基於GAN技術進行語音轉換技術探討

**摘要**

本實驗中，我們以生成對抗網路之概念發想，試著將其運用於語音轉換之領域，而在尋找適當的模型的過程中，我們發現一個由GAN所衍生出名為StarGAN之生成對抗網路。由於StarGAN的諸多優點，使其在語音轉換領域當中脫穎而出，可謂聚便利、準確、快速於一身。於是我們運用StarGAN實作語音轉換，並探討說話者人數對於轉換成果的影響。

我們將實驗所得數據以tensorboard繪出判別器、分類器、生成器之loss函數圖，透過觀察loss函數圖，可看出說話者人數有助於模型之穩定，並且在實現多對多轉換的情況下，仍可維持一定水準的轉換品質。

**壹、研究動機**

語音轉換(Voice Convertion)係指將一個人的聲音特色嵌入一句話語中，一般可運用在聲音沙啞，口腔損壞等發聲障礙者的聲音復原，抑或是非母語的發聲矯正、唱歌音調調整等。基於對相關文獻的探討，我們發現GAN是一個十分適合進行聲音轉換的工具。

生成對抗網路(GAN)的概念在2014年被提出，由於運用生成對抗網路(GAN)所生成出來的圖像有極好的成效，所以被視為深度學習的一大突破，到目前為止，生成對抗網路(GAN)大多數都用於生成圖片或是影片等等。但其實在類似的概念下，也能利用生成對抗網路(GAN)進行語音轉換，因此我們想利用GAN進行語音轉換，並設計實驗探討說話者人數對於轉換成果之影響。

**貳、研究目的**

本實驗中我們選擇利用GAN所衍生出的StarGAN模型進行語音轉換，並探討說話者人數對於轉換成果的影響。

**參、研究設備及器材**

一、硬體

（一）筆記型電腦

（二）台灣杉超級電腦（GPU來源）

二、軟體

（一）Linux（作業系統）

（二）Python 3.6（程式語言）

（三）TensorFlow 1.8（機器學習軟體庫）

（四）相關套件：librosa（音頻處理庫）、pyworld（語音合成套件）、tensorboard （tensorflow製圖套件）、scikit-learn（機器學習套件）

三、訓練數據集

為蒐集實驗所需數據，我們共錄製說話者五人，總音檔535個。

**肆、研究過程**

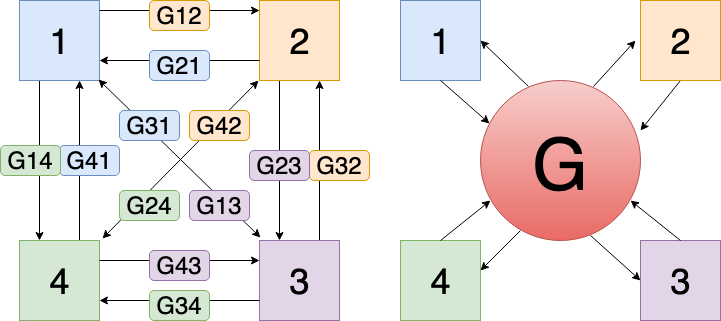
一、StarGAN

（一）生成對抗網路Generated Adversarial Network(GAN)

生成對抗網絡由一個生成器(Generator)與一個判別器(Discriminator)組成。生成器生成的輸出需要儘量模仿訓練集中的真實樣本。判別器的輸入則為真實樣本或生成器的輸出，其目的是將生成器的輸出從真實樣本中儘可能分辨出來；相對的，生成器則要儘可能地欺騙判別器。兩個網絡相互對抗，不斷調整參數，最終目的是使判別器無法判斷生成器的輸出結果是否真實。

（二）多領域生成對抗網路Star Generated Adversarial Network(StarGAN)

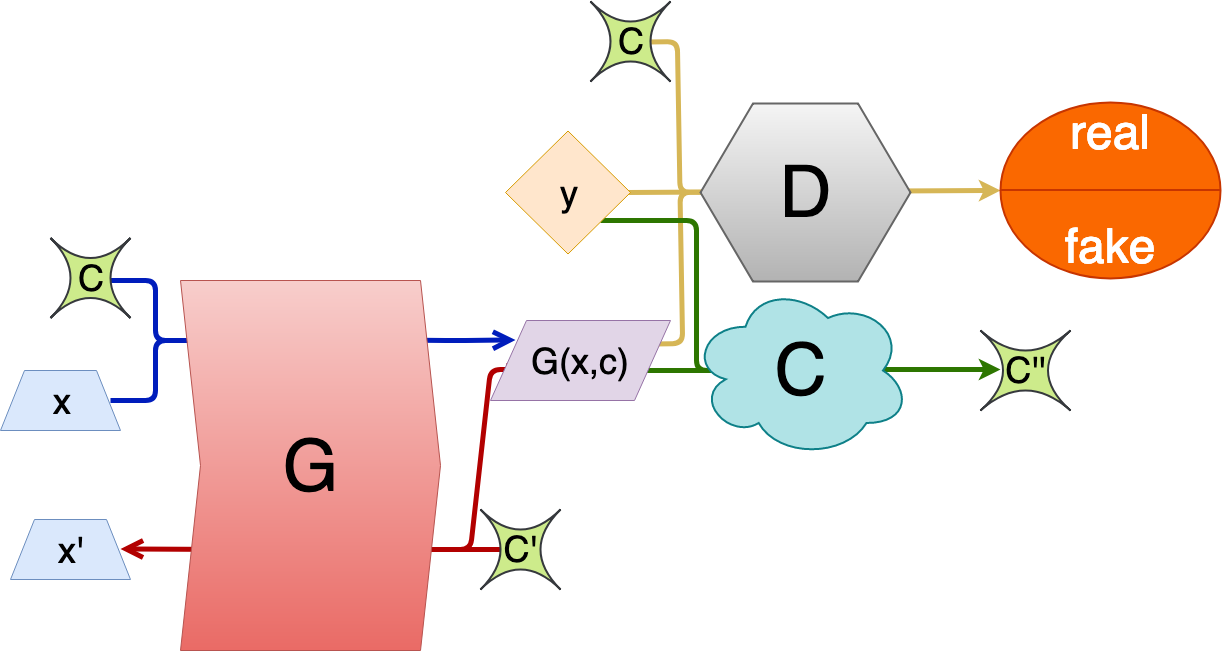
StarGAN屬於GAN的一種變形。StarGAN和其他GAN最大的不同在於其多出了分類器(Domain Classifier)的概念，一般的GAN僅能做到兩個域之間的轉換，而StarGAN透過加上訓練後的分類器能做到多個域中的互相轉換。本研究參考Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, Nobukatsu Hojo等人的StarGAN，並嘗試以調整說話者人數，評估此方法成效。



**圖一：StarGAN示例**

如圖一所示，在一樣有4個說話者的狀況下，左圖為一般的一對一模型共需12個生成器（Gij），右圖則為StarGAN所需的生成器，由圖一可看到一般一對一模型所需的眾多生成器在StarGAN中僅需要一個。

如圖二所示，整個StarGAN框架由三部分組成：生成器G、判別器D、分類器C。c和c'各代表著一個人聲標籤(label)，分別做為模型生成音頻或判斷真偽的依據，以下簡稱為屬性(attribute)。將一段屬性被標示為c'的真實音頻x加上屬性c後，以生成器G產生的新音頻為G(x, c)；接下來將屬性為c的真實音頻y與前述音頻G(x, c)送入判别器，判斷此音頻為真實音頻或是生成器所生成的偽音頻，其中真實音頻x和y指的是數據集中的原始音頻。x'可以G(G(x, c), c')表示，意指將音頻G(x, c)加上屬性c'傳入生成器而生成的音頻。c''則是分類器C所判斷音頻的屬性結果。



**圖二：StarGAN流程圖**

1. 模型架構

整個模型框架採用CNN搭建起來，並用了門限卷積神經網路GatedCNN(Gated Convolutional Neural Network)作為主要手段。圖三為整個框架的設計。

****

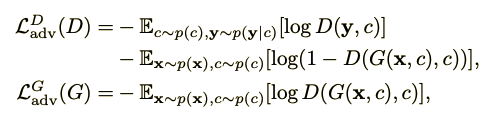
**圖三：StarGAN框架**

2. 損失函數(loss)

(1) Adversarial Loss

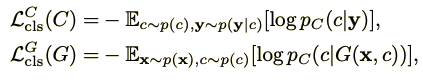
Adversarial Loss損失函數的概念最初由原本GAN作者提出，判別器D負責識別出生成器G生成的假樣本G(x, c)，並將其判為假，原本就是屬性c的音檔y則會被判為真；生成器G以欺騙判別器D為目標，也就是使生成的假樣本G(x, c)能被判別器D判為真。

下列式子D(X, c)表示判別器認為變數X是屬性c之真實音頻的機率，越接近0即是假、接近1則是真。判別器D期望將真實樣本判斷為真（即D(y, c)越高越好），並且期望假樣本能被判斷為假（即D(G(x, c), c)越低越好）；生成器G反而期望D(G(x, c), c)越高越好。故兩者的Adversarial Loss作以下形式。



(2) Domain Classification Loss

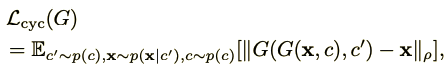
分類器C的任務為準確區分出音頻之屬性準確；對於生成器G所產生的音頻G(x, c)，最理想的狀況為能被分類器分類為c屬性。



其中pC (c | X) 表示在音檔是X的情況下，X被C分到屬性c的條件機率。

(3) Cycle Consistency Loss

若將帶有屬性c' 的x音檔轉換成帶有c屬性的G(x, c)，再將G(x, c)轉回原本的c' 屬性 ，也就是G(G(x, c), c')，這個東西理應和x越接近越好。

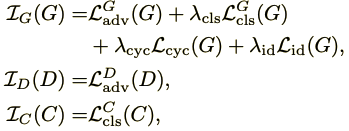


(4) Identity Mapping Loss

為了確保當輸入的音檔x已經屬於目標屬性c'時，生成器G的輸出將保持不變。所以將屬性為c' 的x音檔轉成c' 屬性 (G(x, c'))，理應和x 越接近越好。

**

整個網絡大方向上有生成器G的損失，判别器D的損失和分類器C的損失。最後將這些損失統整，對於G、D、C有以下Loss。



（三）選擇運用StarGAN進行語音轉換之原因

語音轉換之技術發展已有一段時間，但在GAN出現之前，執行語音轉換的過程中往往會遇到一些困難，例如源語音和目標語音需嚴格的成對，這樣的限制使得研究者們在蒐集數據的過程中相當的不方便。此外，源語音和目標語音之時間也需要嚴格對齊，這些因素除了導致實務執行困難，似乎也偏離了聲音轉換的原始想法，但是在已有的基礎上進行非並行的訓練的話，其轉換出的成果顯然無法與前述之並行語音訓練的結果相比擬，於是語音轉換的發展陷入了窘境。

StarGAN的誕生成了解決這些問題的契機，運用StarGAN進行語音轉換時，源語音和目標語音並不需要嚴格配對，時間方面上也無需要對齊，又能得到和並行訓練差不多的訓練結果，這樣一來語音轉換的工程變得簡單一些。除了擁有上述優點之外，它還具備了可以同時學習多個域之間的多對多映射，並且使用對抗性損失對生成器進行訓練，這樣的方法使得生成器能夠輸出更逼真的結果，此外StarGAN也能夠在短時間內完成訓練。

二、前置處理

（一）將預先錄好的語音數據集透過iTunes轉為wav格式。

（二）分別建立train（訓練集）以及test（測試集）資料夾。

（三）分別將欲訓練之語音數據集放入train資料夾，欲測試之語音數據集放入test資料夾。

（四）針對訓練集進行預處理，此步驟的目的在於提取語音數據集的特徵（聲譜圖），並將這些特徵以npy檔案的形式儲存起來，以便進行轉換。

三、進行訓練

（一）讀取先前預處理完的preprocess資料夾中讀取特徵之npy檔案，將其用於訓練模型。

（二）紀錄訓練時的Loss函數變化

（三）讀取測試集中的語音數據，以便在訓練期間使用模型隨機生成一些樣本。

（四）運用前面訓練好之模型，將源語音和目標語音進行轉換。

四、實驗

在實驗中，我們將speaker數量分別定為2, 3, 4, 5人來進行測試及討論，訓練資料為97個音檔，測試資料為10個音檔，將data放入StarGAN模型中進行。

**伍、研究結果**

根據上述實驗結果，我們依不同說話者人數，分別將各組實驗的判別器、分類器、生成器之loss函數列表如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 表1. 說話者人數及不同模型之loss函數分佈圖 | | | | |
|  | 2 speakers | 3 speakers | 4 speakers | 5 speakers |
| D Adversarial Loss | Macintosh HD:Users:student:Downloads:exp:exp2:螢幕快照 2019-02-15 下午2.18.55.png | Macintosh HD:Users:student:Downloads:exp:exp3:螢幕快照 2019-02-15 下午2.20.15.png | Macintosh HD:Users:student:Downloads:exp:exp4:螢幕快照 2019-02-15 下午2.21.57.png | Macintosh HD:Users:student:Downloads:exp:exp5:螢幕快照 2019-02-15 下午2.25.27.png |
| G Adversarial Loss | Macintosh HD:Users:student:Downloads:exp:exp2:螢幕快照 2019-02-15 下午2.19.19.png | Macintosh HD:Users:student:Downloads:exp:exp3:螢幕快照 2019-02-15 下午2.20.36.png | Macintosh HD:Users:student:Downloads:exp:exp4:螢幕快照 2019-02-15 下午2.22.18.png | Macintosh HD:Users:student:Downloads:exp:exp5:螢幕快照 2019-02-15 下午2.25.44.png |
| C Domain Classification Loss | Macintosh HD:Users:student:Downloads:exp:exp2:螢幕快照 2019-02-15 下午2.19.04.png | Macintosh HD:Users:student:Downloads:exp:exp3:螢幕快照 2019-02-15 下午2.20.22.png | Macintosh HD:Users:student:Downloads:exp:exp4:螢幕快照 2019-02-15 下午2.22.05.png | Macintosh HD:Users:student:Downloads:exp:exp5:螢幕快照 2019-02-15 下午2.25.33.png |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 表1. 說話者人數及不同模型之loss函數分佈圖(續) | | | | |
|  | 2 speakers | 3 speakers | 4 speakers | 5 speakers |
| G Domain Classification Loss | Macintosh HD:Users:student:Downloads:exp:exp2:螢幕快照 2019-02-15 下午2.19.32.png | Macintosh HD:Users:student:Downloads:exp:exp3:螢幕快照 2019-02-15 下午2.20.56.png | Macintosh HD:Users:student:Downloads:exp:exp4:螢幕快照 2019-02-15 下午2.22.34.png | Macintosh HD:Users:student:Downloads:exp:exp5:螢幕快照 2019-02-15 下午2.25.59.png |
| Cycle Consistency Loss | Macintosh HD:Users:student:Downloads:exp:exp2:螢幕快照 2019-02-15 下午2.19.14.png | Macintosh HD:Users:student:Downloads:exp:exp3:螢幕快照 2019-02-15 下午2.20.30.png | Macintosh HD:Users:student:Downloads:exp:exp4:螢幕快照 2019-02-15 下午2.22.12.png | Macintosh HD:Users:student:Downloads:exp:exp5:螢幕快照 2019-02-15 下午2.25.40.png |
| Identity Mapping Loss | Macintosh HD:Users:student:Downloads:exp:exp2:螢幕快照 2019-02-15 下午2.19.28.png | Macintosh HD:Users:student:Downloads:exp:exp3:螢幕快照 2019-02-15 下午2.20.47.png | Macintosh HD:Users:student:Downloads:exp:exp4:螢幕快照 2019-02-15 下午2.22.29.png | Macintosh HD:Users:student:Downloads:exp:exp5:螢幕快照 2019-02-15 下午2.25.53.png |

**陸、討論**

一、loss函數之曲線

由上面說話者個數的loss函數圖可以看出，當說話者個數越多的時候，loss函數的曲線趨於平順，在訓練的過程中，loss比較不會有突然間的大幅變化，以這樣的結果看來，我們可以推測說話者個數有助於模型的穩定。

二、轉換之語音品質

從轉換結果可發現，當speaker數量愈多時，品質並沒有隨之下降，因此可見StarGAN的多對多語音轉換極為成功，比起原本GAN模型多對多時需要的眾多生成器及判別器，StarGAN利用分類器僅需一個生成器及判別器，加快了訓練時間。

**柒、結論**

利用StarGAN進行聲音轉換是快速且有效的，訓練時只跑了100個Epoch，便能產生逼近目標的聲音，且配合GPU，四個speaker訓練時間不到一小時，為一個快速轉換多對多語音的模型，即，可以在不損失音頻品質的前提下大幅提升語音轉換的目的。

**捌、參考資料及其他**

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio(2014). Generative Adversarial Networks. stat.ML. arXiv:1406.2661v1.

[2] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, Jaegul hoo1(2014).Generative Adversarial Networks. http://openaccess.thecvf.com/content\_cvpr\_2018/papers/Choi\_StarGAN\_Unified\_Generative\_CVPR\_2018\_paper.pdf

[3] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, Nobukatsu Hojo(2018). StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks. cs.SD. [arXiv:1806.02169](https://arxiv.org/abs/1806.02169)