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

Sian-Yi Chen

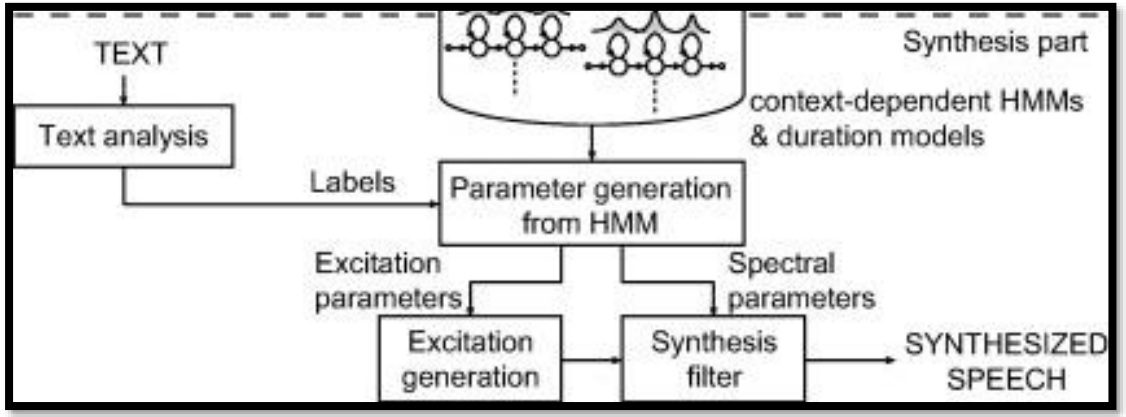
Advisors : Tay-Jyi Lin and Chingwei Yeh

Outline

傳統 TTS：statistical parametric synthesis 統計參數合成 [1]
其中可以分為訓練及合成部分，在合成階段中，首先，輸入文字，將給定的單詞轉換成上下文相關的標籤序列，然後透過決策樹去挑選對應的 HMM 模型，接著根據這個 HMM 模型使用語音參數生成演算法生成頻譜與激勵訊號參數，最後透過梅爾對數頻譜近似濾波器 (Mel Log Spectrum Approximation filter, MLSA filter) 生成語音訊號。

合成語音流程

- 從文字如何轉換成上下文標籤
- 標籤如何與 frame 對齊
- 使用 CART 挑選 HMM 模型
- 生成頻譜 (MFCC) 與激勵 (F0) 參數
- 使用 MLSA 生成波形



Spectral parameter: 梅爾到頻譜係數(MFCC)及其動態特徵
Excitation parameter: logF0及其動態特徵
動態特徵是靜態特徵(MFCC)的線性變換
如果輸入是一向量 c_t
動態特徵則是 $c_t - c_{t-1}$

[1] H. Zen, K. Tokuda and A. W. Black, “Statistical parametric speech synthesis,” *speech communication*, Vol. 51, pp.1039-1064, 2009.

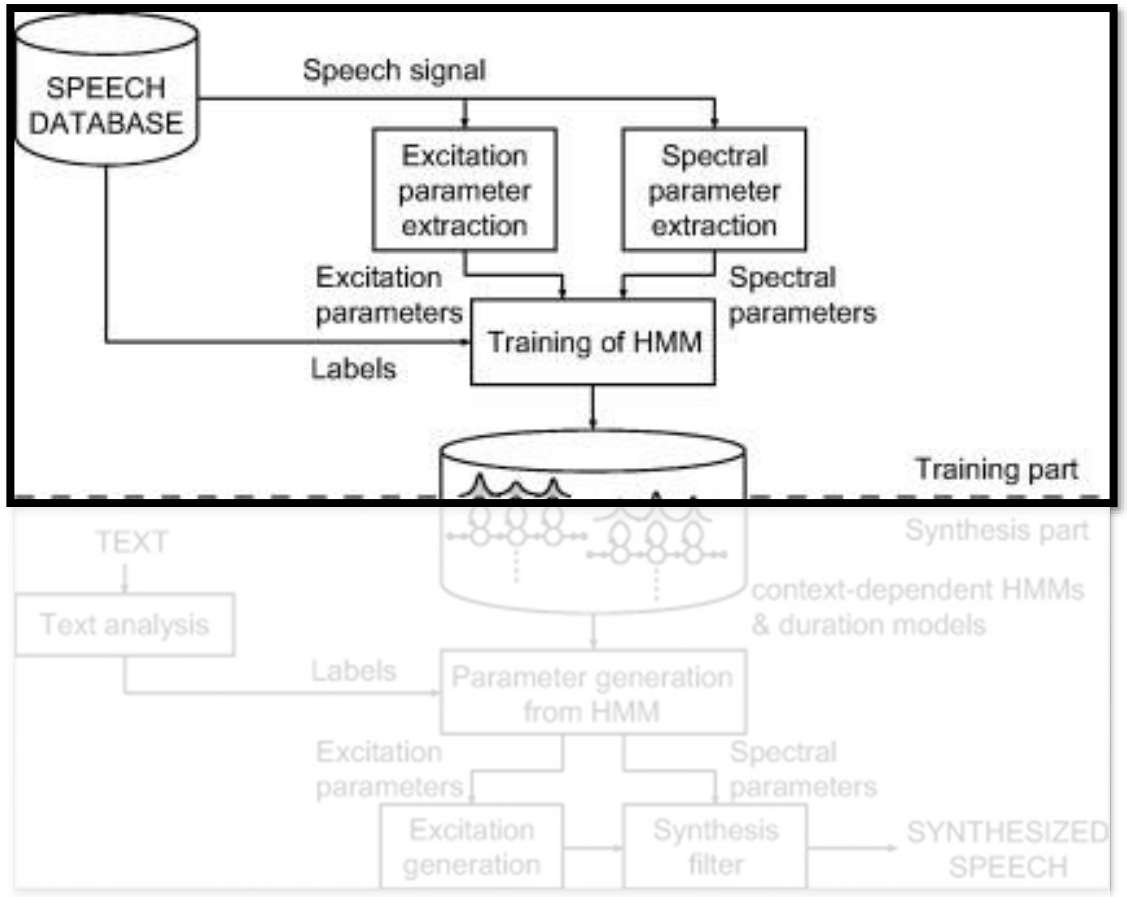
HMM 訓練部分

在訓練部分，從語料提取頻譜參數 (MFCC以及動態特徵) 和激勵參數 (logF0以及動態參數) 並搭配相對應的文本分析器產生上下文標籤，再配合適當的文脈相關問題集，訓練決策樹，最後產生與文脈相對應的HMM模型

其中頻譜、激勵、持續時間參數皆使用決策樹個別聚類，因為它們個自具有上下文依賴性。

動態特徵 (delta) 主要功用為連接每一個HMM模型時變得平滑解決分段問題。

取log目的是為了在計算HMM時，因為機率都是小餘一的數字，當連乘次數越多，有可能造成溢位，因此避免連乘到最後變成0的問題。



Context-dependent label

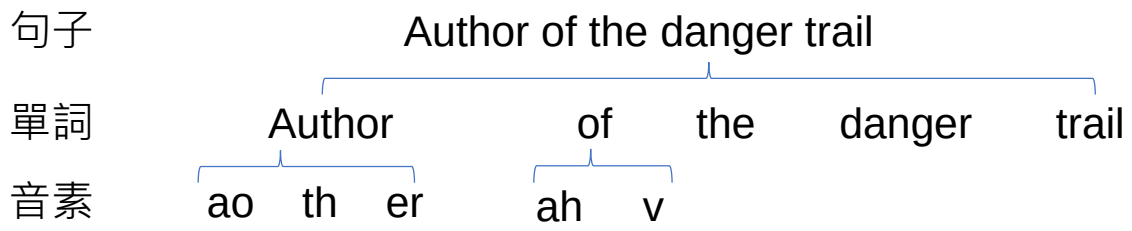
上下文標籤：輸入一段文字，解析其中音素、音節、單詞、短語以及發聲數量之間的關係。

- 音素：之前、現在、之後的音素、位置以及子母音類別等
- 音節：之前、現在、之後的音節個數、重音、位置等
- 單詞：之前、現在、之後的詞性、位置、個數等

以 “Author of the danger trail.” 這段話作為舉例，它會將此句子拆分到音素單位，如右下圖所示，標籤就會依照多少音素生成多少標籤。
其餘資訊，像是位置，子母音為何則緊接在後方，詳細格式可參考附錄

第一個標籤：sil ao th 其餘資訊
第二個標籤：ao th er 其餘資訊
第三個標籤：th er ah 其餘資訊
第四個標籤：er ah v 其餘資訊
.
.
.

上下文音素標籤



句子拆分至音素

文脈相關問題集

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

問題集，問題題數就是向量化的維度，此檔案包含416個問題

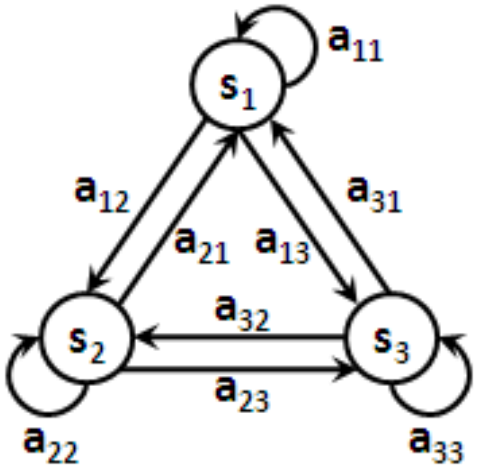
Hidden Markov Model (HMM)

Markov Model：選一個狀態作為起點，然後沿著邊隨意走訪任何一個狀態，會一直走並沿途累計從起點該點的機率。

S (狀態) = $\{S_1, S_2, S_3\}$

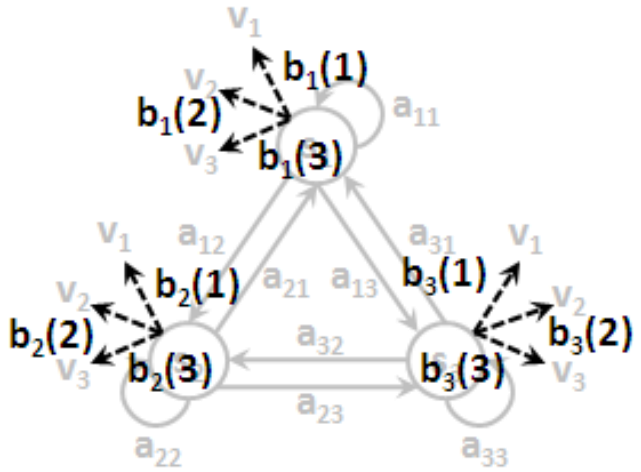
A (轉移機率) = $\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$

Π (起始機率) = 可以取任一點作為起點，機率總和為1



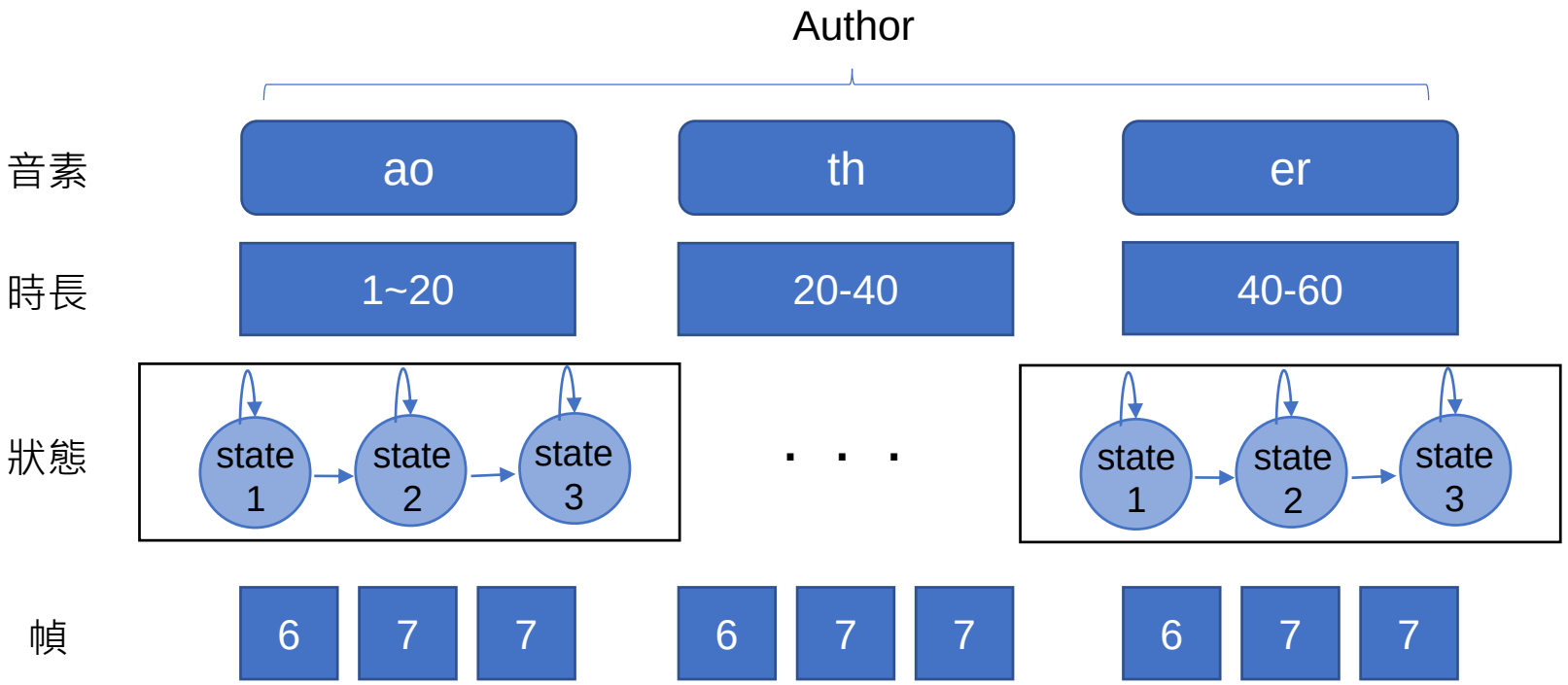
Hidden Markov Model：隱藏馬可夫模型添加了新的要素，每造訪一個狀態就會出現一個新的值 (v)，而每一個新出現的值都有不同的機率 (b)。

舉例來說今天有一位醫生要判斷病人是健康的還是發燒，病人只會回答正常(S_1)、頭暈(S_2)、冷(S_3)，醫生要從這3個答案中判斷是否發燒，是否發燒就是隱藏狀態(無法直接觀察到)；發燒(v)的機率(b)。

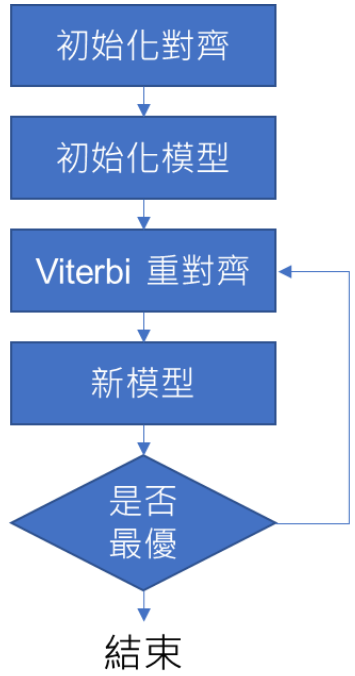


透過HMM將音素與 frame 對齊

單詞可以透過CMU發聲字典轉換成音素，但一開始不知道一段語音的哪些幀對應哪些狀態，因此進行**初始化對齊**，也就是將一段時長平均分配，假設 author 這個詞發聲 1.5 秒，若一個 frame 長 25ms，一次移動 25ms，則可以得到 60 個frame，也就是 “ao”、”th”、”er” 每個音素各對應至 20 個 frame，每個音素又由 3 個狀態所組成，因此每個狀態分配到 6 或 7 個 frame。



音素對齊frame示意圖



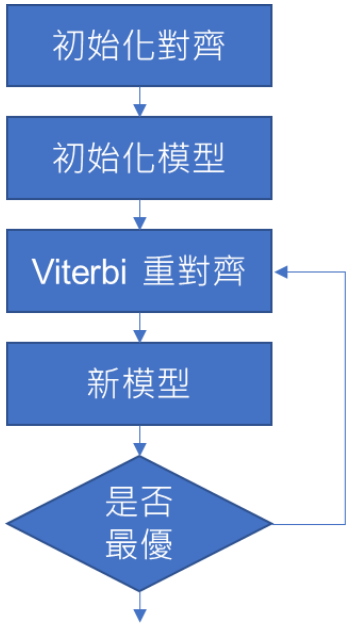
HMM 訓練流程

HMM 初始化模型

輸入：觀察序列 O=[o_1, o_2, …, o_x] (X frame 的 MFCC 特徵)
輸出：通過模型計算每一 frame 對於 “ao”、”th”、”er” 這 3 個音素的某一狀態 (3狀態)的機率

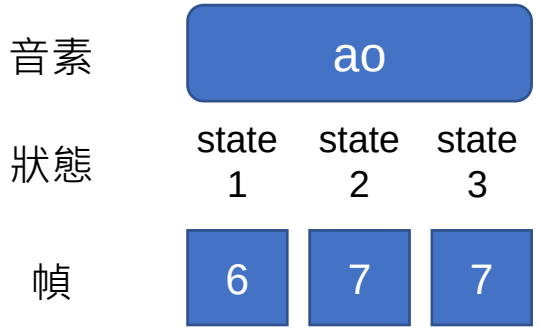
HMM模型 λ=(A,B,Π)

- 其中A是隱藏狀態轉移機率的矩陣
- B是觀測狀態生成機率的矩陣
- Π是隱藏狀態的初始機率分佈



HMM 訓練流程

- 初始化完就可得到轉移機率 A，計算轉移次數 (狀態1->狀態1，狀態1->狀態2)，轉移次數/總轉移次數 = 轉移機率
- 初始機率分佈 Π：HMM 模型是從左到右的模型，一開始在狀態1的機率為100%，所以此參數可忽略
- 狀態生成機率 B：一個狀態對應一個gmm模型，一個狀態又對應好幾個frame，所以好幾個frame對應一個gmm模型，初始化後，可得知狀態1對應6個frame，因此可以透過此計算狀態1的gmm模型 (單高斯模型)，求得平均值和變異數。



音素對應至frame

HMM 新模型

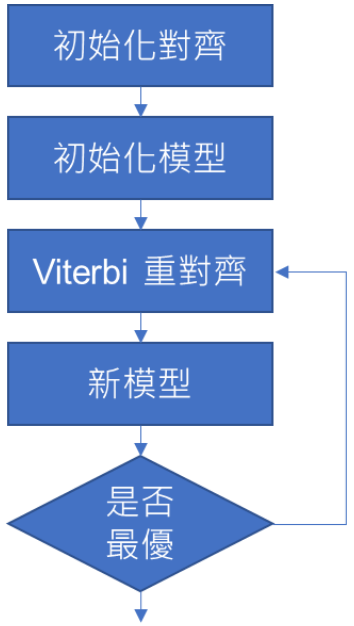
重新對齊

初始化結束需要重新對齊，使用的是viterbi演算法，根據初始化模型λ=(A,B,Π)來計算，記錄每個時刻的每個可能狀態最優路徑概率，同時記錄最優路徑的前一個狀態，不斷向後反覆運算，找出最後一個時間點的最大概率值對應的狀態。

Viterbi是一種動態規劃演算法。它用於尋找最有可能產生觀測事件序列的維特比路徑(隱含狀態序列)

反覆運算

透過重新對齊可以得到新的A(轉移機率)和B(生成機率)，就可以進行下一次的Viterbi演算法，尋找新的最優路徑，得到新的對齊，新的對齊繼續改變著參數A、B。如此迴圈反覆運算直到收斂，則GMM-HMM模型訓練完成。
反覆運算次數可以透過設定固定的迴圈數，也可以藉由觀察似然 (某件事發生的機率) 的變化，如果變化不大就結束。



HMM 訓練流程

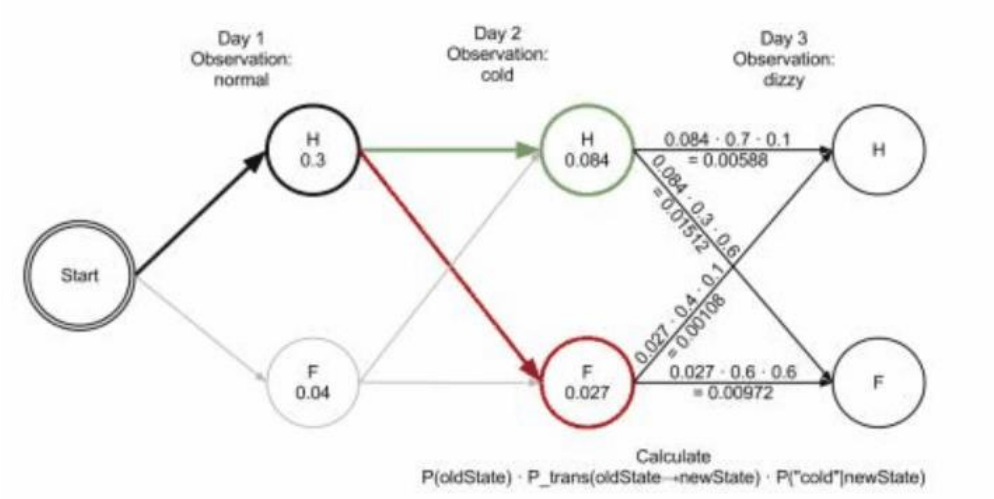
Viterbi演算法概念

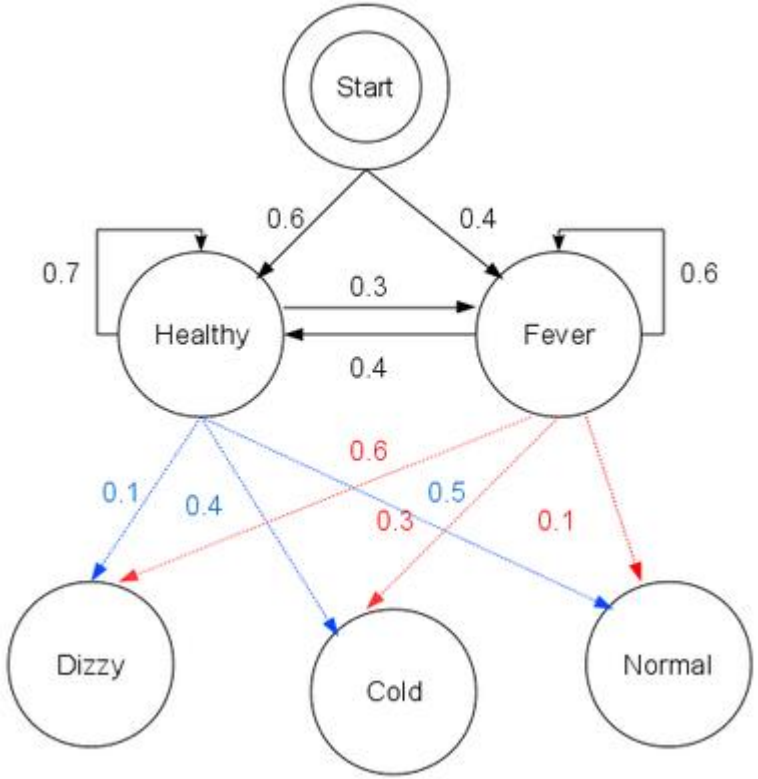
Viterbi是一種動態規劃演算法。它用於尋找最有可能產生觀測事件序列的路徑，以及其機率。

今天有一位醫生要判斷病人是健康的還是發燒，病人只會回答正常、頭暈、冷，醫生要從這3個答案中判斷是否發燒，是否發燒就是隱藏狀態(無法直接觀察到)
右圖為病人各狀態的機率：

- 當天健康的病人隔天只會有30%的機率會發燒
- 如果病人是健康的會有50%的機率覺得正常
- 如果病人發燒了會有60%的機率覺得頭暈

病人連續看醫生3天，得以下結果：[正常、冷、頭暈]
根據viterbi演算法可以計算出3天的狀態分別是：[健康、健康、發燒]





透過CART分類和迴歸樹選擇適合的HMM模型

CART：為決策樹的一種，在條件下輸出的條件概率分佈的學習方法，可用於分類 () 或回歸。
CART是二元樹，每個節點取值方式為判斷 “是” 與 “否”，左邊為 ”是” 的分支，右邊則為 ”否”。

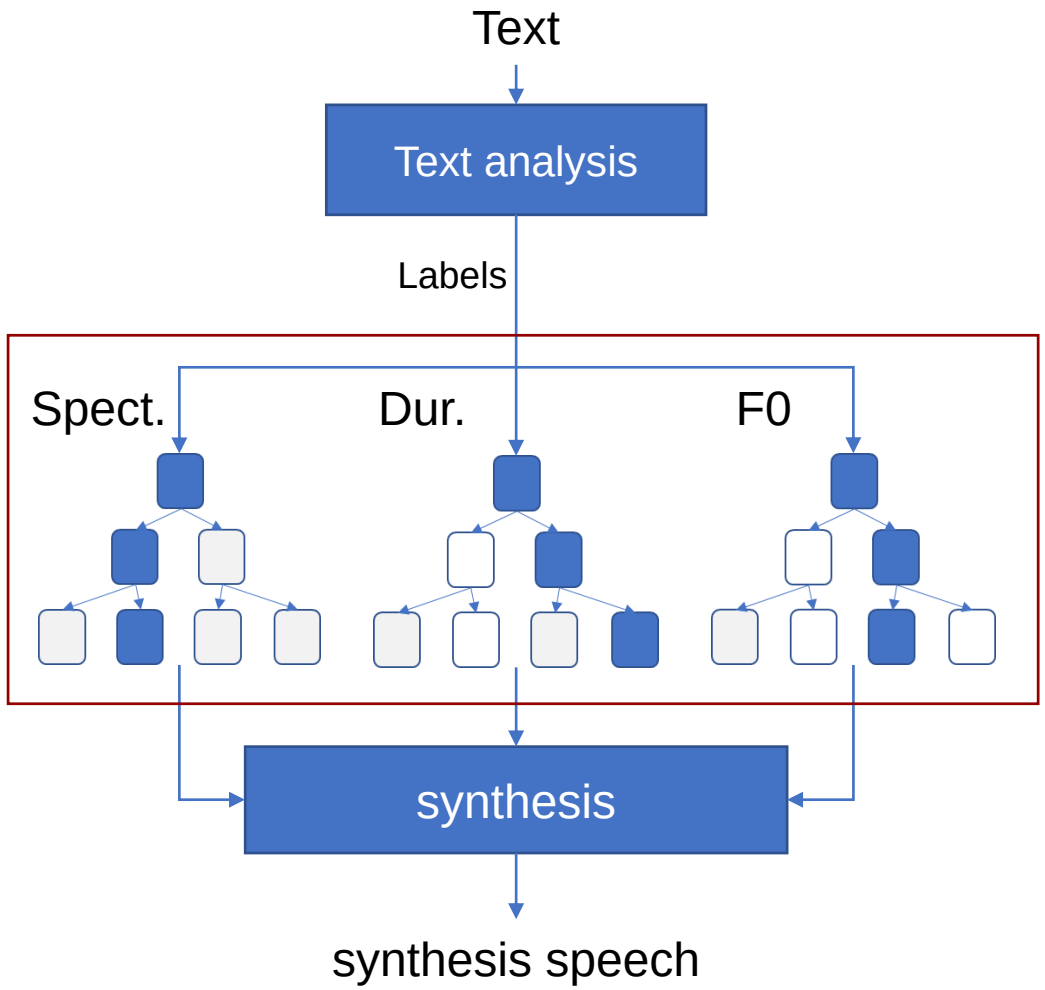
分類樹：根據基尼係數 (樣本集合中選隨機一個樣本被分類錯誤的機率，值越小代表分錯的機率越低) 作為依據將一集合不斷分為兩類。

- 分類依據可視為某個特徵屬性

回歸樹：使用最小的平方誤差作為判斷依據，將一集合不斷分為兩類。

- 分類依據可視為某個東西或是值

輸入：訓練資料，停止計算的條件
輸出：CART決策樹



statistical parametric synthesis

附錄

上下文標籤格式

完整上下文標籤：

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

所有特殊位元皆為固定格式的連接用符號，無意義

可以看到除了 p1~p7 總共分為 9 種關係，各類所代表的物理意義可參照下一頁每一種顏色對應其關係

格式：

1. p1^p2-p3+p4=p5@p6_p7
2. /A:a1_a2
3. /B:b1-b2@b3-b4&b5-b6#b7-b8!b9-b10|b11
4. /C:c1+c2
5. /D:d1_d2
6. /E:e1+e2@e3+e4
7. /F:f1_f2 /G:g1_g2
8. /H:h1=h2@h3=h4
9. /I:i1_i2
10. /J: j1+ j2- j3

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

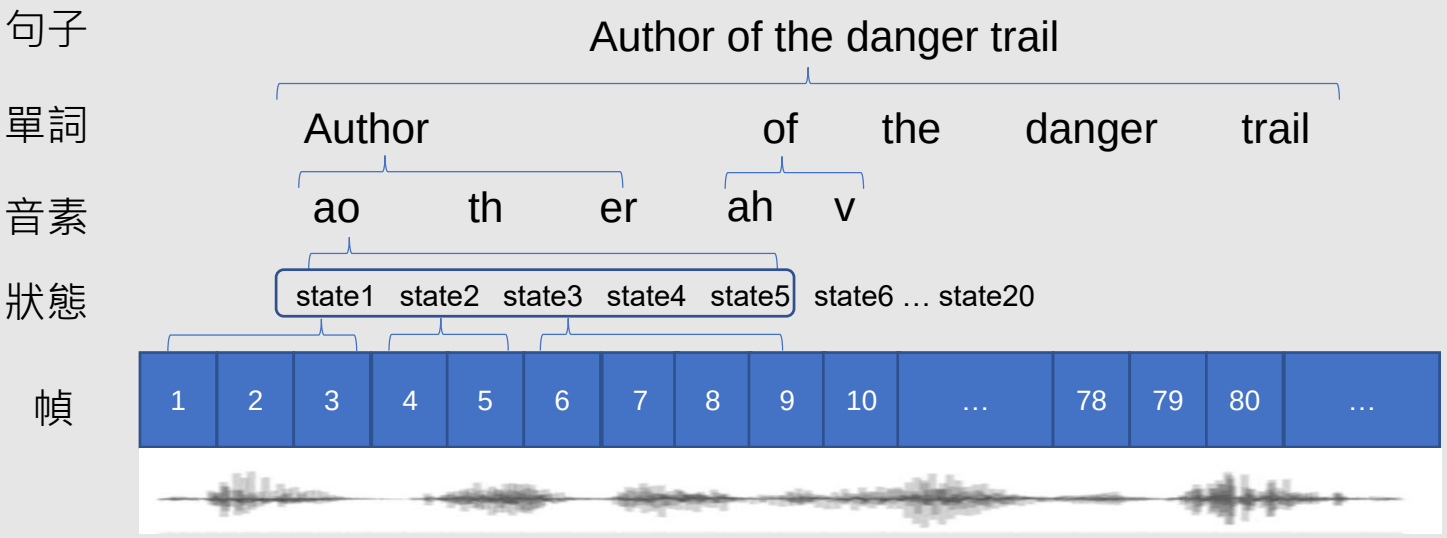
持續時間 sil^sil-ao+th=er 其餘資訊

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、單詞語、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