Speech resynthesis from discrete disentangled self-supervised representations

A. Polyak, et al., “Speech resynthesis from discrete disentangled self-supervised representations,” arXiv:2104.00355 [cs.SD], Jul. 2021

Student : Sian-Yi Chen

Advisor : Tay-Jyi Lin and Chingwei Yeh

Outline

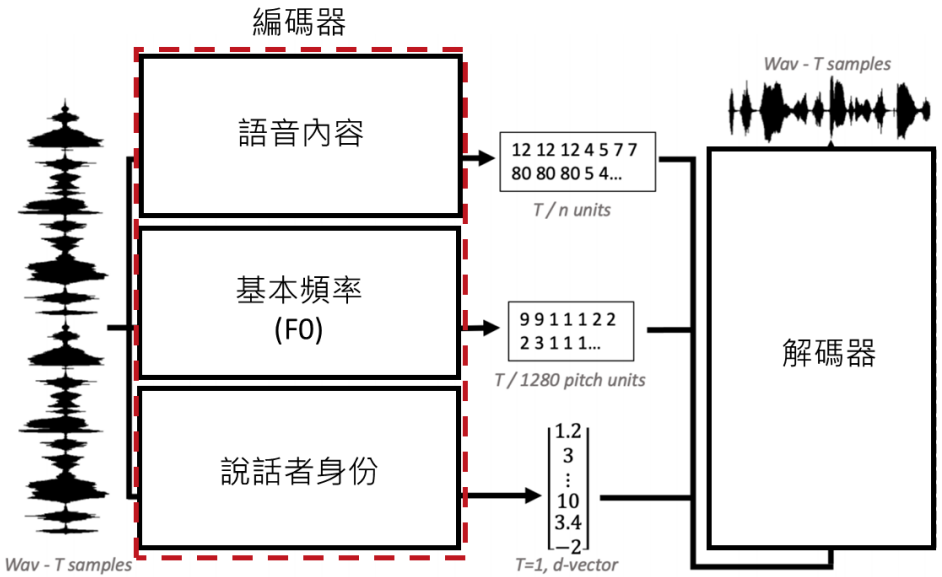
1. Abstract
2. Introduction
3. Method
   - Encoder
   - Decoder
4. Results
5. Conclusion

Abstract

論文中的語音轉換使用了自監督學習 (SSL, Self-Supervised Learning)(p.9) 這種技術。

語音轉換架構中(圖一)使用了三個編碼器以及一個解碼器，三個編碼器分別從原始語音中提取了[1]語音內容、[2]基本頻率(聲音中頻率最低的音)以及[3]說話者身分，最後再使用解碼器重構訊號。

論文中以語音合成的角度對於 SSL 進行廣泛的評估，像是訊號重組、語音轉換、錄音的可理解程度和整體效能。最後論文表示完成了速率為 365bits/s 的輕量級語音轉換器。



(圖一) 語音再合成架構

Introduction

因為 Self-Supervised Learning (SSL) 方法成功後，無論是在連續或是離散的無監督學習中，效果都大幅躍進。

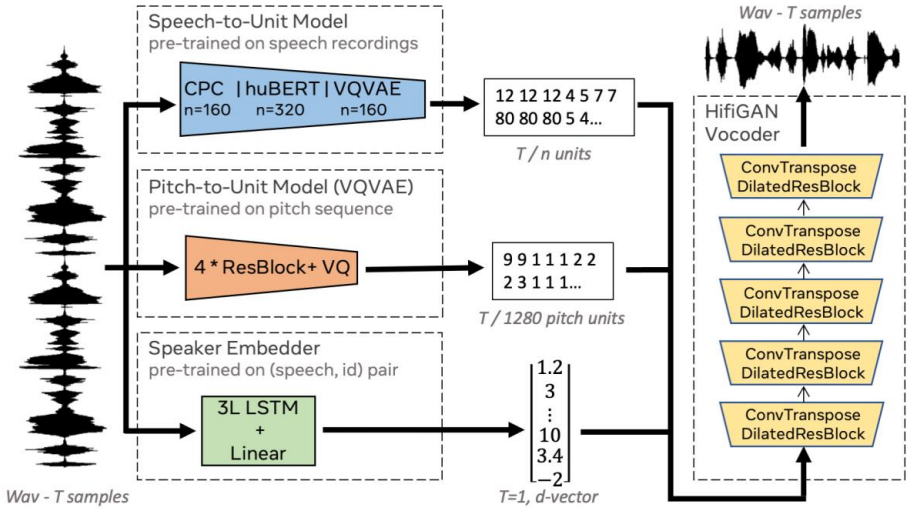
目前對於使用 SSL 方法應用在語音上，大部分都集中在 Automatic Speech Recognition (ASR) 生成語音和評估品質，雖然 SSL 取得了成功，但目前還不確定這些方法是否適合用於語音合成。

語音轉換架構中(圖一)使用了三個編碼器以及一個解碼器，三個編碼器分別從原始語音中提取了語音內容、基本頻率 (F0) 以及說話者身分，並使用多個數據集和編碼器模型進行實驗操作。

最後論文按照提出的方法得到一個速率為 365bits/s的輕量級語音編解碼器。

論文中共有三點貢獻

1. 為了高品質的合成，從離散的語音單元中使用自監督學習
2. 從合成的角度對 SSL(Self-Supervised Learning) speech 進行廣泛的評估，舉例來說訊號重組、語音轉換、F0 的控制
3. 完成了一個超輕量級的語音編解碼器



(圖一) 語音再合成架構

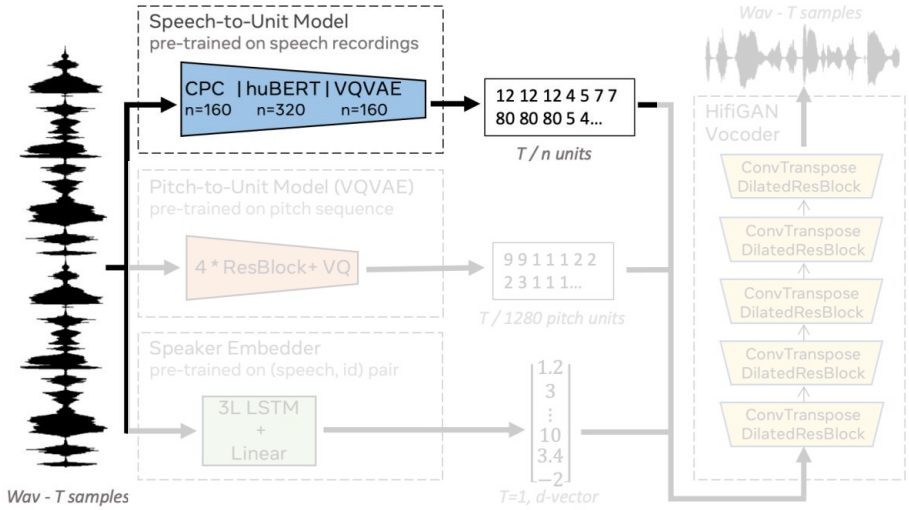
Method

內容編碼器：
它的輸入為語音，輸出則是一系列以低頻採樣的頻譜表示
編碼器是由以下三種最先進的無監督式學習組成

(i) CPC (Contrastive Predictive Coding)，它利用現有的資訊來預測未來的高維度資料，其中重要的精神是在資料標籤很少的情況下可以有效學習到需要的資料。

(ii) HuBERT (Hidden Unit Bidirectional Encoder Representations from Transformers)，它類似於 BERT 的遮蓋預測任務進行訓練，以遮蓋連續音訊作為輸入。目標是通過對原始語音特徵或從早期迭代中學習到的特徵獲得聚類(p.10)。

(iii) 和 VQ-VAE (Vector-Quantized Variational AutoEncoder)，運作原理也是尋找輸入資料的潛在特徵，讓隨機輸入的資料可以在輸出時被還原。



(圖一) 語音再合成架構

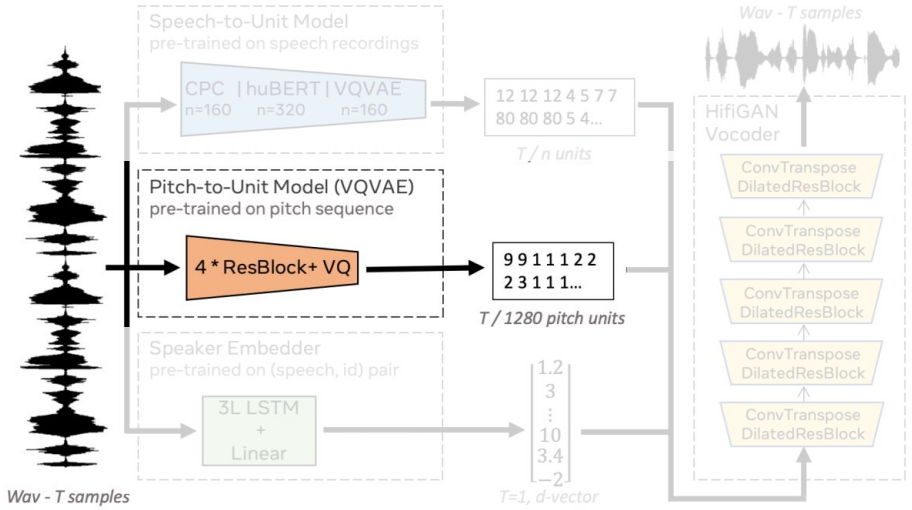
Method

基本頻率 (F0) 編碼器：
為了生成低頻離散 F0，編碼器使用了 VQ-VAE (Vector-Quantized Variational AutoEncoder) 作為框架訓練。

因此在此架構中，有兩個 VQ-VAE 模型，一個用於波形，另一個用於 F0。

另外作者還有使用 YAAPT (Yet Another Algorithm for Pitch Tracking) 演算法來提取 F0。

YAAPT 是一種基本頻率的跟蹤演算法，它對於高品質語音或電話語音有高度準確性和強健性 (Robustness)。



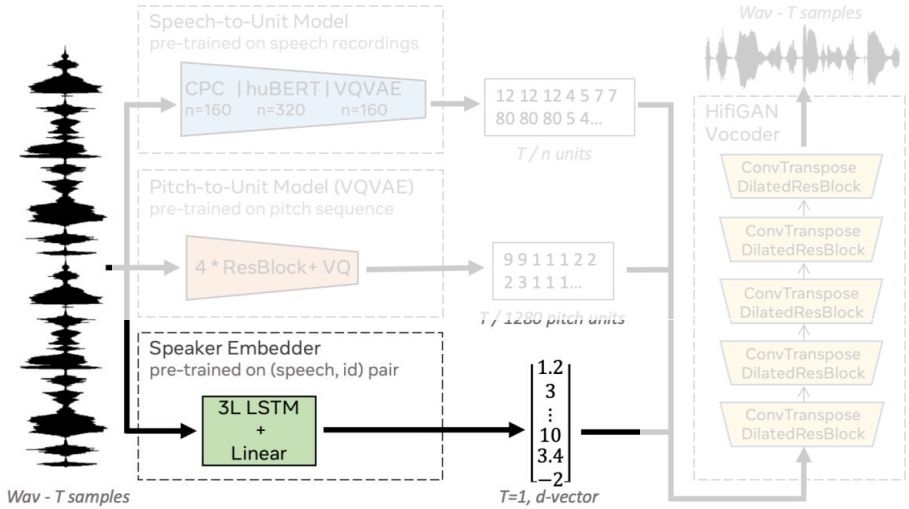
(圖一) 語音再合成架構

Method

揚聲編碼器：
使用了預訓練說話者驗證模型，輸入聲音，提取梅爾頻譜圖，並輸出說話者表示向量

預訓練：是（被)預先訓練過的模型，如果有一個很大很浪費時間訓練的模型，不想從頭訓練，可以下載別人已經訓練好的模型（parameter）來使用

此說話者驗證模型可以用來判斷是誰在說話

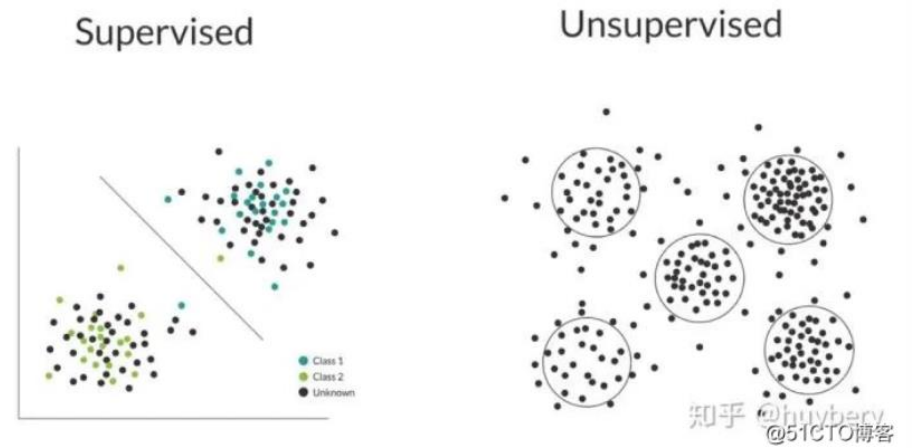


（圖一）語音再合成架構

Self-Supervised Learning

機器學習中的有兩種基本的學習正規化

1. 監督式學習
2. 無監督式學習



監督學習透過人工大量的標註資料來訓練模型，模型的預測和資料的真實標籤產生後進行反向傳播，通過不斷學習，最後獲得辨識新樣本的能力

無監督學習不依賴任何標籤，自行對資料內特徵挖掘，找到樣本間的關係，如上圖，左邊有三種型態，class1、class2還有未知，而右邊分成5區卻沒有標示任何訊息

兩種最主要的差別在於訓練時是否需要人工標記標籤資訊

那什麼是自監督式學習？

自監督式學習主要利用**輔助任務**從大規模無監督式學習的資料中自己去挖掘監督資訊。
在無監督學習下，會產生多筆集合，而輔助任務會將這些沒有標籤的集合，選出需要的集合標上標籤。

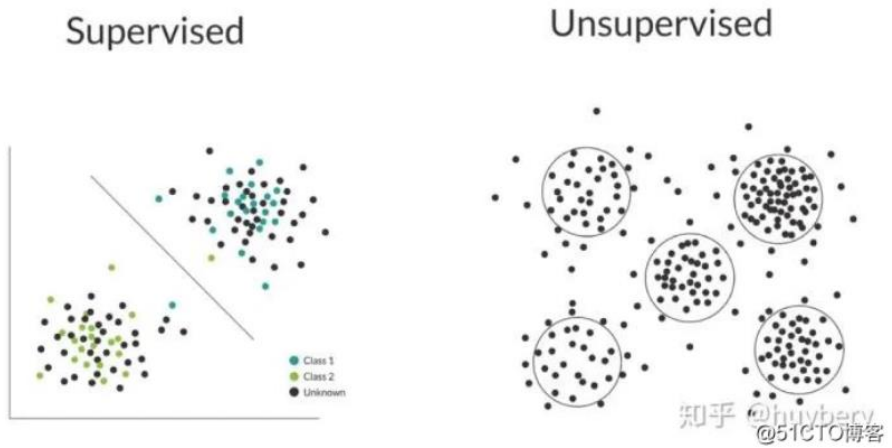
也就是自監督式學習具有傳統手動取得特徵的優點，又具有無監督學習透過模型自行學習的優點。

分類？聚類？

分類與聚類，可以簡單的使用屬於監督學習，或是無監督學習來區分。

分類：
根據現有資料的特徵或是屬性去做區分，也就是說這些類別是已知的，因此如果是在監督學習下，就屬於分類。

聚類：
在無監督學習下，不知道模型會將資料分成多少類，分類的標準也是未知的，但每一聚類都是由物理性質或抽象性質相似的類別所組成



【自監督學習】Self-supervised Learning 再次入門