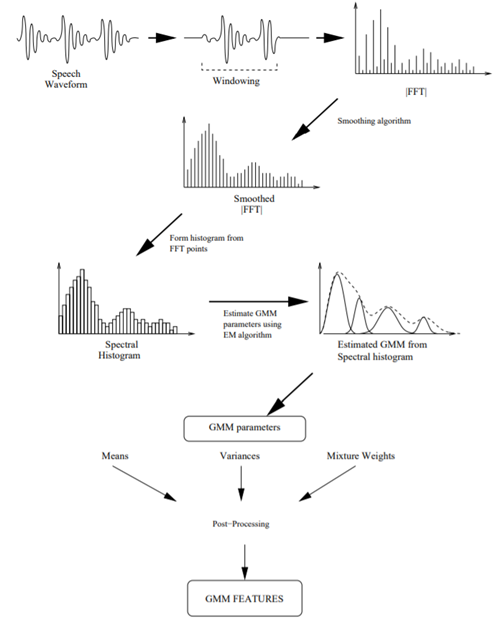
1. **RNN noise suppression**

在聲音轉換前先除雜訊，目的為希望去除雜訊能降低利用不同裝置錄音產生的影響，並提升StarGAN的效果。此部分利用的模型為RNNoise，模型架構圖如下。



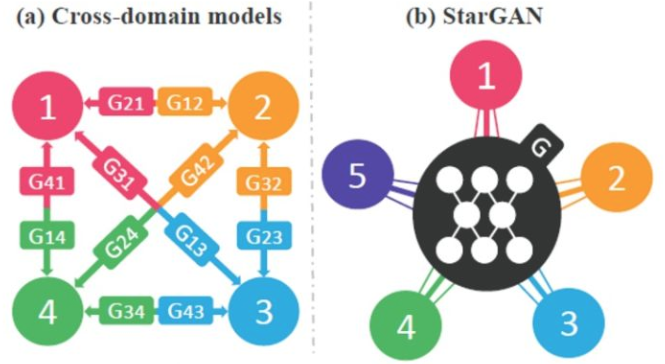
此模型的特別之處在於，輸入並非是純粹的聲音訊號，而是經過傳統DSP除雜訊的前處理，例如DFT轉換、依照頻譜作分割等等，再輸入模型，相比直接輸入純聲音訊號、一切交由模型處理，此方法的模型更小、所需的運算資源也更少，甚至不需要用到GPU，因此轉換的速度也非常快，只要人的聲音夠大，除雜訊的效果不錯。

1. **GMM voice recognition**

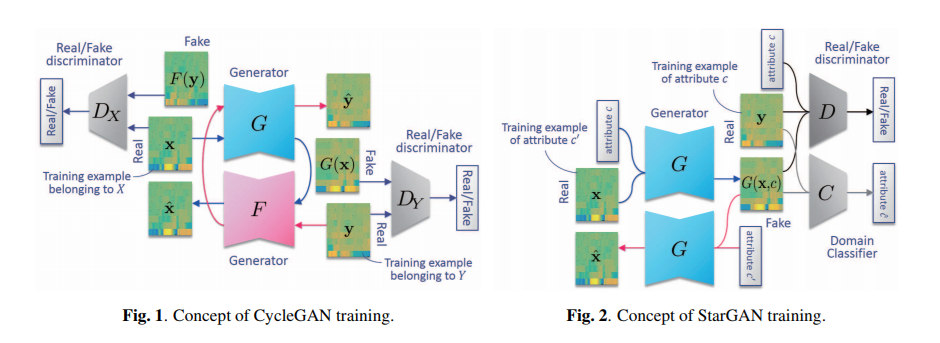
GMM利用多個Gaussian Distribution去擬合語音的頻率分布。透過切割語音再進行頻率轉換來擷取語者的特徵。並以此來判斷是誰在講話。因StarGAN在轉換聲音時，需要知道來源聲音為群組中的哪一個聲音，在我們的系統中利用GMM model的方式自動辨識錄音者的聲音，即可為使用者免去手動選擇的麻煩。

1. **StarGAN**

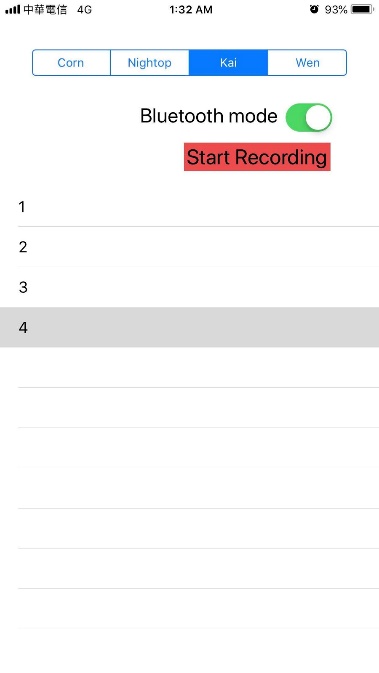
StarGAN是CycleGAN的進化版。StarGAN可以在多domain之間進行轉換，若以CycleGAN來講，4種domain就會需要12種model來互相轉換(下圖a)，而StarGAN卻只需要一個Generator就可以做到所有domain的轉換(下圖b)

****

傳統上StarGAN被用來處理圖片，我們將聲音當成圖片的一種，針對每個音檔經由python的pyworld package downsampled到16 kHz，並取出24 Mel-cepstral coefficients (MCEPs)、logarithmic fundamental frequency (log F0)、每五秒一次的aperiodicities (APs)，最後輸出數個npz, npz檔還有各講者語音的統計特徵作為starGAN的training data。StarGAN除了音檔外，還要input一個target domain的參數，最後透過Generator和Discriminator 之間互相競爭來train出好的model。Discriminator除了分辨輸出音檔是真是假外，還要能夠分辨domain的正確性(Domain classifier)。

最終我們分別錄三名男生以及一名女生的聲音，每個人各162則短句(3-10sec)作為我們的training data set。經過RTX2070兩天的training得到我們的model。****

1. **iOS App、Server**

手機的部分以iphone7plus(xcode 9.0)作為開發環境，在使用者錄音結束後，將整份音檔以HTTP post(Multipart/form)的形式夾帶選擇的target speaker傳送至Server端，Server經過轉檔、去雜音、辨識來源後，由StarGAN輸出轉換到target speaker後的音檔，最後傳送回APP端播放。值得一提的是，我們training的dataset中，其中一男一女的聲音是由此iphone7plus錄製，另外兩名分別由不同android手機錄製。實測結果，由同樣手機錄製的performance會比其他來的好一些，且錄音環境不可以同時有第二個人在說話。根據當前網路環境，大約從錄完音到得到結果撥放大概需要3-10秒的時間。

1. **Results**

因錄音檔案無法直接貼上word或pdf文件，因此將聲音轉換結果附在壓縮檔中，TM1-SF1代表為男性轉女性，雖然結果包含部分雜訊和失真，但仍可聽出部分說話內容和明顯為女性的聲音，SF1-TM1則為女性轉男性，這部分效果差強人意，但聲音明顯為男性，只是說話內容失真較大，已聽不太初原本內容。

1. **Teamwork assignment**

黃昱銘：StarGAN build up、server、整合所有程式碼至server、ios app internet connection、view、報告製作

王晟凱：RNN noise suppression、研究StarGAN、影片製作、報告製作

徐傳恩：GMM voice recognition、ios app、ppt製作