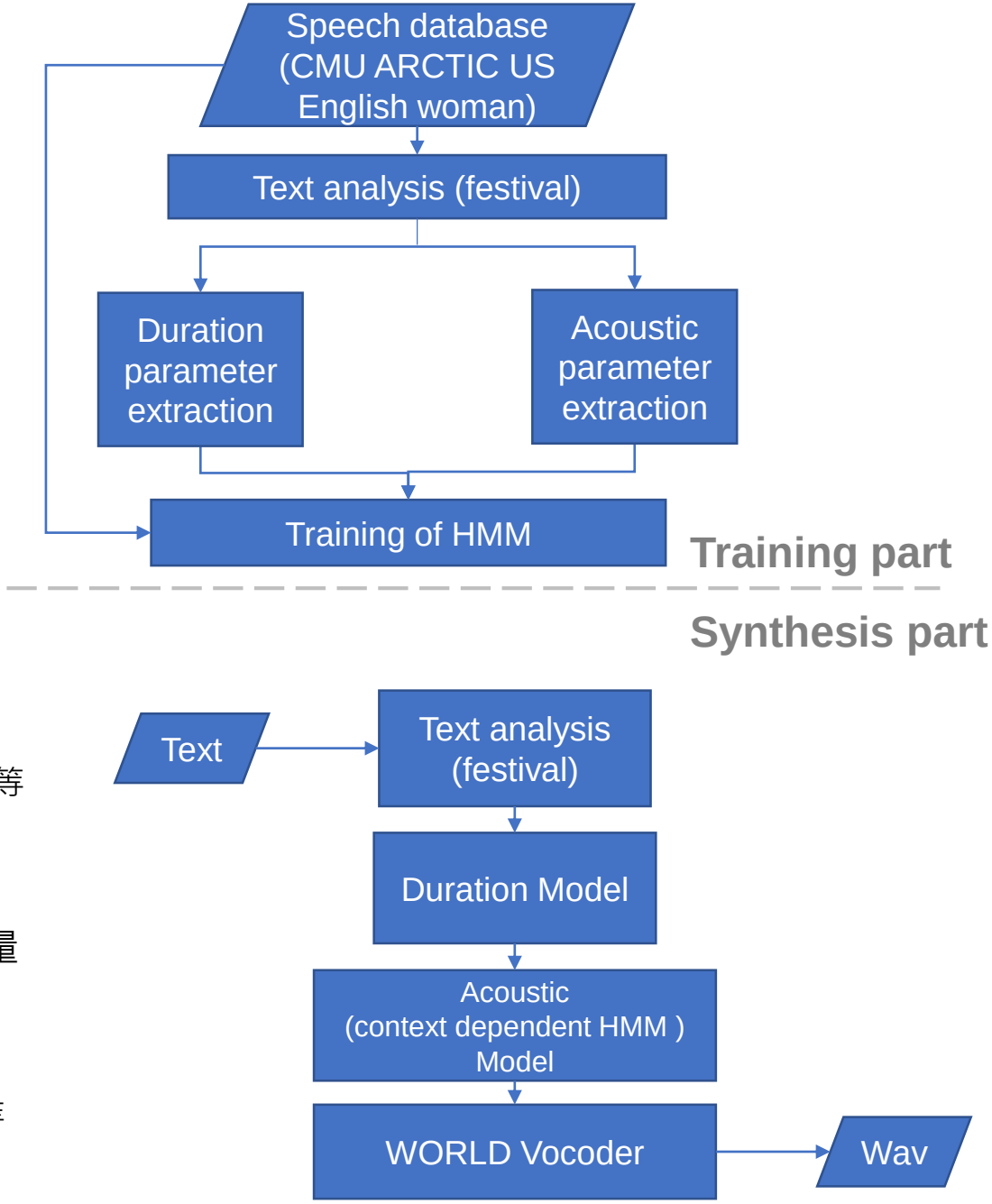
TTS block diagram with clear I/O & processing of each block

Sian-Yi Chen

Advisors : Tay-Jyi Lin and Chingwei Yeh

Outline

1. 訓練
   - 文本分析
     - input：語料庫、語料文本
     - output：訓練用的向量特徵
     - processing
       1) 生成HTS(HMM-based 語音合成系統)上下文標籤
       2) 透過問題集轉換成向量
   - 神經網路模型(持續時間與聲學模型)
     - input：皆為由問題集轉換的向量檔案
     - output
       1) Duration model：各個音素的持續時間
       2) Acoustic model：音色、音高、頻譜包絡(聲道的形狀)等等聲學特徵
2. 合成
   - 文本分析：對應文本生成HTS上下文(音素)標籤再轉換成向量
   - 神經網路
     - Duration model：預測每一個音素的持續時間
     - Acoustic model：加入音色、音高、頻譜包絡(聲道的形狀)等等聲學特徵
   - 聲碼器：合成波形



生成上下文標籤

範例句子：Author of the danger trail

Author

ao th er

第一個標籤

持續時間 x^x-sil+sil=ao 其餘資訊 狀態一

開頭沒有聲音，sil(靜音)

第二個標籤

持續時間 x^x-sil+sil=ao 其餘資訊 狀態二

⋮

第六個標籤

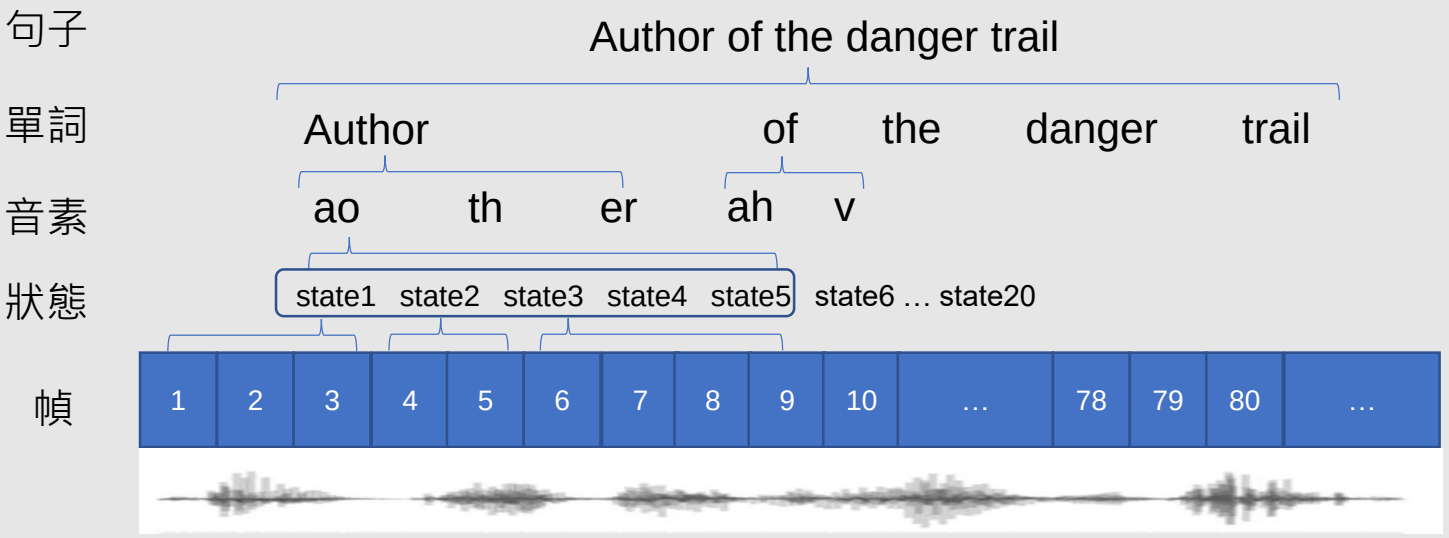
持續時間 sil^sil-ao+th=er 其餘資訊

前前音素 **前音素** **當前音素** **下一個音素** **下下一個音素**

標籤格式(音素、音節、詞語、短語、句子之間的關係)

2050000 2400000 sil^sil-ao+th=er@1_2/A:0_0_0/B:1-1-2@1-2&1-7#1-4$1-3!0-2;0-4|ao/C:0+0+1/D:0_0/E:content+2@1+5&1+2#0+3/F:in_1/G:0_0/H:7=5@1=2|L-L%/I:7=3/J:14+8-2

前後音素、該音素在音節中的位置
前一個音節是否為重音、音素數量
當前音節重音、音素數量、在單詞中的位置、在短語中的位置...等
下一個音節是否為重音、音素數量
前一個單詞詞性、音節數量
當前單詞詞性、音節數量、在短語中的位置、單詞數量、距離...等
下一個單詞詞性、音節數量
前一個短語中的音節數量、單詞數量
當前短語中的音節數量、單詞數量、在語句中的位置
下一個短語中的音節數量、單詞數量
此話語中的音節、單詞、短語數量



HMM 以5個狀態對齊示意圖

問題集

QS代表問題集，雙引號中間為問題名稱，大括弧內容則為問題的內容。
問題中包含前後聲韻母為何？韻律？位置？詞性？聲調？位置？特徵劃分等等
對每一個標籤詢問此問題集的每一題，因此每一個標籤就會有一個416維的向量。

問題一：QS “C-Vowel”　{-aa+}

詢問第一個標籤是否有此元音

label 持續時間　x^x-sil+sil=ao　其餘資訊　狀態一

0 …

416維的向量

問題集，問題題數就是向量化的維度，此檔案包含416個問題

Neural Network Architecture

Input: 416 dimensions label binary file
Duration model (DNN): 4*512 (tanh)

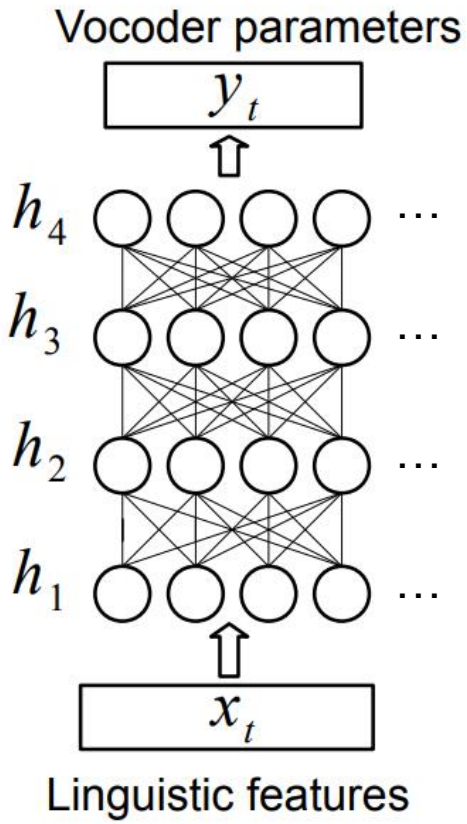
- **Output: 預測出每個音素 5 個狀態的持續時間**

- Batch size: 256
- Learning rate: 0.002
- Train file number: 50
- Valid file number: 5
- Test file number: 5

Acoustic model (DNN): 4*512 (tanh)

- **Output: mgc: 60維; bap: 1維; lf0: 1維;**

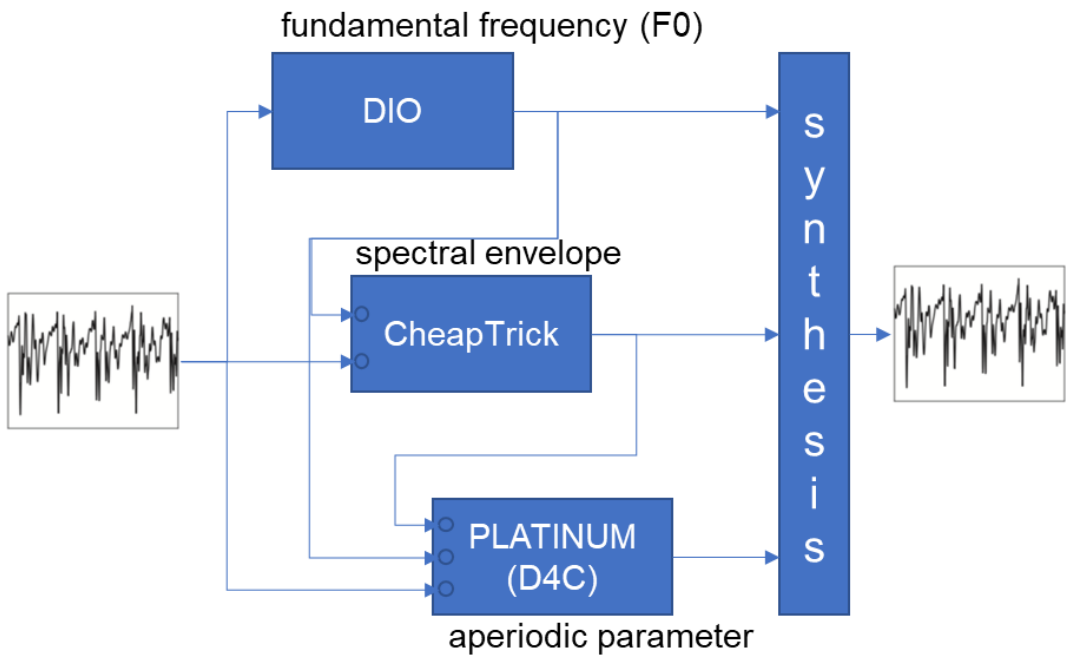
- Batch size: 64
- Learning rate : 0.002
- Train file number: 50
- Valid file number: 5
- Test file number: 5



圖一：前饋神經網路(DNN)

WORLD Vocoder I/O 與物理意義

- Vocoder I/O：聲碼器使用的是WORLD Vocoder
  - 輸入
    1. 每一個frame的基頻(fundamental frequency, f0)
    2. 頻譜包絡(spectral envelope)
    3. 頻段非週期比值(band aperiodicity)
  - 輸出為結合上述資訊合成音檔

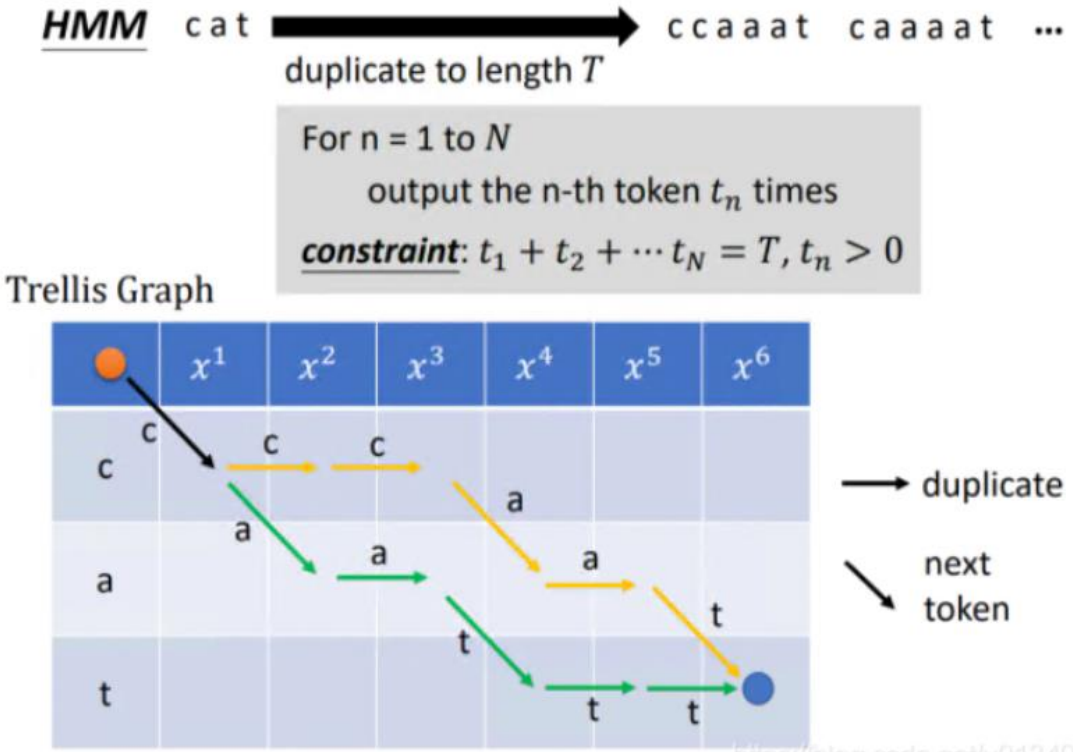


- 名詞的物理意義
  - **基頻：**一個發聲體發聲時，聲音由許多正弦波所組成，其中頻率最低的就是基音，其他較高的皆為泛音，**基音主要用來區分音高，而泛音則決定音色**。
  - **頻譜包絡：**將不同頻率的振幅最點連結起來形成的曲線，可以藉由此來顯示**聲道的形狀**，也就是聲學的特徵。
  - **頻段非週期比值：**可以透過比值控制週期(氣流衝擊聲帶震動產生的濁音)與非周期(聲帶鬆弛不震動的清音)，一般的語音都是有周期信號和非週期信號組成，所以，除了獲取週期信號的參數，我們還需要得到其中的非週期信號參數，才能完美的合成原始信號。

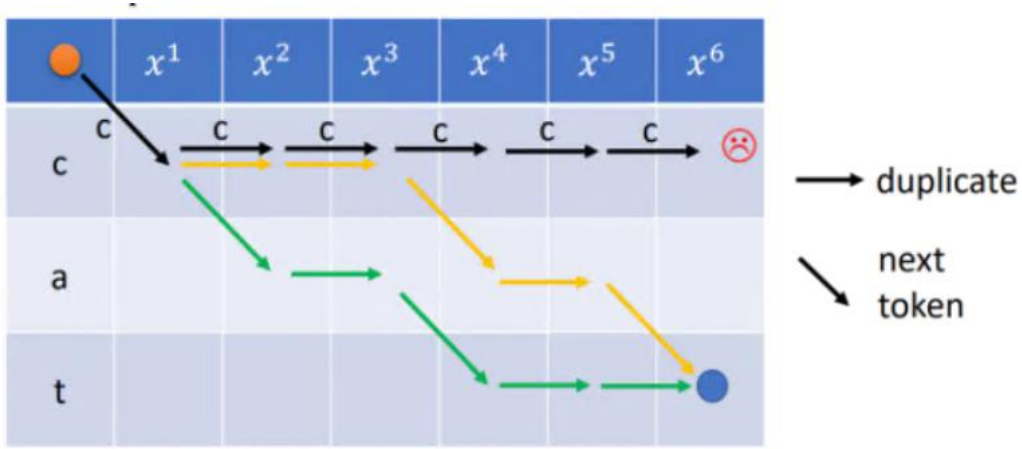
附錄

HMM alignment

HMM (Hidden Markov Model)



一個token可以重複N次，但是所有token重複的次數和要等於acoustic features的長度T，也就是灰色部分所描述的公式。
表中橫軸代表acoustic features，縱軸表示token。
從左上角開始走到右下角，每步只能有兩個方向：向右走或是向下走，但終點一定要走到右下角才算是合法的路徑。
從起點一直走到終點所有合法的路徑就是所有可能的alignment。



WORLD Vocoder 演算法與spectral envelope和MFCC之間的關係

- 演算法
  - 提取F0的演算法(DIO)：通過低通濾波器對原始訊號進行基頻的提取，具體流程是取4個週期計算標準差，並選最低的作為基頻

  - 頻譜包絡Spectral Envelope：有三種方法可以取得，LPC、Cepstrum、CheapTrick
    - LPC：一個語音可以用多個語音過去值的加權線性組合來逼近
    - Cepstrum：訊號->FFT->絕對值->對數->相位展開->IFFT->倒頻譜
    - CheapTrick：音高同步分析
      1. F0-adaptive windowing：語音分段不以frame為單位，以f0對應的週期為單位，以保證波形和頻譜的平滑連續，使用hanning window
      2. smoothing of the power spectrum：對時域訊號做FFT，並在三角窗內對訊號進行平滑
      3. liftering in the quefrency domain：將功率頻譜看做是普通訊號，求出訊號的包絡就是找到其低頻
         1) 對功率頻譜做IFFT
         2) 過濾訊號的到低頻
         3) 頻譜恢復：消除先前平滑帶來的變異
         4) 取得包絡

  - 非週期信號參數：一般的語音都是有周期信號和非週期信號組成，所以，除了以上獲取週期信號的參數，我們還需要得到其中的非週期信號參數，才能完美的合成原始信號。

- spectral envelope和MFCC之間的關係
  - 作者表示WORLD在取得Mel-cepstral與常見方法不同
  - 一般會經過FFT再經過三角濾波器，因為MFCC是在頻譜圖上進行，因為未經過平滑處理，所以需要濾波器。
  - WORLD是在頻譜包絡上進行，已經有平滑處理，所以流程上看似沒有濾波器但使用效果卻與常見方法想同。

Synthesis voice by WORLD Vocoder

Synthesis part