Speaker-Diarization 語者自動分段標記

指導老師：洪維廷

組員：1070423 鄭元富、1070402 劉品彥

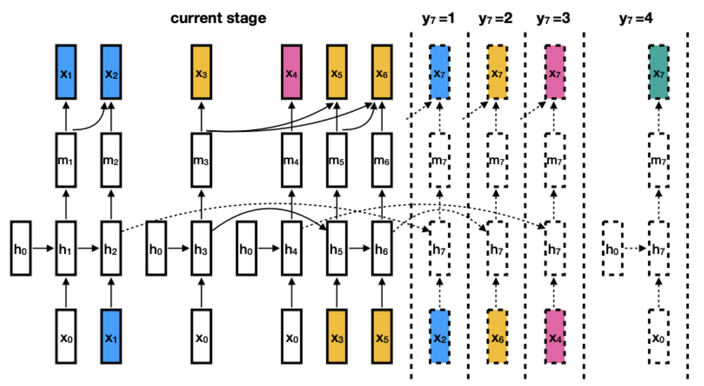
一學期電機系乙組畢業專題展

簡介

- 語者自動分段標記（Speaker Diarization）可以從一段語音中辨識不同的說話者以及他們說話的片段，並用編寫日記的方式(Diarize)將結果輸出。
- 為了解決 who spoke when 的問題，大部分的說話者辨別都有多個相對應的模型，包含但不僅限於：
  1.分割模型(Speech segmentation module) 通常去除沒有說話的片段
  2.提取模型 (Embedding extraction module) 提取說話辨別的特徵
  3.聚類模型 (Clustering module) 決定有多少說話者
  4.重分割模型 (Resegmentation module) 優化聚類結果來提升說話人分類的精度。

理論

- Unbound Interleaved-State Recurrent Neural Networks(UIS-RNN) ：UIS-RNN模型以RNN取代Cluster來預測說話者，而RNN為監督式學習，代表我們可以藉由增加訓練集來提升訓練效果，達到資料訓練的最佳化。
- UIS的U代表Unbounded，不需要提前知道說話者的數量，可由模型的訓練得到。
- UIS的IS代表Interleaved-state，不同的說話者有不同的RNN狀態，在相同的時間軸上交錯運行。
- 在辨識之初，給定一個初始化的參數h0，經由第一個語音向量得到關於第一位說話者的隱藏層參數h1，輸出時同時給第一位說話者標記顏色，當後面出現之前出現的說話者時更新該說話者的隱藏層參數，如果出現新的說話者時重新訓練得到新的隱藏層參數。
- 當有一段新的語音輸入的時候，分別以當前的說話者的代表隱藏層計算說話者變換的機率。
- 以 y7 預測為例，這時候會有三大種情況：仍然是小黃說話、變回小藍或小紅說話還是一個全新的說話者小綠在說話。模型分別計算這三種的機率值，取最大的為結果。

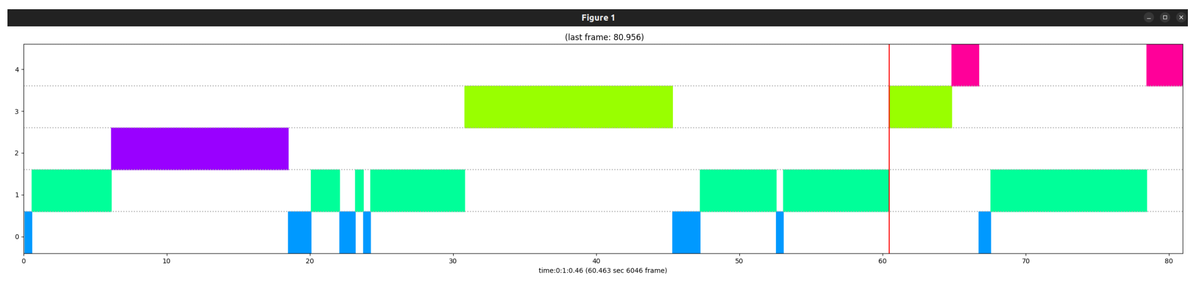


- UIS-RNN流程圖，顏色表示說話者語音片段的標籤，y7可能會有四種輸出可能

實驗結果

```
========= 0 =========
0:00.64 ==> 0:00.564
0:18.496 ==> 0:20.72
0:22.72 ==> 0:23.168
0:23.732 ==> 0:24.232
0:45.317 ==> 0:47.233
0:52.553 ==> 0:53.53
1:06.716 ==> 1:07.536
========= 1 =========
0:00.564 ==> 0:06.128
0:20.72 ==> 0:22.72
0:23.168 ==> 0:23.732
0:24.232 ==> 0:30.808
0:47.233 ==> 0:52.553
0:53.53 ==> 1:00.437
1:07.536 ==> 1:18.420
========= 2 =========
0:06.128 ==> 0:18.496
========= 3 =========
0:30.808 ==> 0:45.317
1:00.437 ==> 1:04.801
========= 4 =========
1:04.801 ==> 1:06.716
1:18.420 ==> 1:20.956
```

▲ 圖一：自動分段標籤



▲ 圖二：依照影片時間將結果視覺化

#iNEWS最新 普丁出招反制西方制裁！"不友善國家"購買天然氣須以盧布支付 德法怒批敲詐勒索 並為俄羅斯天然氣斷供做準備 ｜【國際局勢】20220401｜三立iNEWS

```
Finished diarization experiment
Config:
  sigma_alpha: 1.0
  sigma_beta: 1.0
  crp_alpha: 1.0
  learning rate: 0.001
  regularization: 1e-05
  batch size: 10

Performance:
  averaged accuracy: 0.822673
  accuracy numbers for all testing sequences:
    0.779661
    0.866071
    0.786885
    0.641921
    0.905405
    0.802721
    0.857143
    0.747368
    0.851852
    0.882979
    0.711111
    0.758621
    0.682028
    0.888889
    0.794643
    0.602941
    0.708861
    0.716346
    0.786517
    0.835443
    0.880435
    0.731343
    0.831325
    0.913043
    0.762376
    0.862069
    0.758170
    0.816901
    0.859375
    0.711864
    0.800000
    0.884211
```

▲ 圖三：模型訓練結果

結論

- 我們使用Deep Neural Network(DNN)來提取frame level feature，DNN 架構是使用在頻域和時域都具有 convolution 的 2D CNN，再利用 Average Pooling Layers 聚合 frame-level feature vectors 達到固定長度的 utterance-level embedding ，比較不會受到無關雜訊的影響以便當做後續模型的輸入。
- 當使用相同的說話識別模型時，UIS-RNN 的性能比起一般的 Unsupervised Clustering Model ，處理說話辨識模型的複雜性更能有效提升效能，通過 Bayesian non-parametric process 自動學習每段話中 speaker 的數量，並通過 RNN 隨時間傳遞訊息，最後在測試集上也都有良好的表現。
- 處此之外，我們也嘗試使用語音增強 (Speech Enhancement) 將一段音頻的環境雜音去除，在語者自動分段標籤的任務階段，更能專注於人聲的辨識，減少環境雜音對辨識準確率的影響與模型的執行效能。
- 這項技術的目的在於從音頻流中分離出不同人說話的語音，並將分離出的語音歸併到所屬的說話人上，可以有效的應用於索引或分析各種類型的音頻數據，例如：來自媒體的廣播音頻、會議的對話、個人視頻等，亦或是社群平台或手持設備的個人視頻、法庭訴訟、商業會議、各種行業的收益報告等。

College of Electrical and Communication Engineering Yuan Ze University

本活動由教育部高教深耕計畫補助

元智大學電機通訊學院