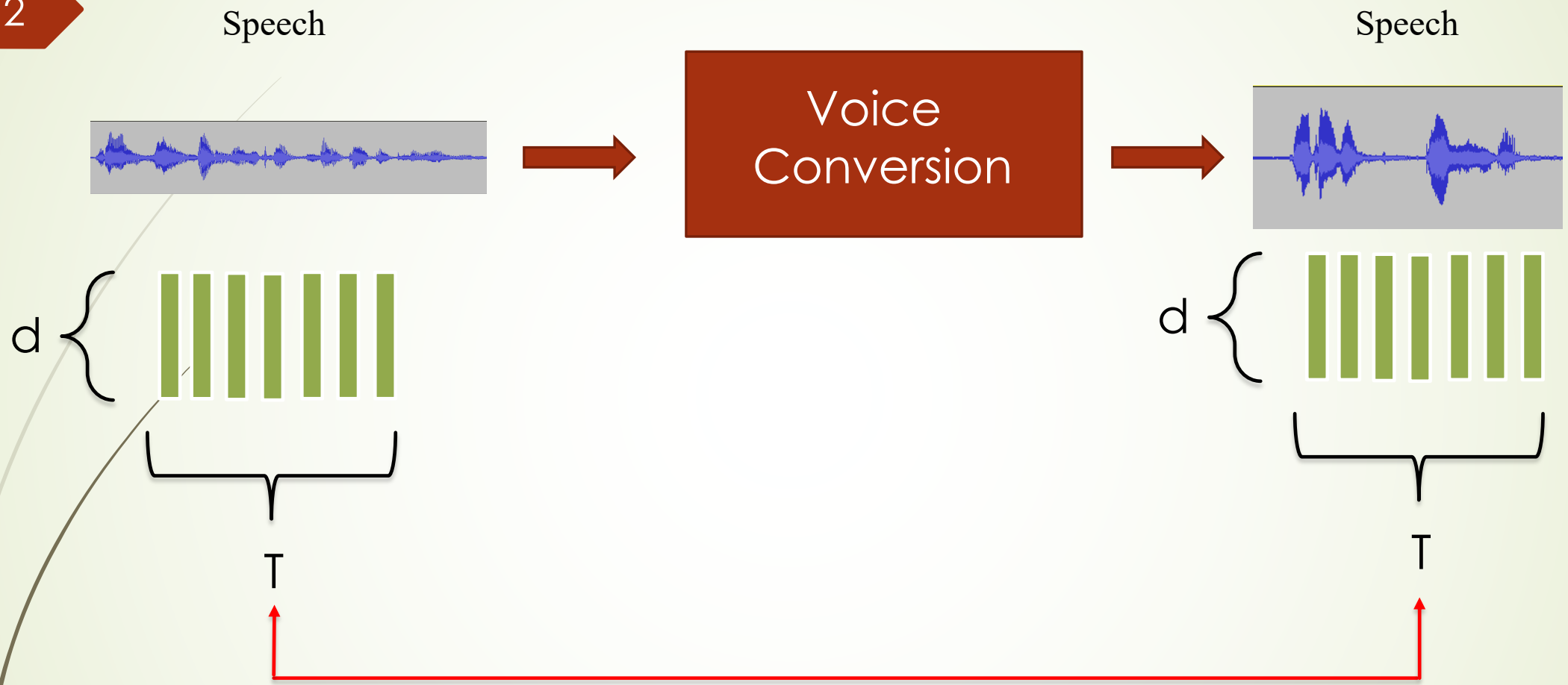
# Multi Target voice conversion and cross-language



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?

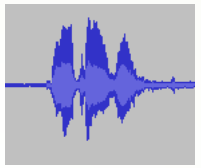
內容

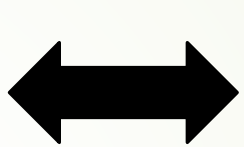
許多都可以，例如語者

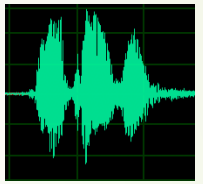
https://www.kocpc.com.tw/archives/182613

# Voice Conversion 技術發展分類

with平行語料







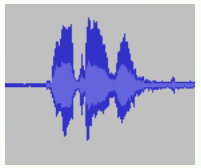
你好

source

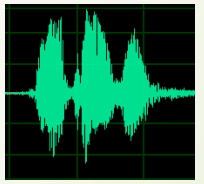
你好

target

without平行語料



你好

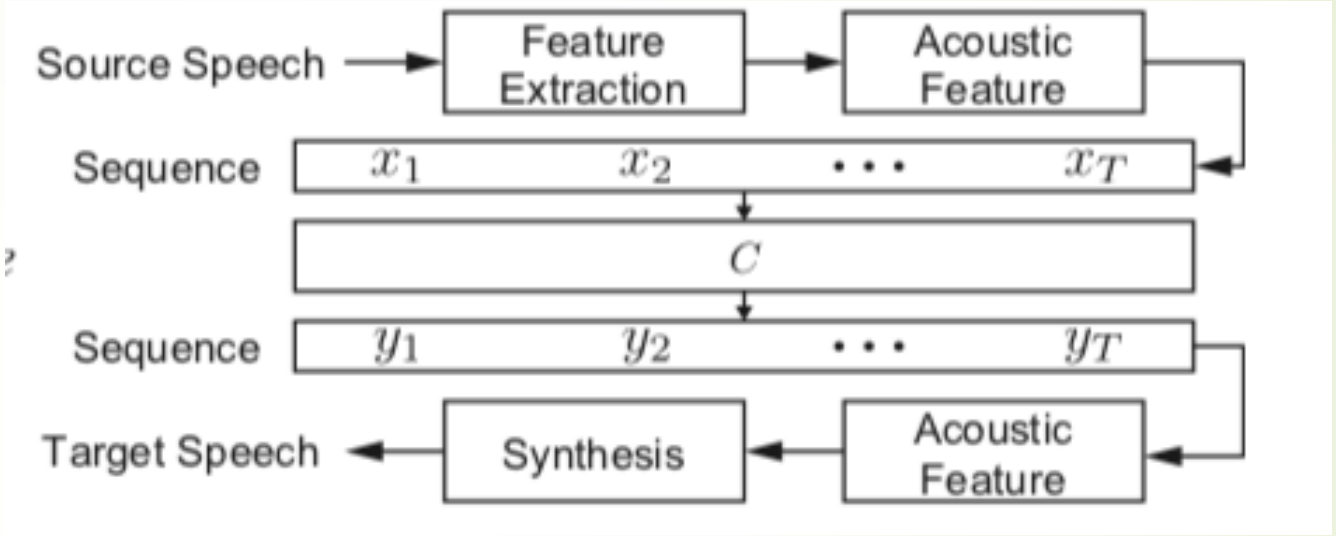


天氣不錯

平行語料

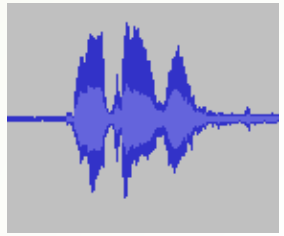
非平行語料

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

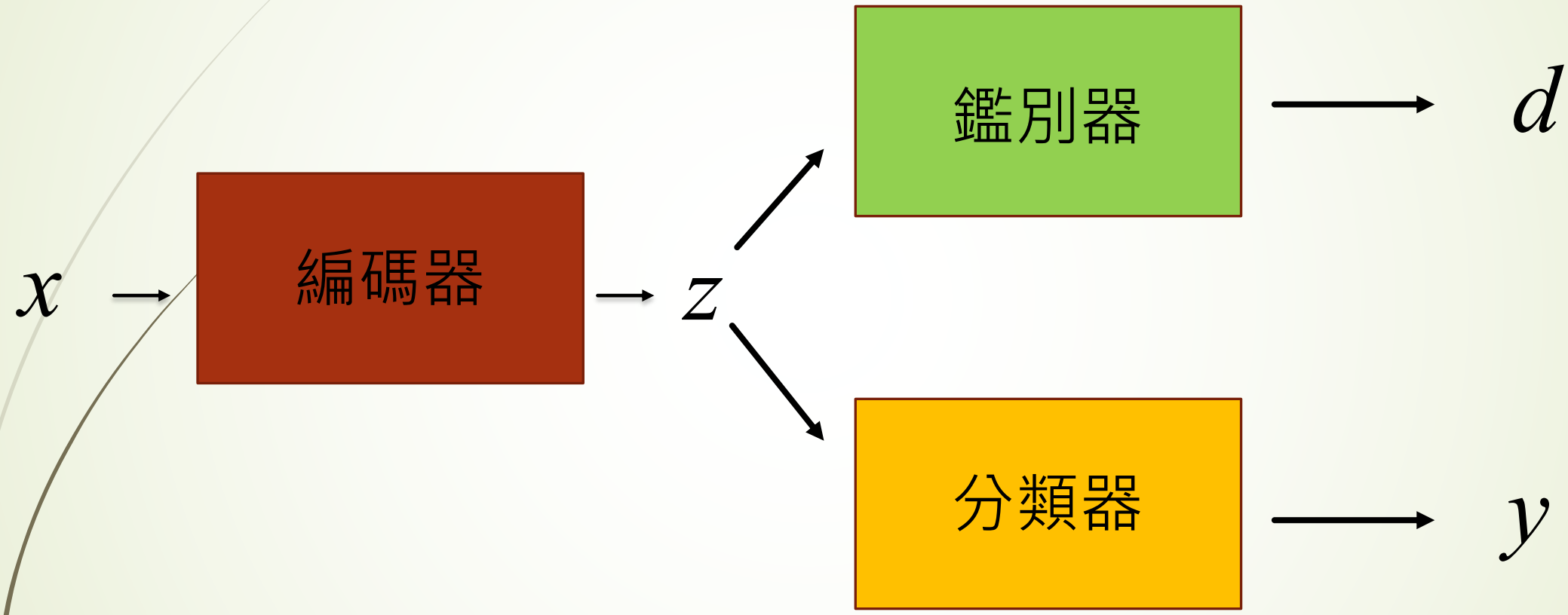
# 非平行語料

Feature Disentangle (特徵解纏)

Direct Transformation (直接轉換)

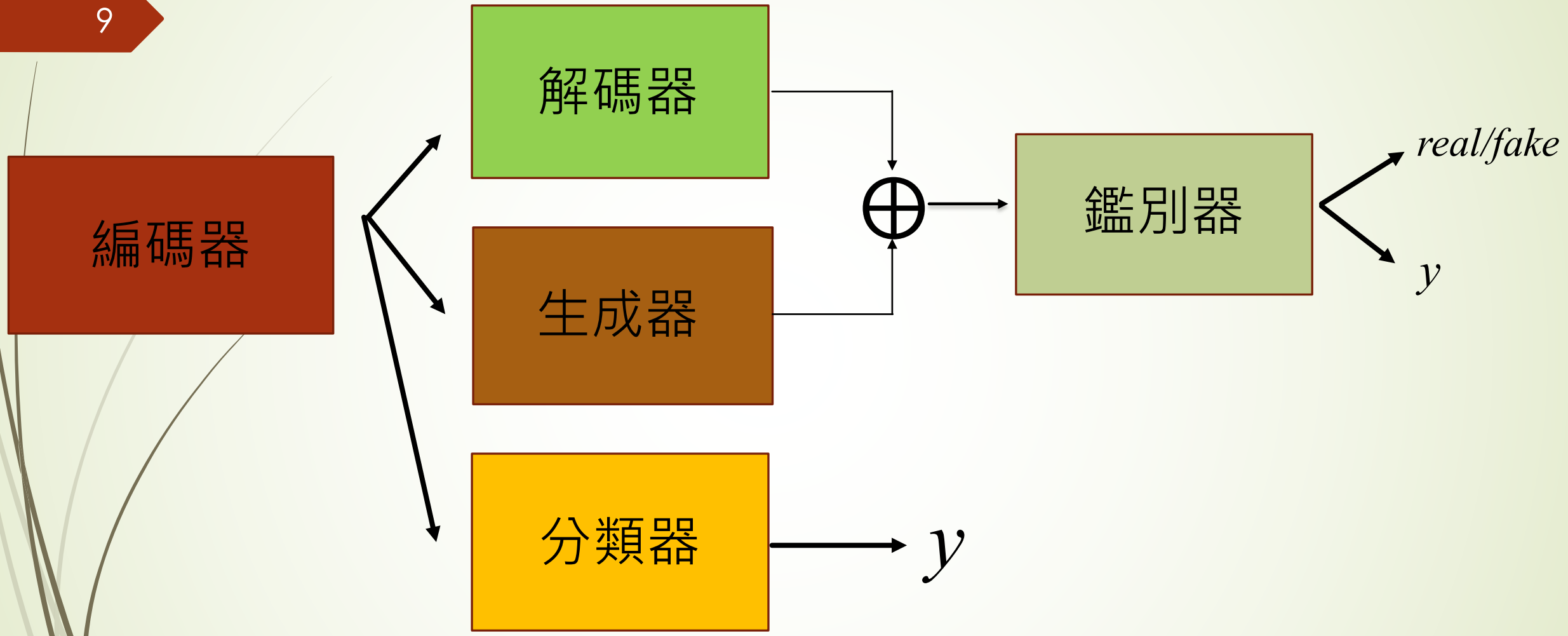
# 使用對抗訓練方法進行解纏特徵學習






x為資料，y為分類器的輸出，z為學習之潛在特徵

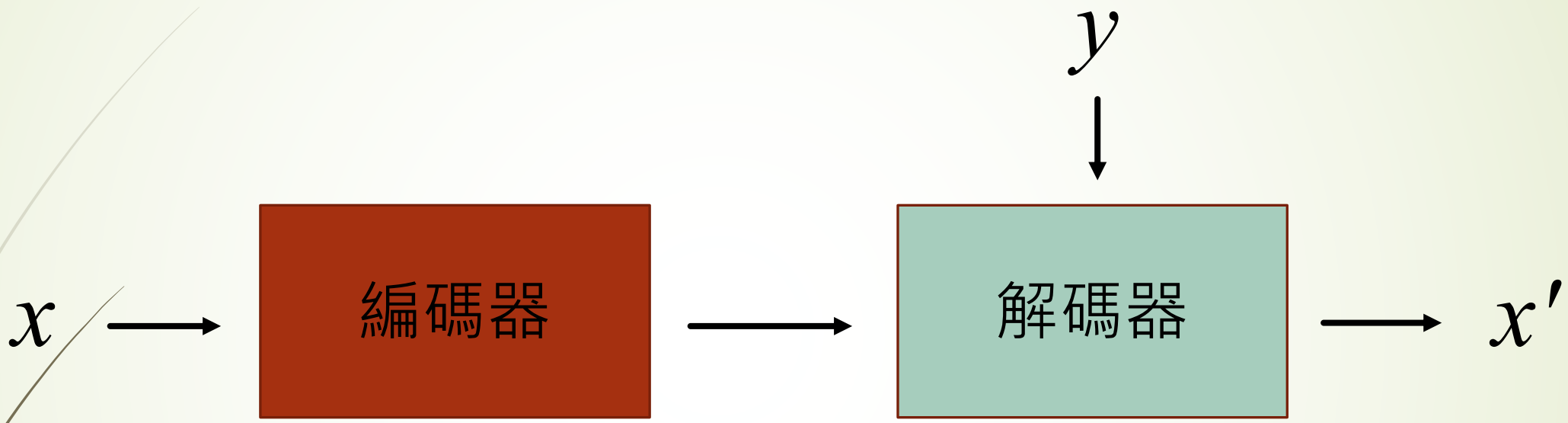
# 如果只有非平行語料，又想實現多目轉換




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。

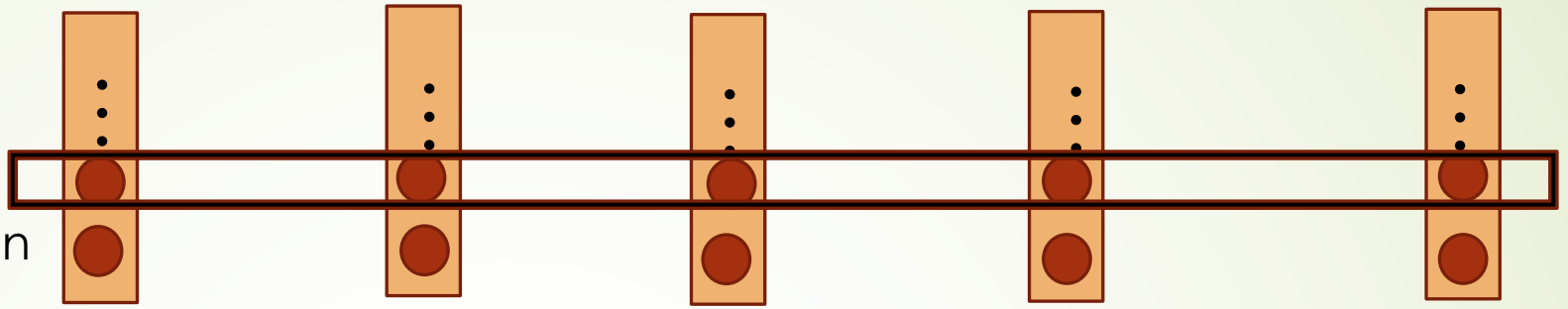



預訓練階段示意圖。在這個階段視同訓練一個語音訊號的自編碼器，但會同時在解碼器輸入語者的編號。x為語音訊號，y為語者編碼

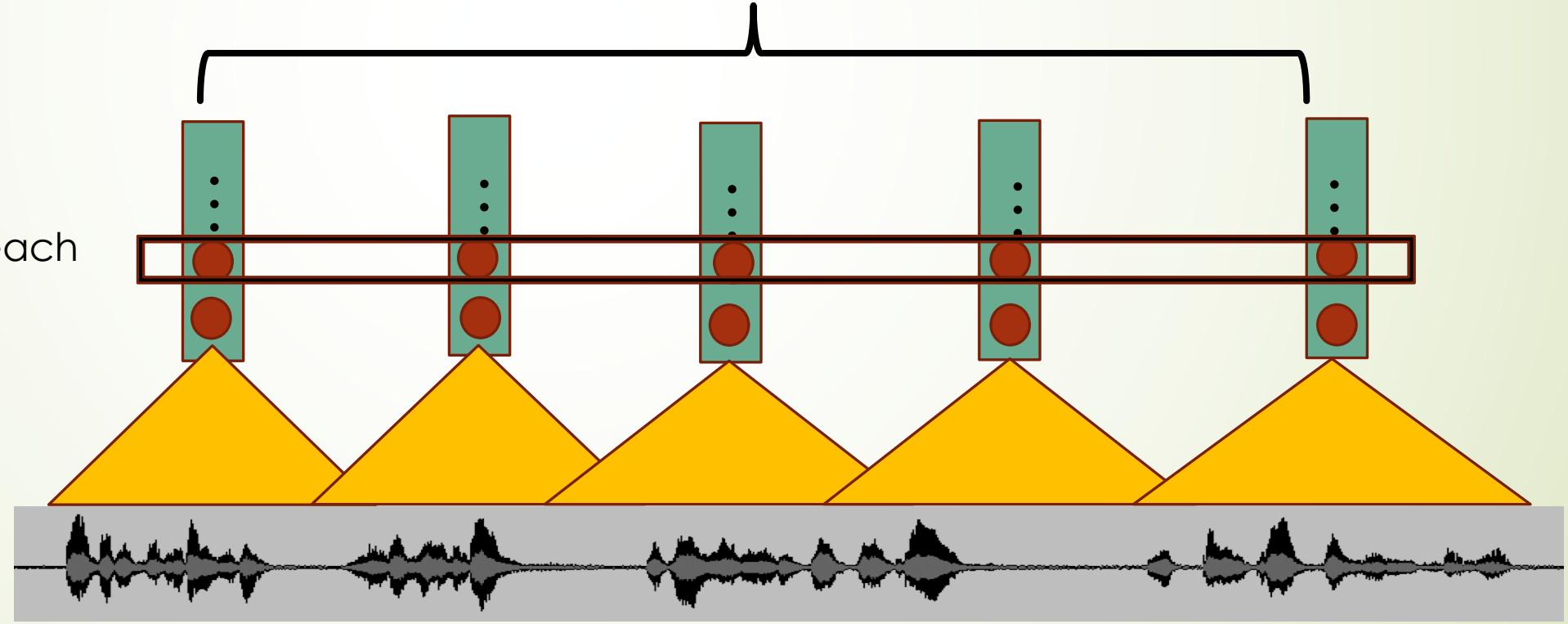
# instance normalization



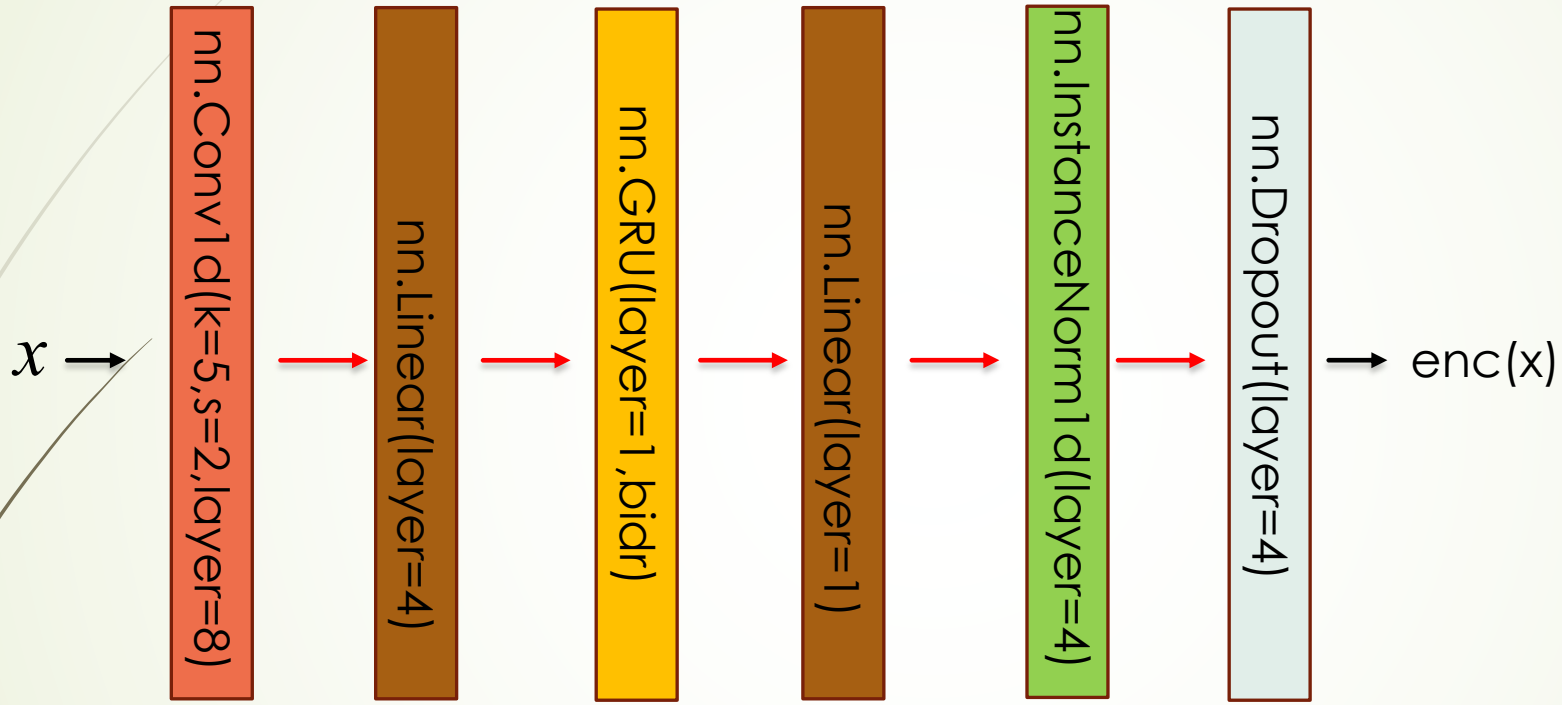
Each channel has zero mean and unit variance




Normalize for each channel

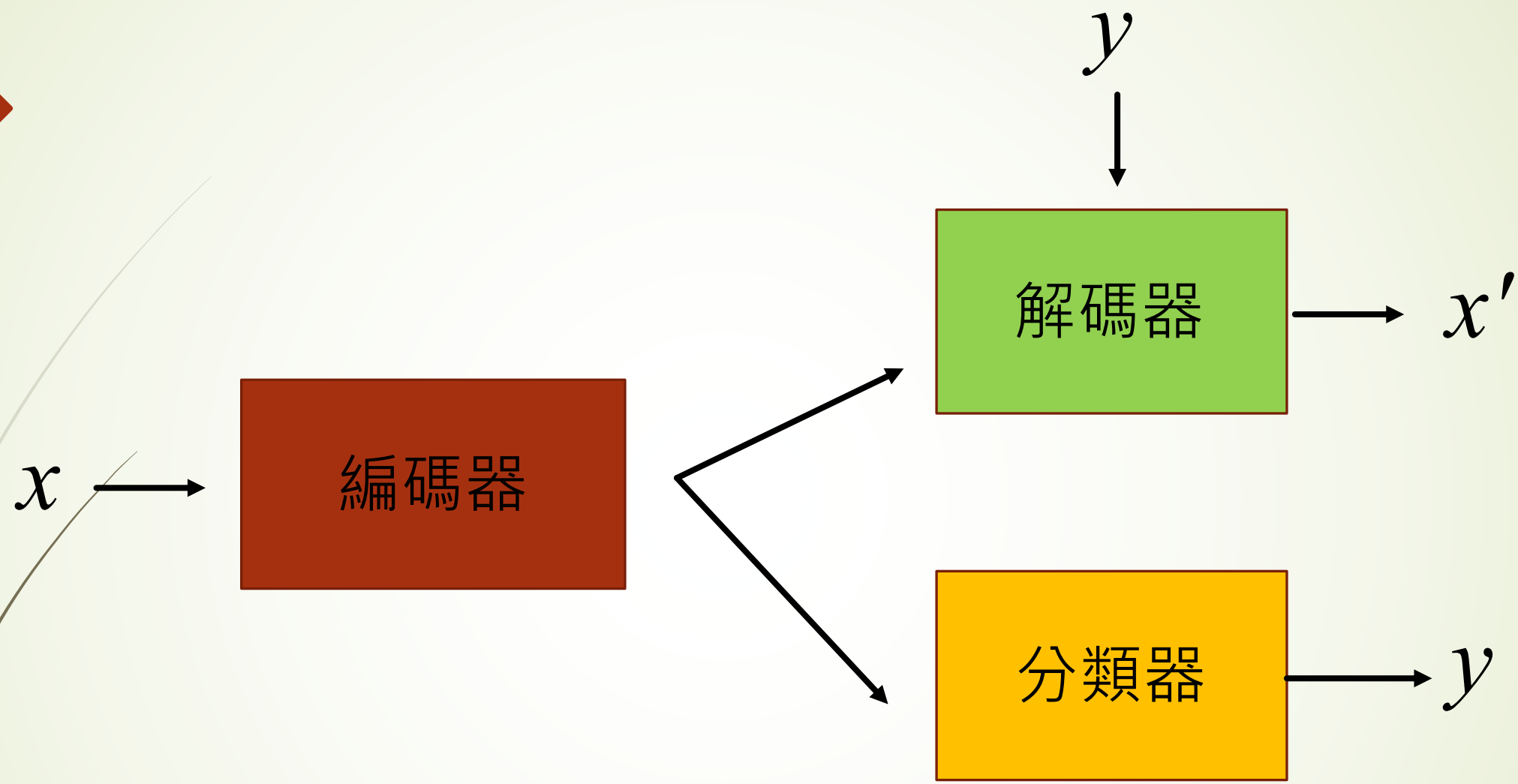


# 編碼器


x
nn.Conv1d(k=5,s=2,layer=8)
nn.Linear(layer=4)
nn.GRU(layer=1,bidr)
nn.Linear(layer=1)
nn.InstanceNorm1d(layer=4)
nn.Dropout(layer=4)
enc(x)

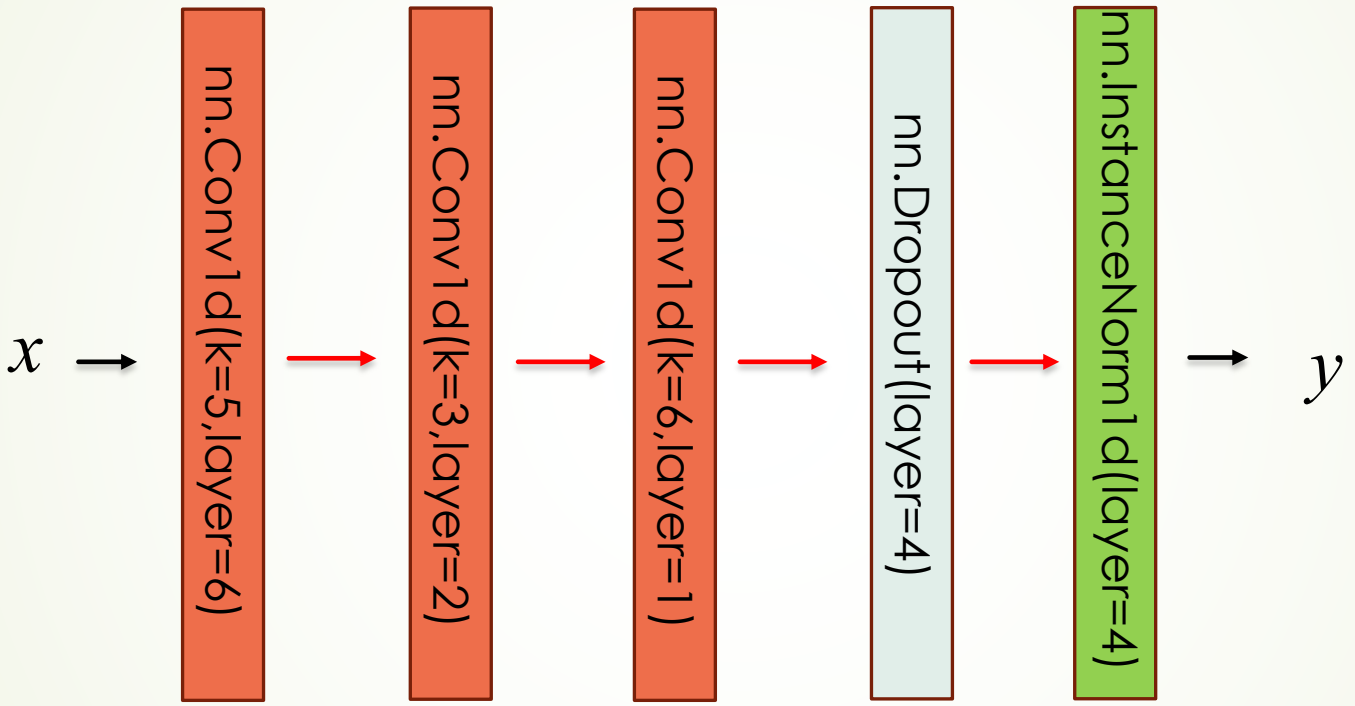
解碼器

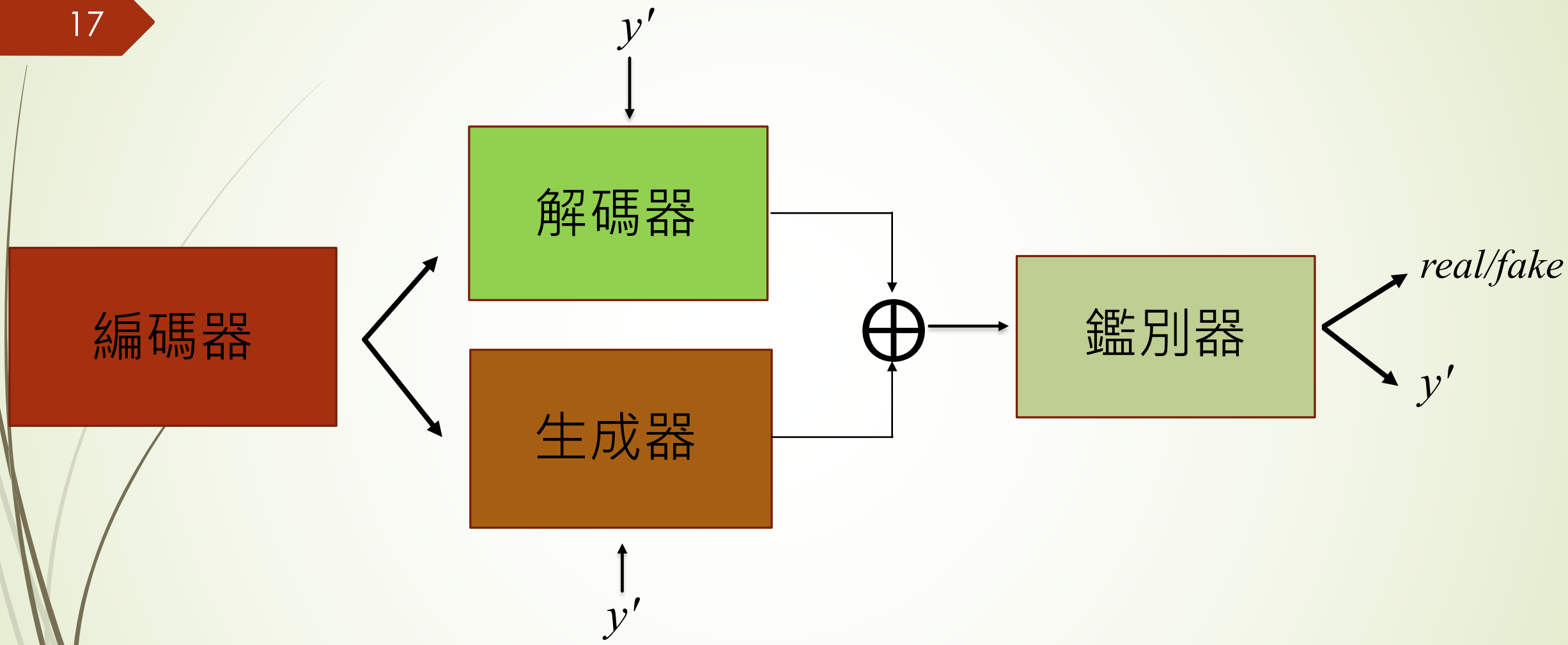

enc(x)
nn.Conv1d(k=3,layer=6)
nn.Linear(layer=4)
nn.GRU(layer=1,bidr)
nn.Linear(layer=1)
nn.InstanceNorm1d(layer=4)
nn.Embedding(layer=4)
$x'$




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

# 輔助分類


x
nn.Conv1d(k=5,layer=6)
nn.Conv1d(k=3,layer=2)
nn.Conv1d(k=6,layer=1)
nn.Dropout(layer=4)
nn.InstanceNorm1d(layer=4)
y




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

# 鑑別器


x
nn.Conv2d(k=5,s=2layer=5)
nn.Conv2d(k=1,layer=1)
nn.Conv2d(k=(17,4),layer=1)
Conv_classify(k=(17,4),layer=1)
nn.Dropout2d(layer=6)
nn.InstanceNorm2d(layer=6)
real/fake
y'

VCTK Corpus
+
自製音檔
( 小學老師&蔡英文 )

**preprocess**

**train**

**test**

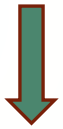
1. make_dataset_vctk.py

產生.h5檔案

2. make_single_samples.py

產生我們所指定語者index.json檔案

main.py



產生以training_step
為名的binary model檔案
檔案中包含decoder,
Encoder, generator
info.

test_back_one_ver.py

Input:來源與者內容(.wav)
Target:在此選擇
Output:轉換後檔案(.wav)
必要檔案：訓練好的model檔，
Vctk.json

# 資料庫描述

- VCTK Corpus
- 共109位語者，44257 utterances，48KHz mono 14.1GB

- 自有資料(48KHz, mono)
- 台語資料：陳豐慧老師02:33:38(885.2MB)，淑琪老師01:12:55(420MB)
- 國語資料：蔡英文總統01:13:51(42536MB)

- 在國語資料蔡英文總統的部分，我們首先在youtube收集蔡總統的語音，如元旦演講或是就職演說，這些收集完後我們對原本的音檔進行一些處理，先將該檔案處理成48KHz，mono。接著我們使用一套專業軟體將原本裡面的過多的殘響去除，去除然後我們使用audacity進行人工手動分割音檔，尋找話與話之間的間隔來斷句。
- 在台語資料這個部分我們也使用audacity進行人工手動分割音檔，尋找話與話之間的間隔來斷句

# DEMO

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