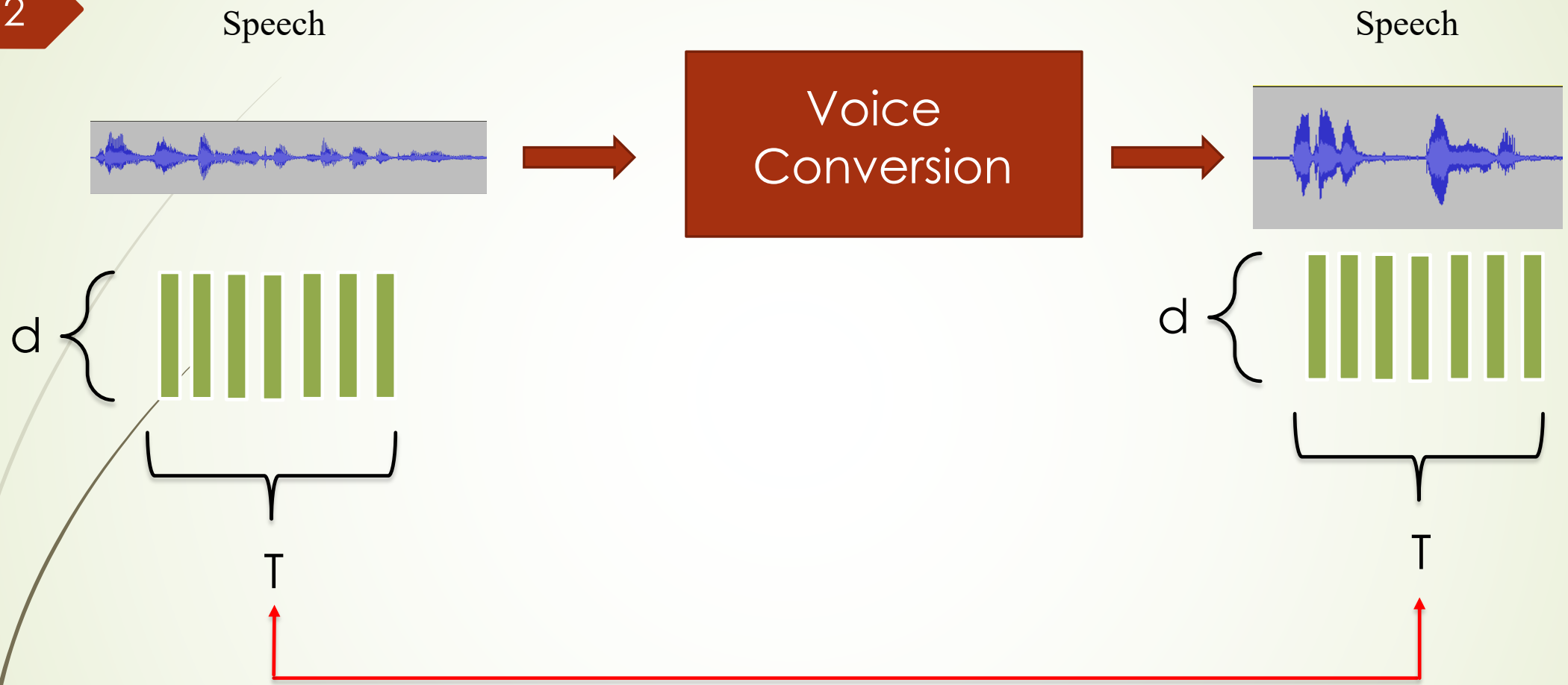
Multi Target voice conversion and cross-language

長庚大學 資工所 劉祈宏
指導教授 呂仁園

What is Voice Conversion



What can be preserved?

What is changed?

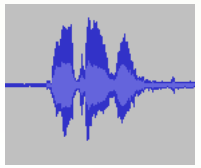
內容

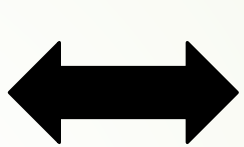
許多都可以，例如語者

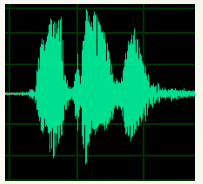
https://www.kocpc.com.tw/archives/182613

Voice Conversion 技術發展分類

with平行語料







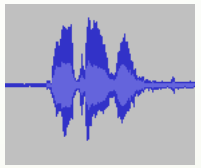
你好

source

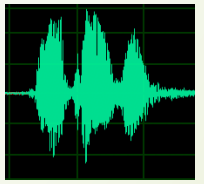
你好

target

without平行語料



你好

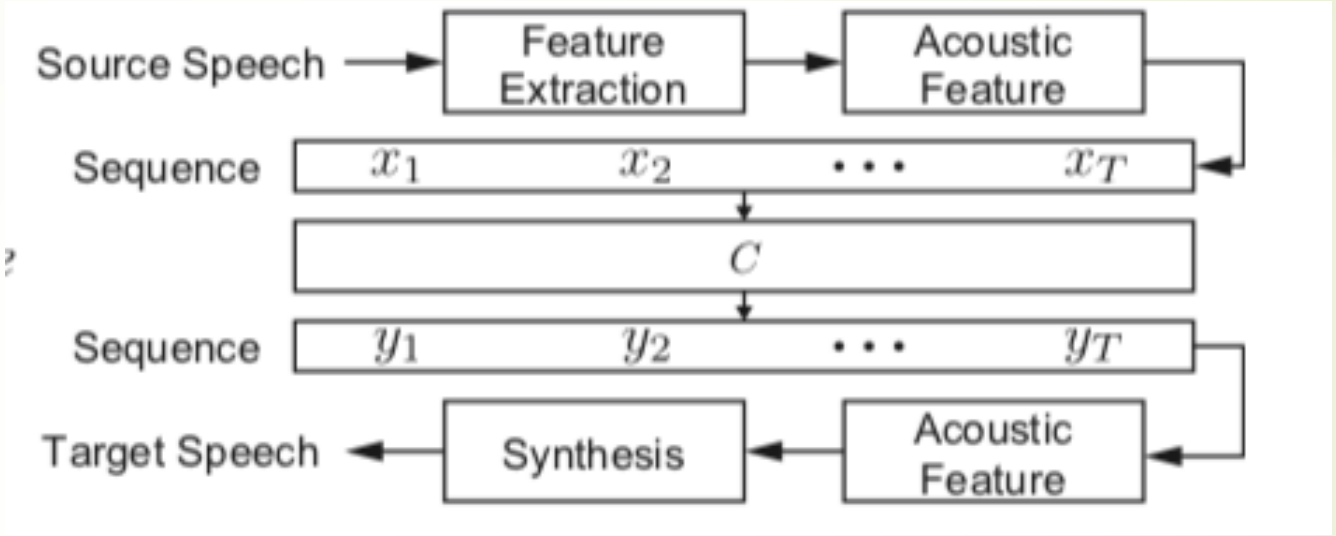


天氣不錯

平行語料

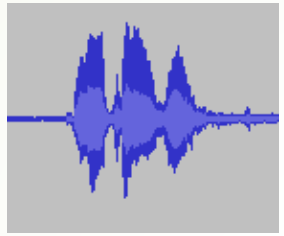
非平行語料

Sequence to sequence model



Sequence-to-Sequence Voice Conversion with Similarity Metric Learned Using Generative Adversarial Networks , *INTERSPEECH 2017*

平行語料



Content Encoder

Phonetic information

Speaker Encoder

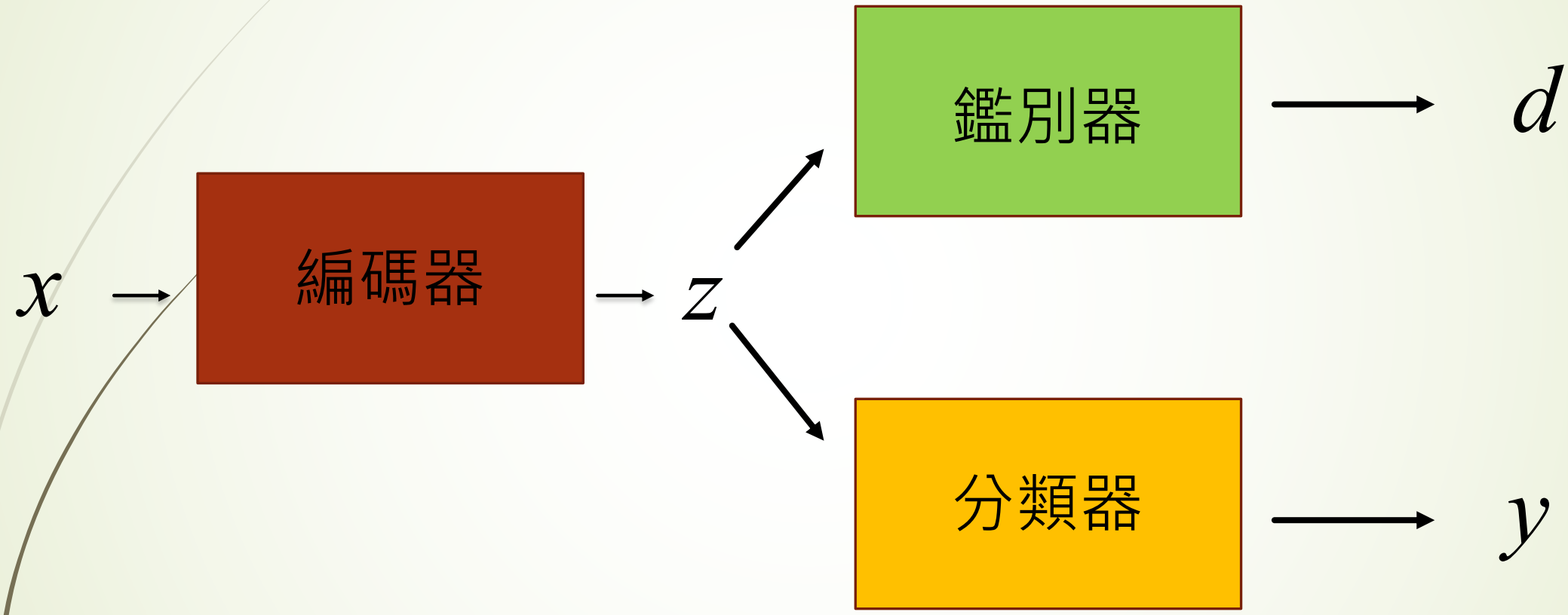
Speaker information

非平行語料

Feature Disentangle (特徵解纏)

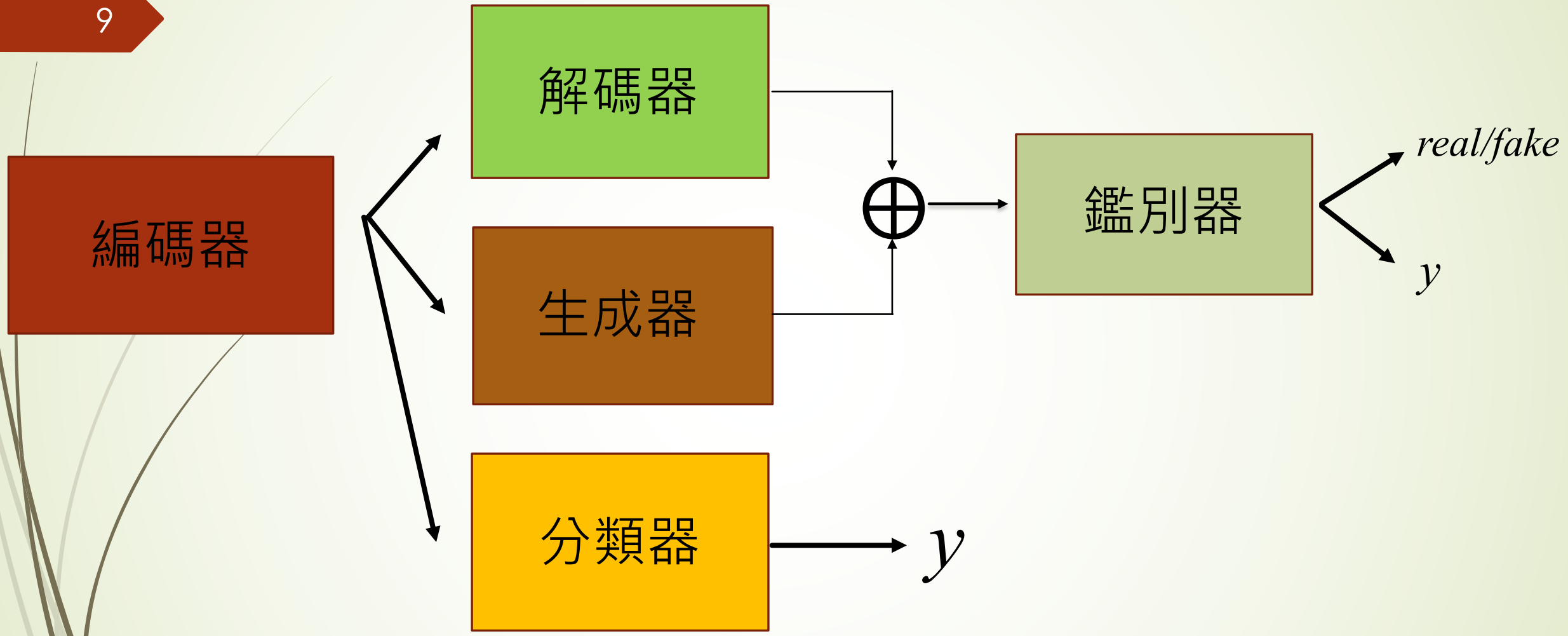
Direct Transformation (直接轉換)

使用對抗訓練方法進行解纏特徵學習



x為資料，y為分類器的輸出，z為學習之潛在特徵

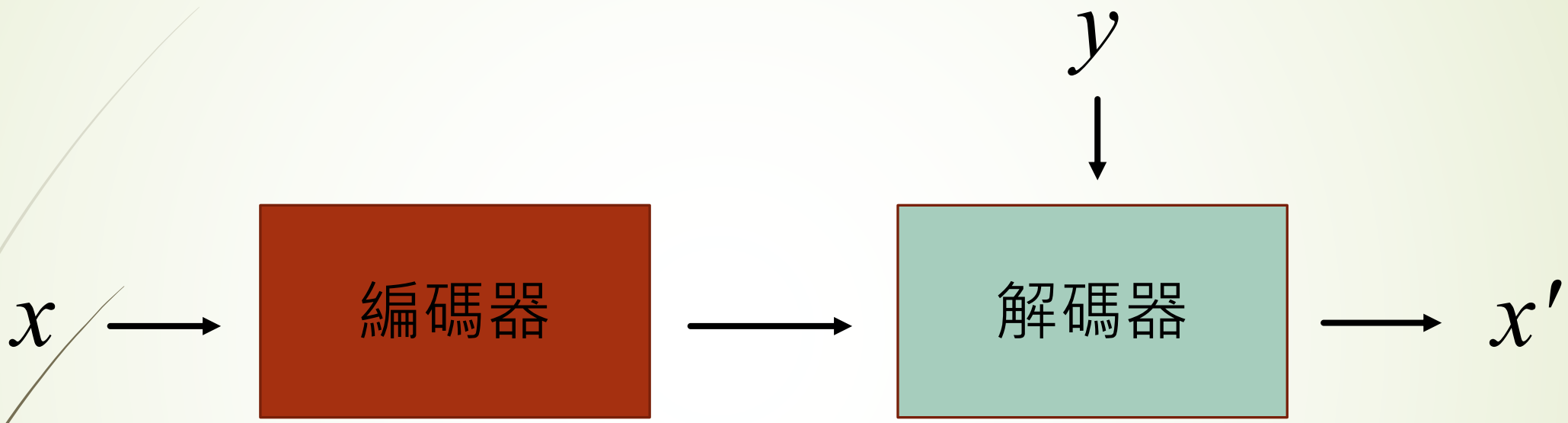
如果只有非平行語料，又想實現多目轉換



其中y為語者編號，⊕為個別元素相加(Elementwise Addition)。

模型架構與訓練過程

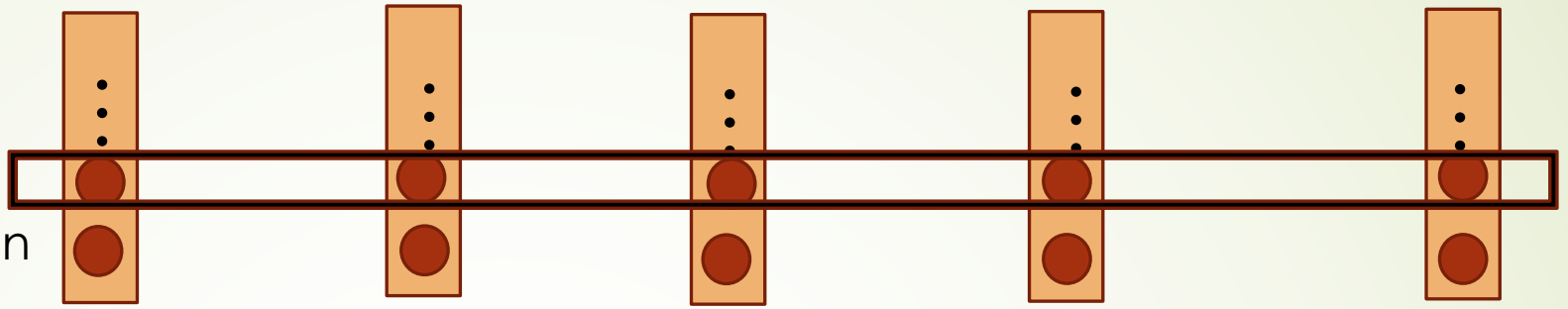
- 預訓練階段:此階段如同訓練一個自編碼器。
- 解纏特徵學習階段:此階段會使用一個輔助分類器來幫助編碼器壓縮出不含語者資訊的特徵。
- 生成對抗網路階段:此階段的訓練會將編碼器以及解碼器參數固定，並且訓練平行於解碼器的一個生成器來生成解碼器的殘差訊號(Residual Signal)，進 而能夠生成較為尖銳、與真實資料較相近的輸出。



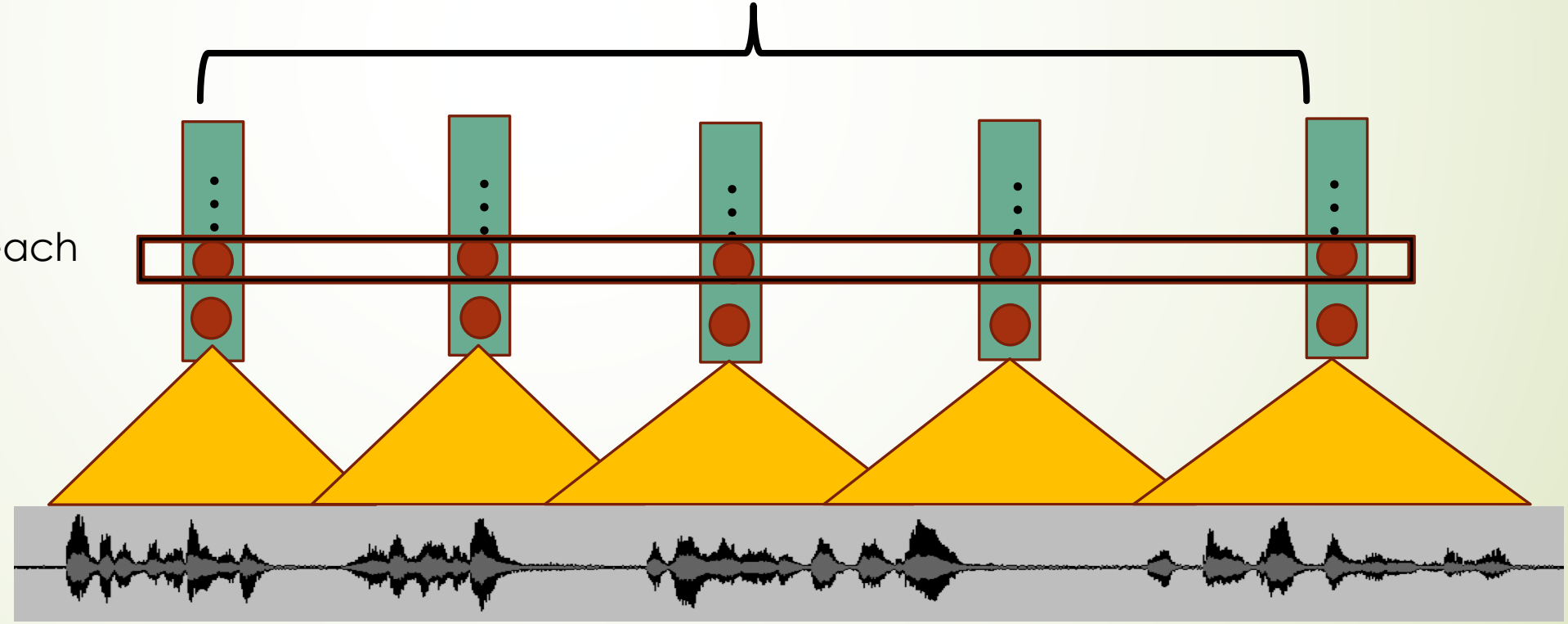
預訓練階段示意圖。在這個階段視同訓練一個語音訊號的自編碼器，但會同時在解碼器輸入語者的編號。x為語音訊號，y為語者編碼

instance normalization

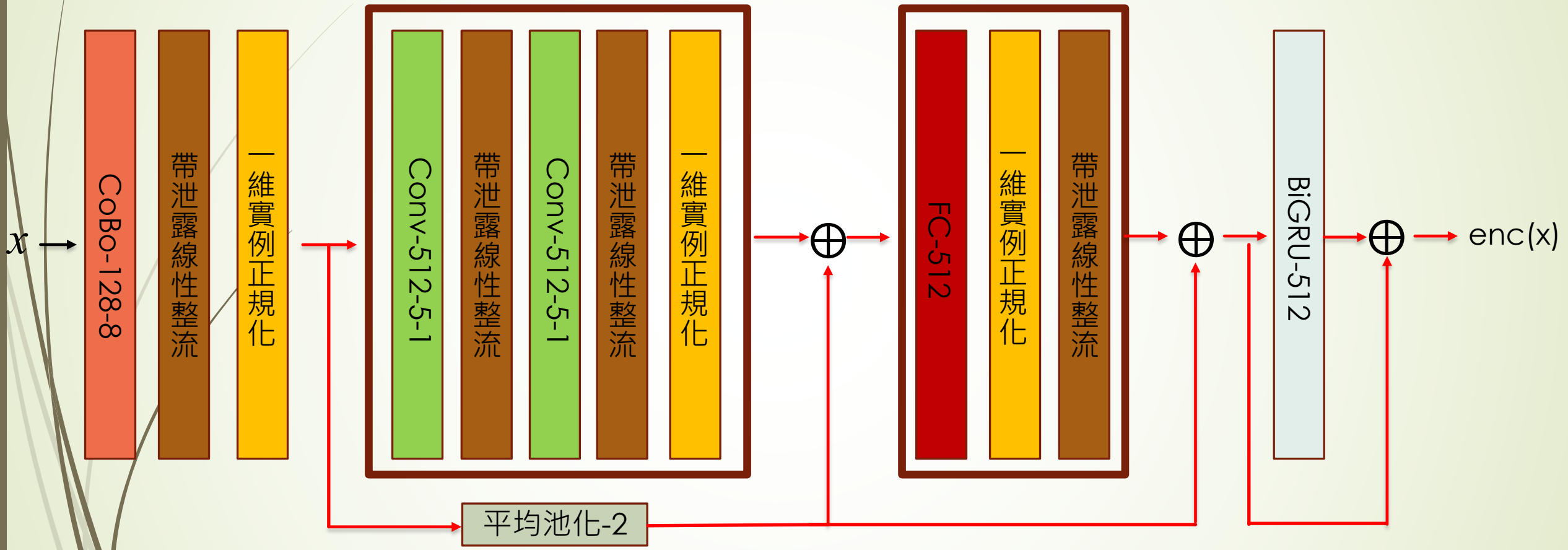
Each channel has zero mean and unit variance

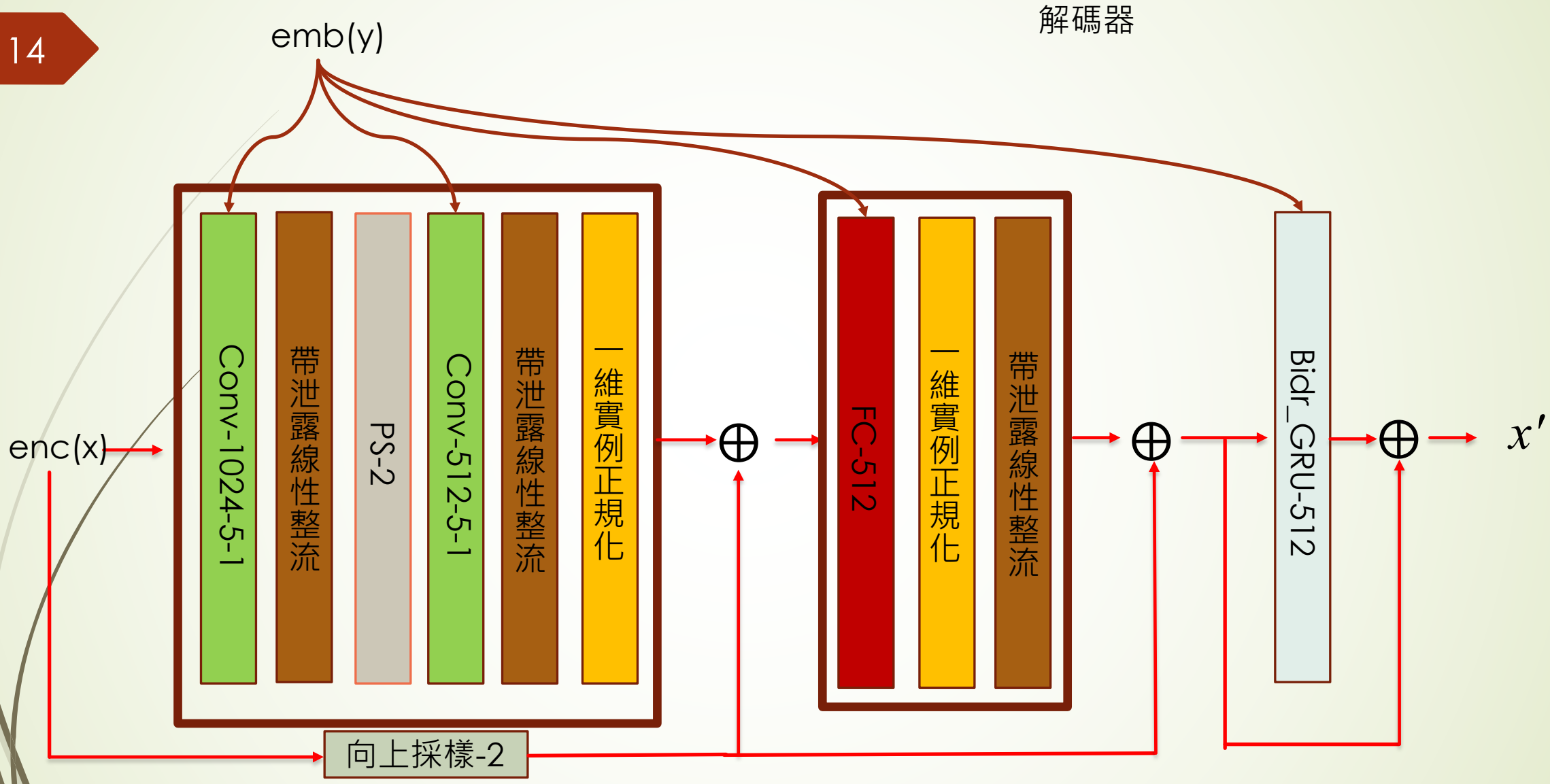


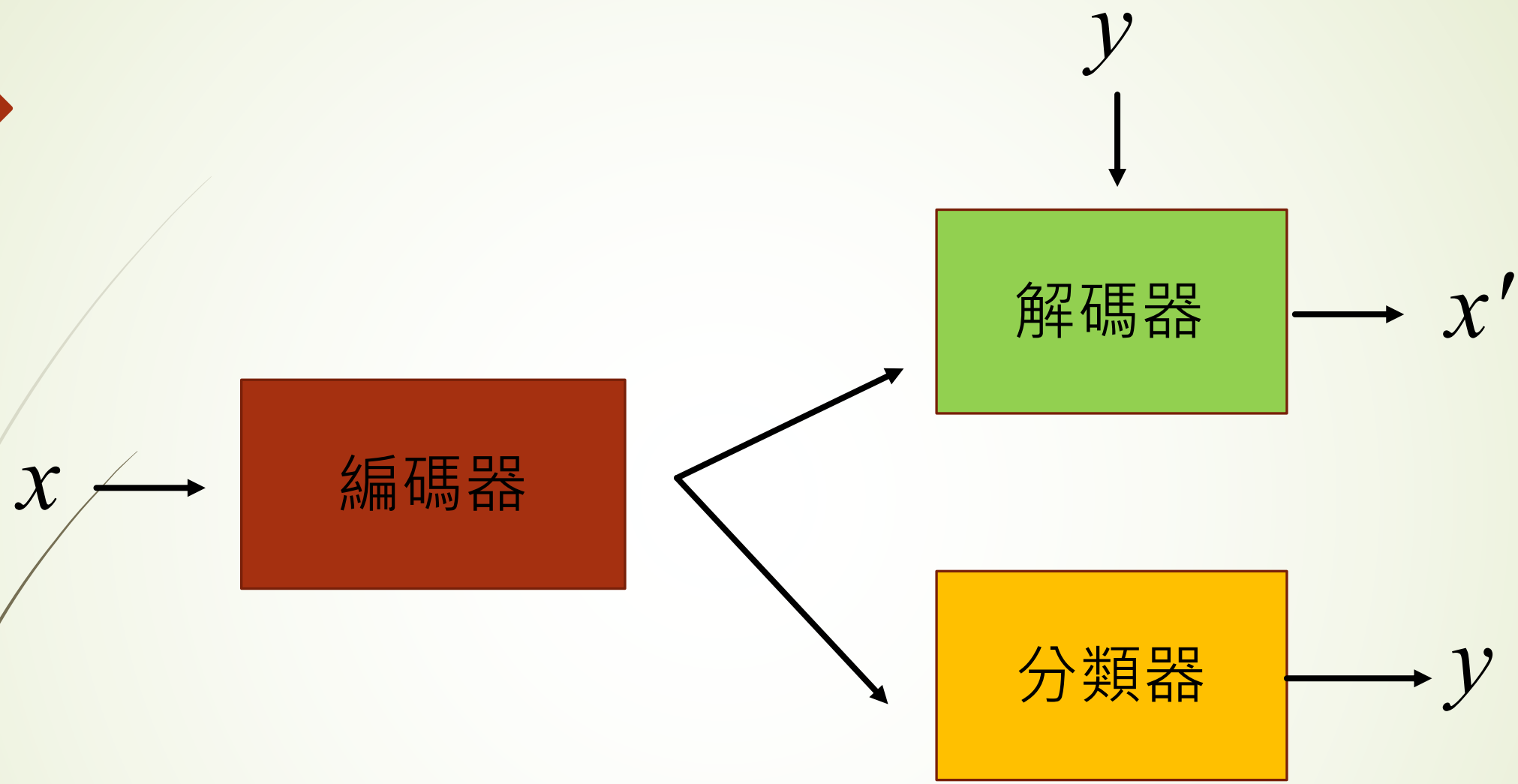
Normalize for each channel



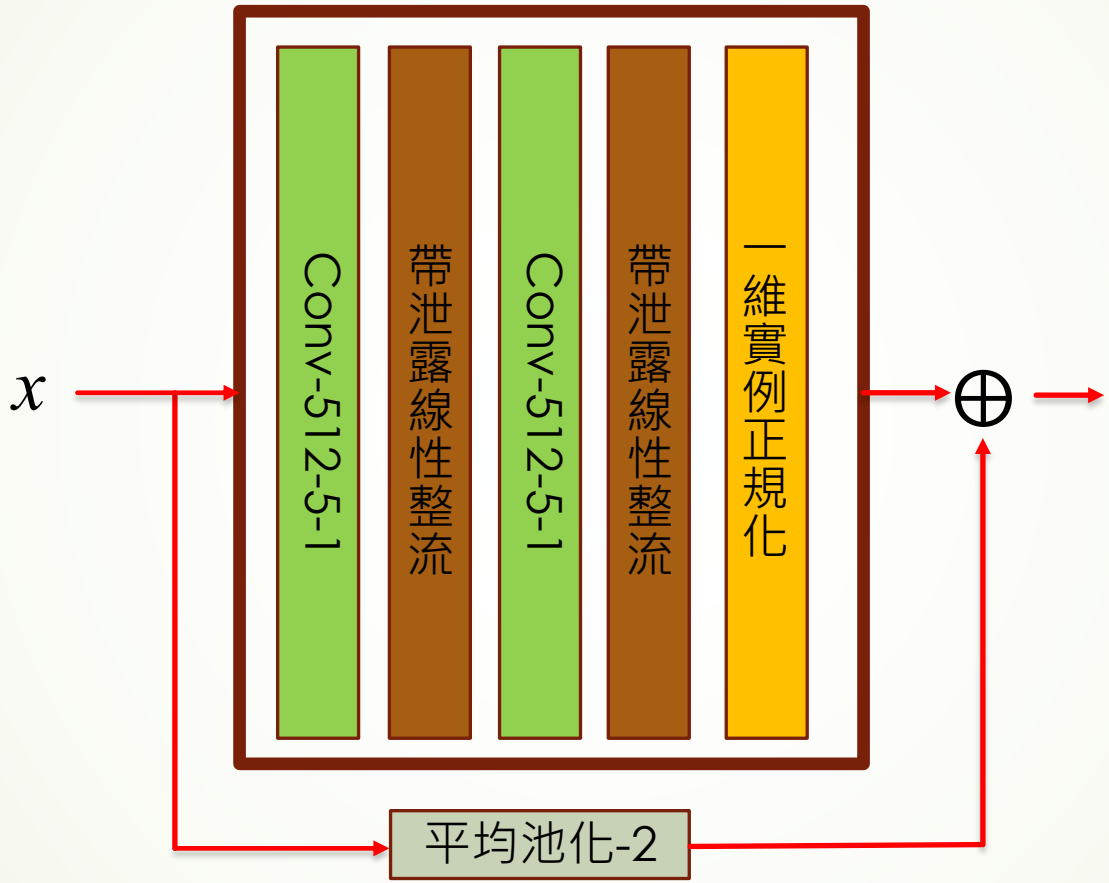
編碼器

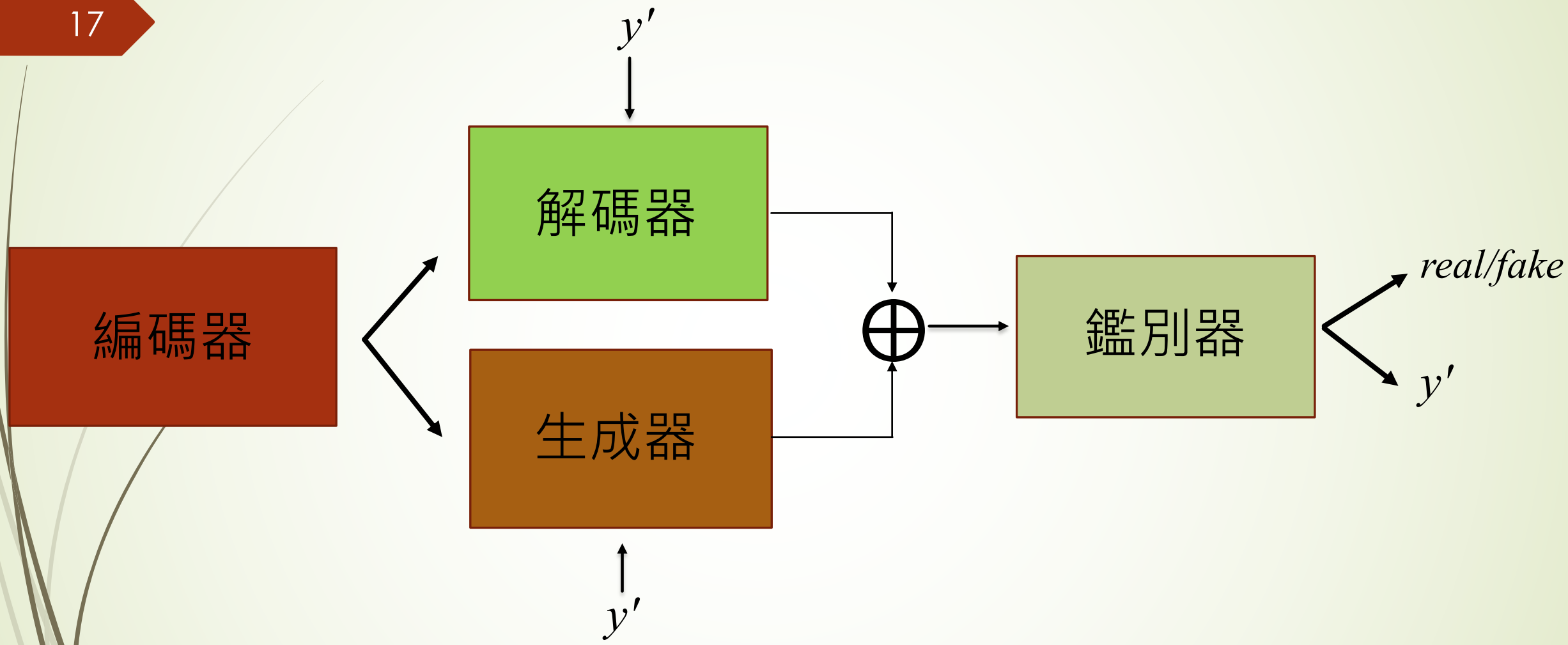

x
CoBo-128-8
帶泄露線性整流
一維實例正規化
Conv-512-5-1
帶泄露線性整流
Conv-512-5-1
帶泄露線性整流
一維實例正規化
平均池化-2
FC-512
一維實例正規化
帶泄露線性整流
BiGRU-512
enc(x)


解碼器
emb(y)
enc(x)
Conv-1024-5-1
帶泄露線性整流
PS-2
Conv-512-5-1
帶泄露線性整流
一維實例正規化
FC-512
一維實例正規化
帶泄露線性整流
Bidr_GRU-512
向上採樣-2
x′

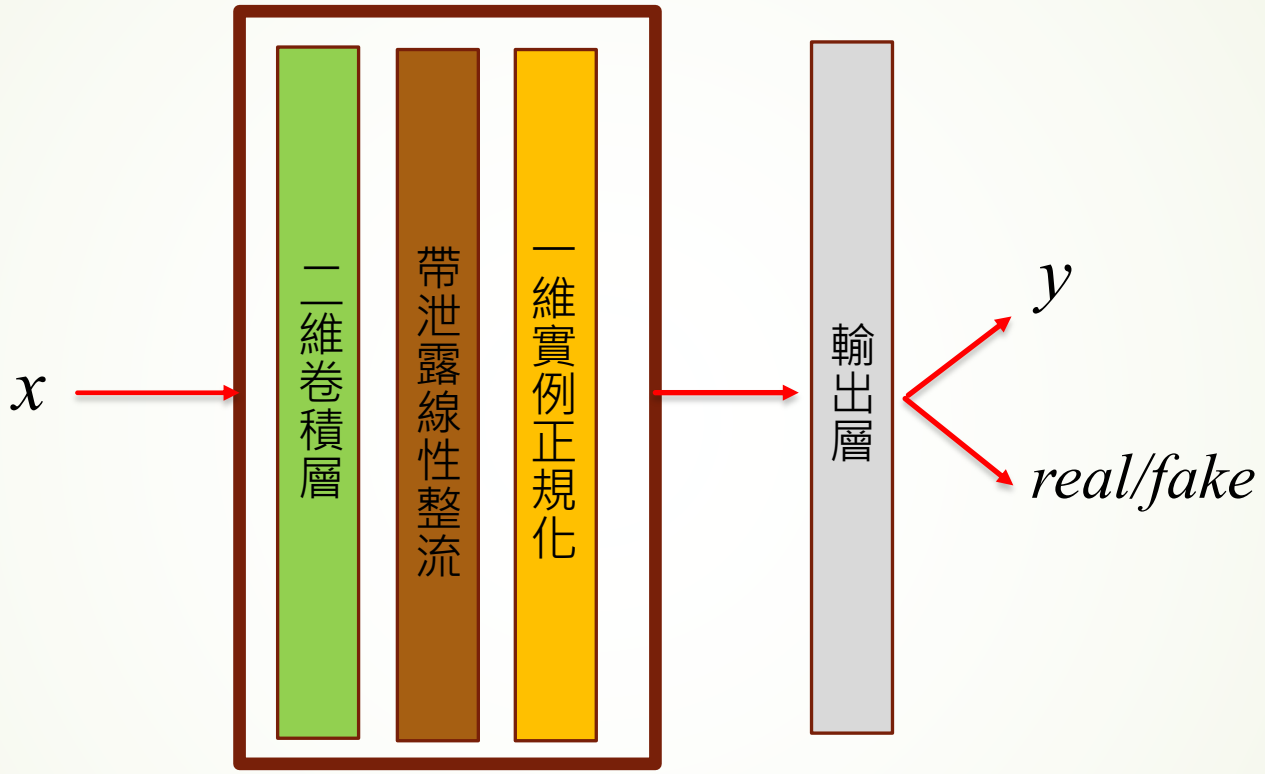


解纏特徵學習階段示意圖。
在這個階段會引入分類器來規範化(Regularize)編碼器生成的語音特徵


x
Conv-512-5-1
帶泄露線性整流
Conv-512-5-1
帶泄露線性整流
一維實例正規化
⊕
平均池化-2



生成對抗網路階段。在這個階段會引入一個生成器以及一個鑑別器。透過生成器與鑑別器的交互訓練，來提升模型所生成的語音品質


x
二維卷積層
帶泄露線性整流
一維實例正規化
輸出層
y
real/fake

DEMO

https://eric4404123.github.io/voice_conversion/