多語聲學單位分類之最佳化研究

Authors: 呂道誠、呂仁園、江永進、許鈞南

Professor: 陳嘉平

Reporter:　吳柏鋒

摘要

- 簡介
- 多語聲學單位分類相關研究
- 聲學單位的分類與數目的最佳化
- 實驗與結果

簡介

- 介紹專家知識與資料分析，兩種多語聲學單位分類方法。
- 提出了一套整合專家背景知識與實際語音分析的方法，來產生一組新的聲學單位
- 針對這組聲學單位的數目，使用差分貝式資訊法則來做最佳的處理。

簡介

- 從訓練好的隱藏式馬可夫聲學模型中，計算其單位間的相似度矩陣
- 之後透過語音學和音韻學的知識，限定了各個聲學單位能群化的上限，根據不同限定的群化上限，使用聚合階層式分群法，來建立不同的結構樹。
- 利用差分貝式資訊法則，將每個結構樹中發音相近的聲學單位做合併，當差分貝式資訊法則的值小於零的時候，就停止合併，而新合併成一群的聲學單位則為新的聲學單位。

多語聲學單位分類相關研究

- 多語聲學單位分類的方法，大致上可分為兩種：

(一)以專家知識的方法

(二)從資料分析的角度(data-driven)

合併多語言之相似音素。

多語聲學單位分類相關研究

(一)以專家知識的方法

1. **語言相關:**

   聲學單位是結合各自語言的音素而成的，依據此方法，聲學模型的訓練上，各個語言間具相同發聲的音素彼此之間並不共用訓練語料

2. **語言獨立:**

   利用一套能包含所有語言的音標符號，如IPA、SAMPA和 Worldbet等，將不同語言，但相同發音的音素標記成相同的符號。

多語聲學單位分類相關研究

(二)利用資料驅動的方法

- 此方法是以真實語音資料的發音特性為考量，根據現有的語料，運用**群聚技術**定義出一組多語聲學單位。即被凝聚在同一群的發音會有某些特性是相近的且會被標記成一致的發音符號

多語聲學單位分類相關研究

1. **分裂階層式分類法**

把整個資料集合看成一個群聚，然後逐次分裂，每次都會在其中一個群聚裡切割相似度最低的連結，成為二個較小的群聚，直到群聚數目達到事先所設定的數目為止

2. **聚合階層式分類法(AHC)**

將每一筆資料視為一個群聚，然後每次將特性最相近的二個群聚合而為一，直到群聚數目達到事先所設定的數目為止

多語聲學單位分類相關研究

例:先算出所有將要群聚的聲學單位的相似程度矩陣，而此矩陣的數值是利用訓練好的聲學模型參數來算出彼此間的距離，下列為兩種計算距離的方法:

1. Bhattacharyya distance

$$D_{bata} = \frac{1}{8}(u_p - u_q)^T \left[\frac{\Sigma_p + \Sigma_q}{2}\right]^{-1}(u_p - u_q) + \frac{1}{2}\ln\frac{\left[\frac{\Sigma_p + \Sigma_q}{2}\right]}{\sqrt{|\Sigma_p||\Sigma_q|}}$$

2. Kullback-Leibler divergence

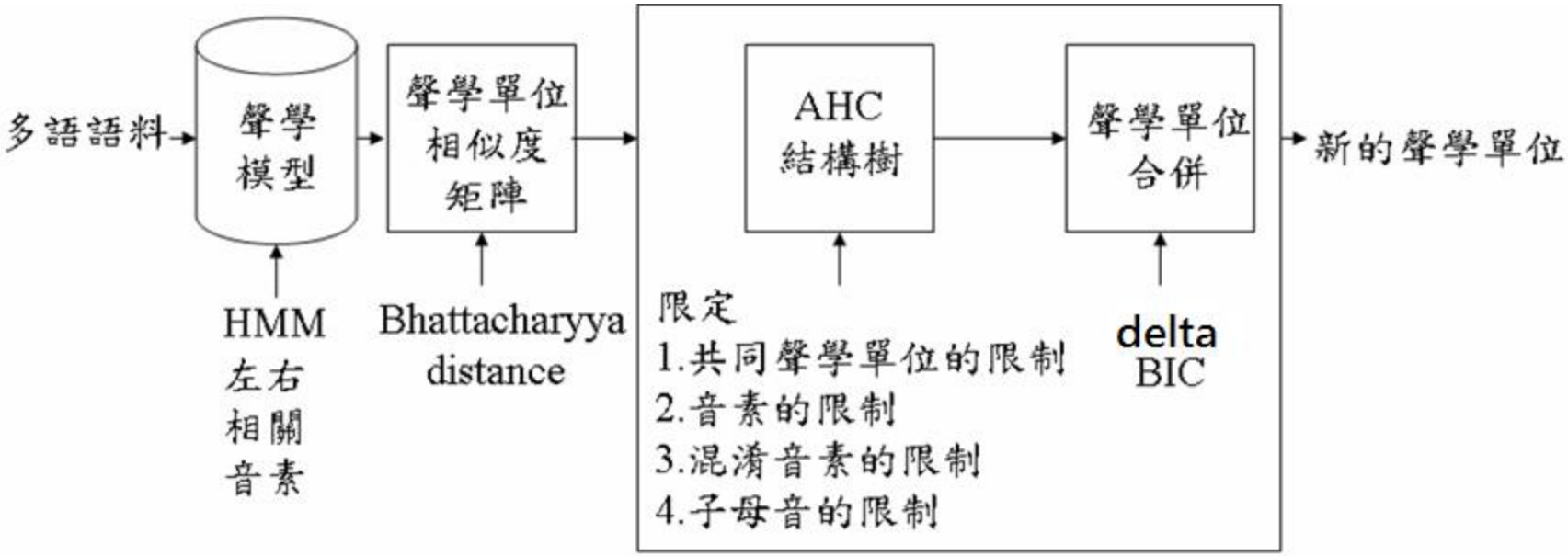
$$D_{KL} = \frac{1}{2}\left(\ln\frac{\Sigma_p}{\Sigma_q} + trce\left(\Sigma_p^{-1}\Sigma_q\right) + (u_p - u_q)^T \Sigma_p^{-1}(u_p - u_q) - d\right)$$

聲學單位的分類

- 依據聲韻學和語音學的語音知識，限定了群聚技術裡的分類，讓發聲相近的聲學單位，透過相似度的篩選機制，建立了以AHC為方法的結構樹，在相同的結構樹中，聲學單位有機會做合併

- 在此利用delta-BIC模型選擇的方法，將其應用在找出最佳的聲學單位

聲學單位的分類



圖二、結合專家知識和資料驅動方法產生最佳聲學單位數目的流程圖

聲學單位的分類

- 在AHC之前，我們先將所有的聲學單位分成子群，而這些子群的成立，是根據以下的四種分群限定，產生不同的結構樹:

1. 共同聲學單位的限制
2. 音素的限制
3. 混淆音素的限制
4. 子母音的限制

聲學單位的分類

1. 共同聲學單位的限制

在這個限定下，AHC結構樹的數目，只針對各語言間的具有相同的IPA標音的聲學單位，而每個子群裡的聲學單位為左右相關音素

2. 音素的限制

這個限定，是在觀察左右相關音素與左右獨立音素之間的關係。因此，每棵樹的最底層是左右相關音素，而最上層為左右獨立音素

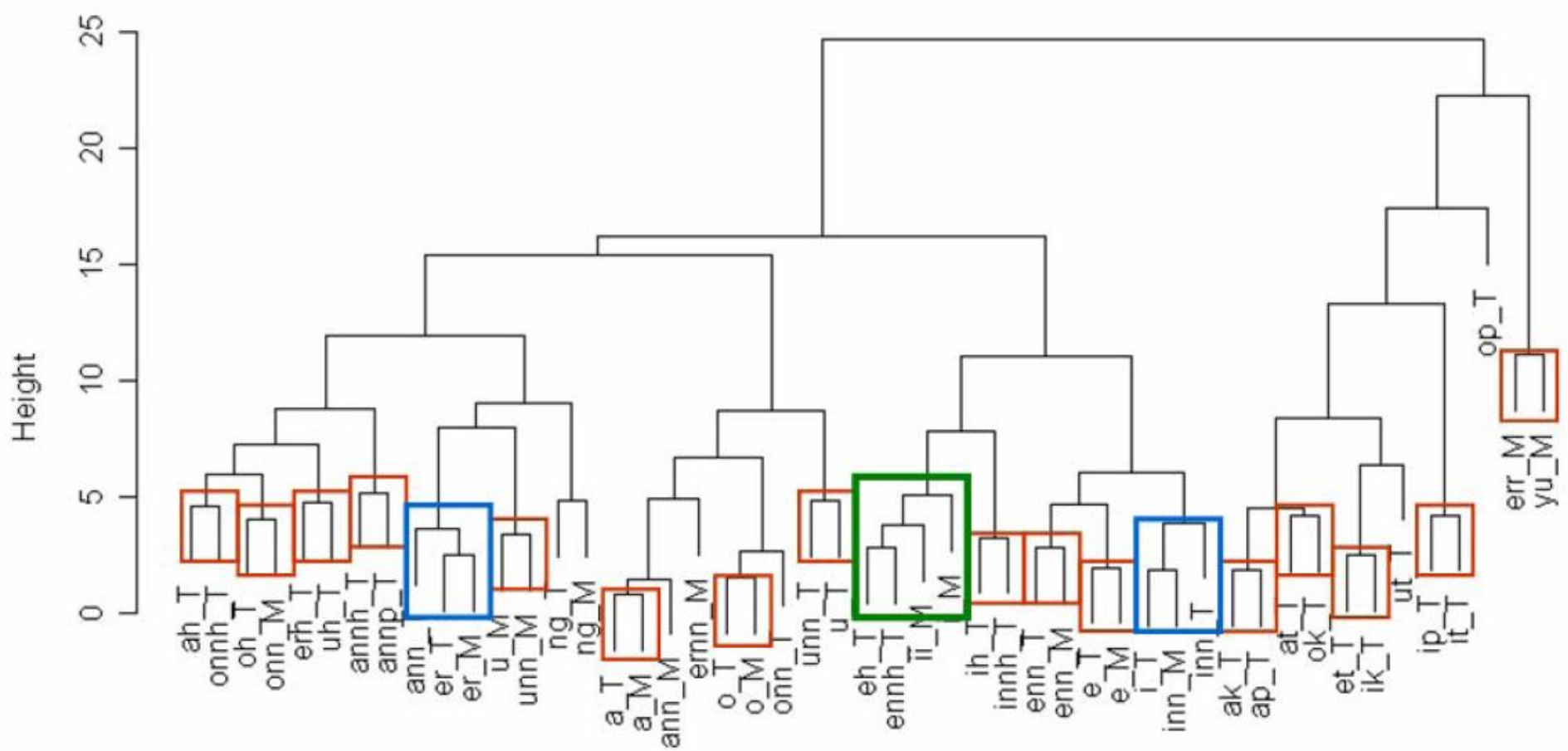
聲學單位的分類

3. 混淆音素的限制

我們了解到，有些音素是很容易混淆的，比如華語的捲舌音和不捲舌音，或台語的入聲音，如-p和-k結尾的音素，因此，這個限制裡，我們擴大可以合併的範圍到混淆音素。

4. 子母音的限制

在這個階段，我們只將分類樹的樹頭分為兩類，子音類和母音類。而樹根則為華台語所有的左右相關聲學單位。因此在最極端的情況下，華台語的聲學單位，最後則會分成只有字音和母音兩個



圖一、標準 AHC 所產生的華語與台語母音與子音的樹狀圖

聲學數目的最佳化

- 利用最大概似度(Maximum Likelihood)的方式，從p個模型中找出最能代表n筆資料 $X = x_1, \ldots, x_n, x_i \in R^d$ 的最佳模型。

  BIC公式如下：

$$BIC_p = \log L_p(X) - \frac{1}{2}\lambda d_p \log n$$

其中 L_p 是模型 p 的最大概似度，λ 是一個微調值，而 d_p 是 p 模型裡的參數數目。

聲學數目的最佳化

- 依據差分貝式資訊法則（delta BIC):

（判斷delta-BIC值大於0，合併

值小於0，停止合併）

$$\Delta BIC_{pq} = BIC_p - BIC_q$$

$$= -\frac{n_p}{2}\log|\Sigma_p| - \frac{n_q}{2}\log|\Sigma_q| + \frac{n_r}{2}\log|\Sigma_r| + \frac{1}{2}\lambda(d + \frac{d(d+1)}{2})\log n_r$$

其中 n_p, n_q 分別爲模型 p 和 q 所對應的訓練語料數目

$n_r = n_p + n_q$ ，而 Σ_p, Σ_q 分別爲模型 p 和 q 的共變異數矩陣的行列式

d 爲參數數目

實驗與結果

- 使用ForsDAT語料庫(麥克風)

	語言	人數	語音句數	總時間(小時)
訓練語料	華語	100	43078	11.3
	台語	100	46086	11.2
測試語料	華語	10	1000	0.28
	台語	10	1000	0.28

實驗與結果

- 三種多語聲學單位實驗方法(建構在四類限定上)
  (1)語言相關(Lang-De)
  (2)語言獨立(Lang-In)
  (3)聲學單位數目最佳化

實驗與結果

- 每個聲學模型使用HMM來做訓練，而模型的單位為左右相關音素，每個HMM有3個狀態，每個狀態下的高斯分佈模型(簡稱GMM)數目

(1)動態GMM

每個狀態下的GMM數目的增加是根據訓練語料的多寡來決定

(2)固定GMM

每個狀態下的GMM數目是以固定的倍數增加

實驗與結果

- A. 以專家知識為本的固定GMM與動態GMM辨識結果

		8-mix	16-mix	32-mix	64-mix
動態 GMM	Lang-De (1503)	60.7	63.9	62.1	60.2
	Lang-In (1242)	62.5	**64.7**	64.3	63.0
固定 GMM	Lang-De (1503)	59.3	62.8	60.2	58.6
	Lang-In (1242)	61.4	63.1	62.5	61.6

表三、 Lang-De 與 Lang-In 的聲學模型用來做固定 GMM 與動態 GMM 的語音辨識正確率。

實驗與結果

- B. 以HMM為單位的最佳化聲學模型的結果

	8-mix	16-mix	32-mix	64-mix
C-I (1242)	62.5	64.7	64.3	63.0
C-II (527)	51.7	56.7	60.4	59.4
C-III (1083)	59.5	64.2	**65.7**	**66.1**
C-IV (862)	56.4	59.6	61.8	61.5

表五、限制下最佳化聲學模型的辨識結果

實驗與結果

- C. 以狀態為單位的最佳化聲學模型的結果（將合併單位從HMM轉狀態）

	8-mix	16-mix	32-mix	64-mix
C-I (3726)	62.5	64.7	64.3	63.0
C-II (1581)	51.7	56.7	60.4	59.4
C-III (3569)	61.9	65.1	66.4	**66.7**
C-IV (2760)	59.2	61.5	62.3	62.5
DT(3374)	62.2	63.4	64.7	64.9

表七、以狀態爲單位的最佳化聲學模型的辨識結果，其中最後一行是決策樹的結果。