多語聲學單位分類之最佳化研究

呂道誠　Dau-cheng Lyu
長庚大學電機工程學系
d9221003@stmail.cgu.edu.tw

呂仁園　Ren-yuan Lyu
長庚大學資訊工程學系
renyuan.lyu@gmail.com

江永進　Yuang-chin Chiang
國立清華大學統計學研究所

許鈞南　Chun-nan Hsu
中央研究院資訊科學所
chunnan@iis.sinica.edu.tw

摘要

由於全球化的形成，人與人之間的溝通不再限於同一種語言，因此多語的語音辨識也變的格外的重要。如何有效整合多語的聲學模型是一個關鍵議題，因爲一組好的多語聲學單位將影響辨識結果。本論文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，並且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。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。之後，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爲新的聲學單。我們將用 ForSDAT01 華台雙語語料庫來實驗評量，而實驗結果顯示，本論文所提出的新方法比只用專家知識所定義的聲學單位所訓練出的辨識器有較高的辨識效果。

關鍵詞：多語語音辨識、音素群化、差分貝式資訊法則

一、緒論

語音是人與人溝通最直接也是最原始的一種工具之一，透過語音的傳遞，能夠拉近彼此的距離。近來，由於地球村的形成，人與人之間的交談並不再限制單一的語言。這種現象，在國家彼此相鄰的歐洲和人種混合的亞洲都會時常發生。如台灣，目前華語和台語是最主要的兩種語言。因此，自動語音辨識的研究領域裡，從原本的單一語言的語音辨識，漸漸的，已朝向雙語或多語的方向了[1]。同樣的，如果一套自動語音辨識機器，能夠一次辨識兩種語言或兩種以上的語言，這會比只能處理單一語言的機器來的更強大。所以說，多語言的語音辨識，是個當前重要且必須探討的議題。

另外，在多語的語音辨識領域裡，對於不同語言間聲學單位的定義，是個值得研究的方向。這個議題，主要在探討如何將多個語言的發音符號做有效的定義，好讓最後，給定一句未知的語音，在根據這些定義好的發音符號所訓練出來的聲學模型，能夠有最佳的辨識效果。換句話說，如何定義一組發音符號，能夠有效的整合不同語言間具有相

同發音或相似的發音。而整合的觀念，也可以說是群類化(Clustering)的意思，這代表著說，藉由群類化的技術，將不同語言間發音接近的單位群化，讓這群化後的發音有共同的標音方式，而最後得到多語語言間整合的效果。根據之前學者的實驗結果[2,3]，一組好的多語聲學單位，將有助於最後語音辨識的結果，因此由此可知，聲學單位的選取與定義在多語的語音辨識裡，有著舉足輕重的地位。

依據之前的報告，有學者是使用國際音標系統(International Phonetic Alphabet，簡稱 IPA)[4]來統一標記不同語言間的發音。這套系統是透過語音學或音韻學專家的知識，將不同語言的發音都一對一的對應到 IPA 裡的標音符號，透過這套機制，可將而不同語言間專家所認定相同發音的語音都標記成相同的符號[16]。利用這種方法的聲學模型有個好處，就是相同標音符號但不同語言的訓練語料，能夠彼此分享，使得聲學模型更加強健。但，大部分的語料庫，其劇本都是事先規劃好的，之後再請語者進行錄音，由於資料的龐大，錄完音之後，並沒有做標音符號與真實發音之間的再確認，最後使得，真正所錄下來的發音，和原來的劇本裡所標的發音符號並不完全的一致。如以華語來說，”師”這個發音，往往語者會發成沒捲舌”斯”的情形。因此，有其他學者，用另一套方法，先分析既有的語音資料，而後根據發音相似度的量測，將相似的發音歸類成一群，最後找出最能符合這批語料的聲學單位[2]。近年來，也有學者利用聲學與文脈的分析，產生了多語語音的聲學單位[3]。可是，這些學者所提的方法，並沒有將這些所找到的多語語音聲學單位數目上做最佳化的分析，而是利用不同的門檻值，產生不同的聲學單位數目來做實驗。

本篇論文，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，並且對這組聲學單位的數目做最佳的處理。聲學單位是以左右相關音素爲主，實驗在華語及台語的語料庫上。此方法，首先對每種語言，訓練以隱藏式碼可夫模型(Hidden Markov model，簡稱 HMM)爲主的聲學模型，之後透過相似度的量測，產生所有聲學單位的相似度矩陣。然後，加入了語音學以及音韻學的知識，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(Agglomerative Hierarchical Clustering，簡稱 AHC)，建立了不同的結構樹，此外，在這裡引進了差分貝式資訊法則 (delta Bayesian Information Criterion，簡稱 delta-BIC)[5]，將同一棵結構樹裡，由下而上，把兩群類的 delta-BIC 值大於 0 的聲學單位做融合，直到 delta-BIC 值小於 0 則停止融合，根據貝式理論，以現有的資料來說，這樣的方式可以找到最佳的聲學單位數目，因此我們才用這種方法來做聲學單位的最佳化。最後，將這些定義好的聲學單位做訓練 HMM 的模型訓練，在每個融合後的聲學單位裡，彼此分享訓練語料，這樣可達到多語聲學的整合效果。

本篇文章的架構如下，第二章：介紹多語聲學單位分類的相關研究，裡面包含了兩種(專家知識和資料驅動)方法。第三章是本論文所提出的新方法，將前章的兩種技術做結合，並且探討了如何把哪些聲學單位該分爲一群，以及最後該如何決定聲學單位的數目。實驗所用到的語料、環境設定和結果分析都在第四章做完整的描述。而最後一章是結論。

二、多語聲學單位分類相關研究

多語聲學單位分類的方法，大致上可分爲兩種：(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。現分別介紹如下：

（一）利用專家知識的方法

聲學單位利用專家知識的方法，可分爲 1.語言相關(Language Dependent)與 2.語言獨立(Language Independent)的方式。其中，語言相關的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。如華語和台語，這兩種語言都有/a/的發音，透過專家的認定，將台語語音 /a/ 和華語語音 /a/ 的發音，標記成[a_T]與[a_M]，因此但每個音素符號會分別帶上各個語言的標記。但此作法的缺點是具相同發音的音素在不同的語言裡，彼此的語料並不能共用，而可能會使得某些訓練語料相對的不足，而造成聲學模型在做參數估計的時候會不夠強健。

相反的，語言獨立的方法，則是利用一套能包含所有語言的音標符號，如 IPA、SAMPA [6] 和 Worldbet [7]等，將不同語言但相同發音的音素標記成相同的符號，因此，如台語的 /a/ 和華語的 /a/ 的發音通通都標記成 [a] 的發音符號。此種作法可以有效地將相同發音部分的音素做合併，以減少語音音素的數目，相對的，在同樣多的多語訓練語料上，此種方法其每個聲學單位所能分配到的訓練語料會比其語言相關的方法來的多，因此，所訓練出來的聲學模型參數也會更加強健。<表一>裡記載了華台語用福爾摩沙標音系統(ForPA)[8]所定義出來的音素。ForPA 是根據專家知識所定義的一套可標記台灣主要三種語言(華語、台語和客語)的標音系統，同時，這套標音系統裡的每一個標音都可以和 IPA 做一對一的對應。

然而，使用語言獨立的方法的缺點爲：此方法沒有辦法反映出實際語料的發音特性。原因是，其音素的定義是完全建立在專家的知識上，而非從資料特性上做考量。理論上，假設所有的語料都發音的和標記的完全相同，這樣以專家知識上所定義的聲學單位分類是完美的。但，在真實情況下卻不是這樣的。往往，事先標記好的劇本，錄音員並不能百分之一百的完全照著劇本的發音音素正確的念出來，而造成發音和音素標記上會產生不匹配的現象。因此，我們也要從真實語料上做統計和分析，這樣才能確實地反應真實語料上發音音素的特性。

	子音	母音
OBT	[bh] [gh] [r]	[ah] [ak] [annh] [annp] [ap] [at] [eh] [ennh] [erh] [et] [ih] [ik] [innh] [ip] [it] [oh] [ok] [onnh] [op] [uh] [ut]
TM	[b] [c] [d] [g] [h] [k] [l] [m] [n] [p] [s] [t] [z]	[ann] [a] [enn] [e] [er] [i] [inn] [ng] [o] [onn] [unn] [u]
OBM	[ch] [f] [rh] [sh] [zh]	[ernn] [err] [ii] [yu]

表一、以 ForPA 爲標記的華台語裡的音素表。在這個表中，可分爲 3 個部分，其中 OBT 是台語特有的音素，OBM 是華語特有的音素，而 TM 則是華台語都有的音素

（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爲考量，根據現有的語料，定義出一組多語聲學單位。主要的觀念是運用群聚技術，將資料依據彼此的相關程度，分成不同的群組；而被凝聚在同一群的發音會有某些特性是相近的，也就是說透過真實語料的分析，有相同特性的發音會被標記成一致的發音符號。在此所用到的群聚技術是以階層式的爲主，

這個技術可分爲兩類，1. 分裂法(Divisive algorithm) [9] 2. 聚合法(Agglomerative Algorithm) [10] 二種。分裂法是先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，成爲二個較小的群聚，直到群聚數目達到事先所設定的數目爲止。而聚合法是先將每一筆資料視爲一個群聚，然後每次將特性最相近的二個群聚合而爲一，直到群聚數目達到事先所設定的數目爲止。以後者爲例，在做聚合階層式的群聚(AHC)技術的演算法之前，會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，而距離有兩種普遍的計算方法，分別爲 a. Bhattacharyya distance [2] 和 Kullback-Leibler divergence [11],其公式如下：

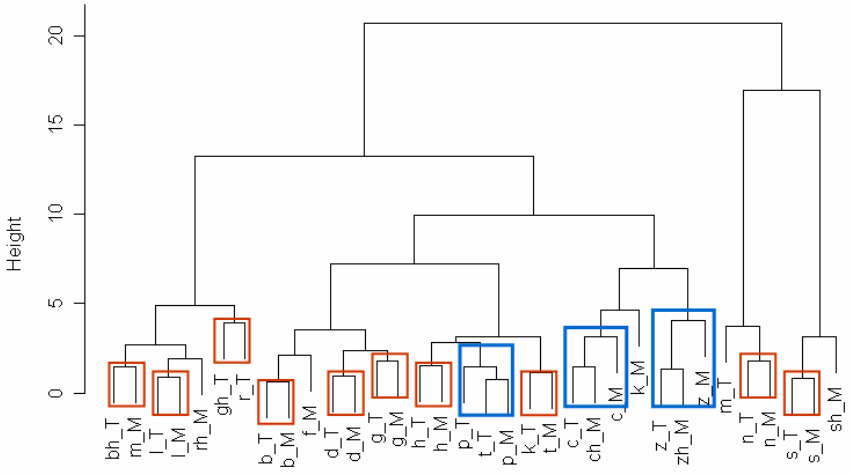
$$D_{bata} = \frac{1}{8}(u_p - u_q)^T \left[\frac{\Sigma_p + \Sigma_q}{2}\right]^{-1} (u_p - u_q) + \frac{1}{2}\ln \frac{\left[\frac{\Sigma_p + \Sigma_q}{2}\right]}{\sqrt{|\Sigma_p||\Sigma_q|}} \qquad \text{(式一)}$$

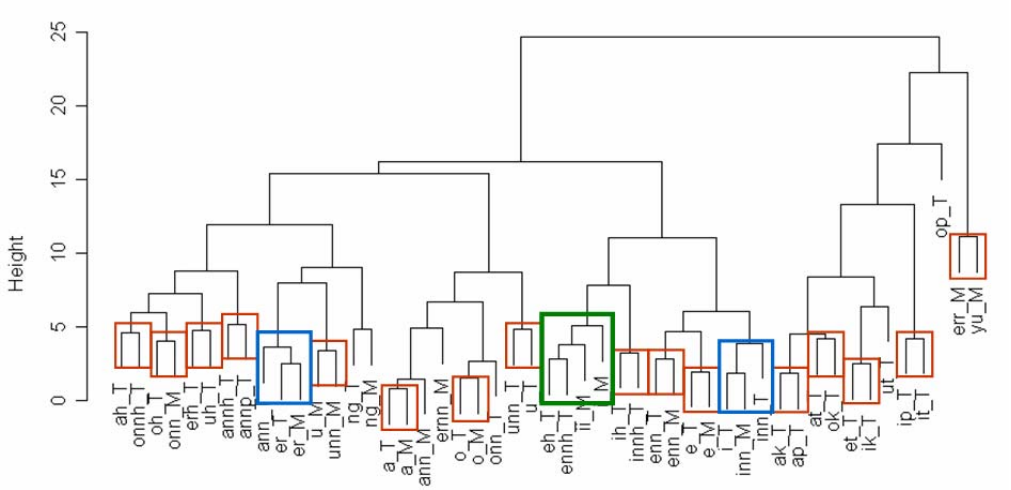
$$D_{KL} = \frac{1}{2}\left(\ln \frac{\Sigma_p}{\Sigma_q} + trce\left(\Sigma_p^{-1}\Sigma_q\right) + (u_p - u_q)^T \Sigma_p^{-1}(u_p - u_q) - d\right) \qquad \text{(式二)}$$

而其 AHC 的評估方式有四種[12]，分別爲 A.重心連結聚合演算法、B.平均連結聚合演算法、C.完整連結聚合演算法以及 D. 單一連結聚合演算法。

不論是由上而下(top-down)的分裂法或由下而上(button-up)的聚合法，其最後的聲學單位數目，都是事先決定好的，之後產生固定的聲學單位來訓練聲學模型。之前的學者使用這兩種階層式方法，並沒有對這些方法做出聲學單位數目的最佳化。除此之外，這些聚合方法，捨棄了專家知識，而直接只採用現有的語音資訊去作分析去定義最後的聲學單位，如果此批語料摻雜了許多的雜訊或某些聲學單位的訓練語料不足的時候，往往可能會產生比用專家知識方法還差的辨識結果。

<圖一>中爲用標準 AHC 所產生的華語與台語其母音與子音的樹狀圖，其中距離的計算是採用歐基里德距離，而 AHC 的評估方式爲完整連結聚合演算法。而在 AHC 之前的相似度矩陣值由每個聲學單位其 HMM 裡的三個狀態下的 Bhattacharyya 距離的平均值。



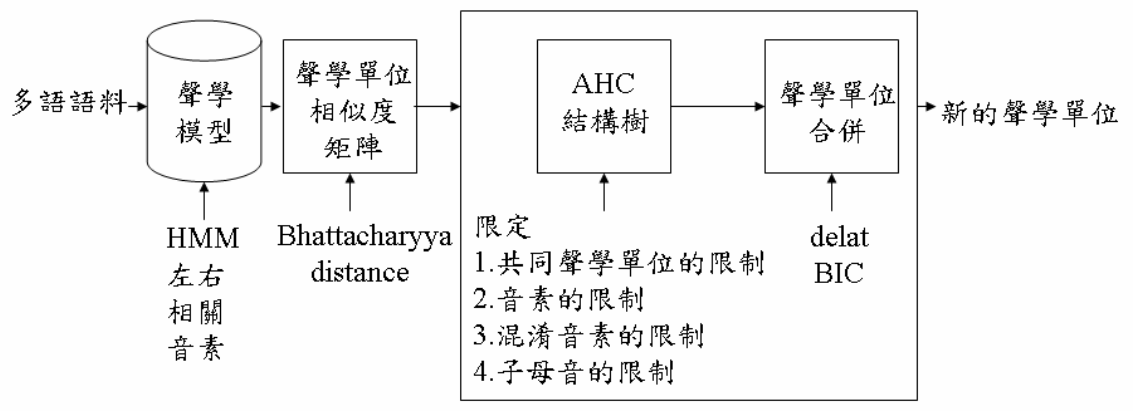


圖一、標準 AHC 所產生的華語與台語母音與子音的樹狀圖

三、聲學單位的分類與數目的最佳化

本篇論文提出了一套，結合了專家知識和資料驅動的方法，從既有的多語語料當中，尋找一組新的聲學單位且讓其數目最佳化。我們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 爲方法的結構樹，在相同的結構樹中，聲學單位有機會做合併。然而，不同的結構樹中的聲學單位彼此之間不能做合併。這樣的機制，避免了發聲差很多的聲學單位，因爲語料品質的影響（如雜訊），而相互融合。

另外，如何在相同的結構樹中，最後找出最佳的聲學單位數目，也就是說，如何決定哪些聲學單位該合併，哪些群不該合併。在此，我們引進了 delta-BIC 技術。根據貝式定理，從現有的訓練語料中，利用 delta-BIC 模型選擇的方法，可找出最佳的模型。相同的，我們將這個觀念，應用在如何找出最佳的聲學單位。而整個尋找聲學單位數目的最佳化過程顯示在<圖二>。



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

因此，由<圖二>可看出，要尋找出一組新的聲學單位，有兩元件程是必須探討的：AHC 結構樹的限制與聲學單位合併的機制，以下，我們就分別細說這兩項元件。

（一）限定結構樹範圍 - 依據專家知識

根據純資料驅動的做法，從聲學單位相似度矩陣到產生 AHC 結構樹時，並沒有做什麼機制來防止因語料品質的缺陷，而造成 AHC 在建立結構樹的時候，將一些發聲差別很大的聲學單位反而安排的很近，因爲標準的 AHC 是完全按照相似度矩陣來做聲學單位的分類，而相似度矩陣是由語料庫所建立的聲學模型所產生的，因此 AHC 結構樹長的好不好，有很重的部份是依賴語料庫的品質優劣。因此，我們在這一節將做一些限制，來改善因語料的錄音品質缺陷所造成 AHC 長的不好的問題。

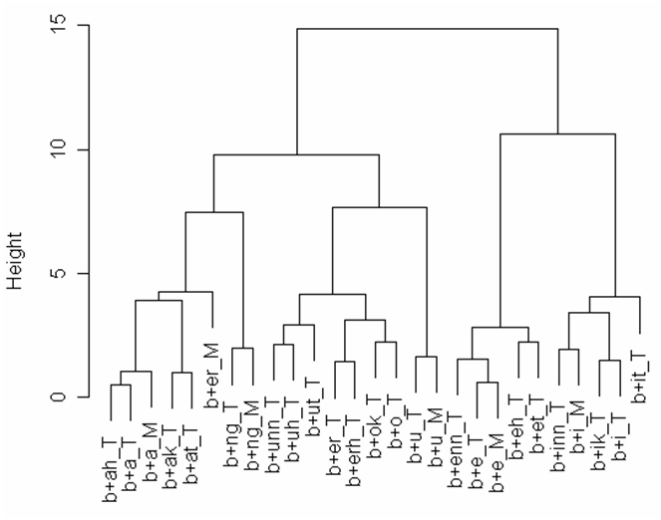
由 IPA 或 ForPA 等標音系統所定義出的各語言聲學單位，在語言學或聲韻學上是有道理的，因爲這些聲學單位，都是由一些專家所定義出來的。根據這些定義，我們可以將這些單位，粗分爲子音、母音、或由不同的發聲構造來分成，鼻音和喉頭音等。因此，在 AHC 之前，我們先將所有的聲學單位分成子群，而這些子群的成立，是根據以下的四種分群限定。我們依此道理來將我們從由 Bhattacharyya 所算出的每個聲學單位裡的狀態混淆矩陣，到由 AHC 所產生的分類樹之間，做四種階層的限制。而不同的限制，也將會產生不同結構的分類樹。

I. 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爲左右相關音素。以華台語舉例來說，[b-a+ng_T, b-a+ng_M]，這樣的結構樹共有 284 個。因此，這些結構樹，只有兩層，一層就是華台語的左右相關音素，第二層就是他們兩個的合併 [b-a+ng]。這個限制，主要是在觀察，華台語之間共同左右相關音素是否應該合併，如果完全合併，則是語言獨立的方法，相反的，如果全部不合併的話，就退回語言相關的方法了。

II. 音素的限制：

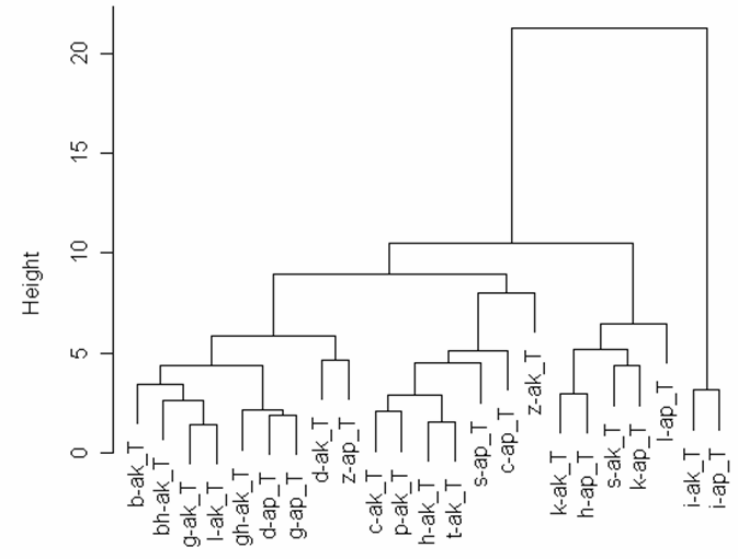
這個限定，是在觀察左右相關音素與左右獨立音素之間的關係。因此，每棵樹的最底層是左右相關音素，而最上層爲左右獨立音素，如<圖三>所示。因此，如果到最後在這棵樹裡的每個左右相關音素都能夠合併，則原本以[b+*]爲主的左右相關音素，就會退化成語言獨立的左右獨立音素的[b]



圖三、以[b]爲例的 AHC 結構樹

III. 混淆音素的限制：

根據[13]，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，如帶-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。如<圖一>的字音和母音 AHC 的結構樹圖中，我們用 delta-BIC 的技術(下一節會談到)，將結構樹中，delta-BIC 值大於零的聲學單位群組化，而被群組化的聲學單位就被認定是一組混淆音素組，之後，以此組音素所衍生的左右相關音素就可用 AHC 的方式產生一棵結構樹。如<圖一>裡我們可以看出，子音共有 13 組混淆音素（用框框圍住的），而母音則有 19 組。這些混淆音素，最小的是由兩個音素所組成的，而最大的是四個音素。如-ap 與-ak 就是一組混淆音素，而以-ap 和-ak 爲主的左右相關音素所產生的 AHC 結構樹顯示在<圖四>中。



圖四、母音中帶-ap 和-ak 結尾音素的 AHC 結構樹

IV.子母音的限制：

在這個階段，我們只將分類樹的樹頭分爲兩類，子音類和母音類。而樹根則爲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。

（二）決定最佳數目 - 依據差分貝式資訊法則 (delta Bayesian Information Criterion)

貝式資訊法則(簡稱 BIC)，由 Schwarz 在 1978 所提出[2]，是一種非對稱模組選取的準則。其利用最大概似度(Maximum Likelihood)的方式從 p 個模型中找出最能代表 n 比資料 $X = x_1,...,x_n, x_i \in R^d$ 的最佳模型。假設，每筆資料都是相互獨立的。而第 p 個模型的 BIC 公式如下

$$BIC_p = \log L_p(X) - \frac{1}{2}\lambda d_p \log n \qquad \text{(式三)}$$

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。

delta-BIC 則是將兩群不同的貝式資訊法則值做相減，簡化之後，可以轉化成兩個不同群的最大概似度值做相加，之後減去兩群組合併後的最大概似度值，再加上模型參數目的函數值。這項技術常常被拿來當作語者交換點或語言交換點偵測的一個判斷[5]，通常一般法則是當 delta-BIC 大於 0 的時候，就判定兩群組的邊界資料爲一個交換點。而本篇論文將此方法拿來作爲多語聲學模型單位最佳數目的根據。根據[5]，針對第 p 個、第 q 個模型和兩個合併之後的 r 模型的 delta-BIC 公式如下所示：

$$\Delta BIC_{pq} = BIC_p - BIC_q$$
$$= -\frac{n_p}{2}\log|\Sigma_p| - \frac{n_q}{2} lpg|\Sigma_q| + \frac{n_r}{2}\log|\Sigma_r| + \frac{1}{2}\lambda(d + \frac{d(d+1)}{2})\log n_r \qquad \text{(式四)}$$

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。

因此，由(式四)我們可看出，如果模型 p 和 q 以及合併之後的 r 模型，其最大概似度還大於合併之前個別模型 p 和 q 的最大概似度總和，則我們相信，這兩個模型是可以合併的。所以，針對上述的聲學單位分類範圍，透過 delta-BIC 的方法，最後可找出最佳的聲學單位數目。而如果其中的 λ 值爲零的話，(式四)就退化成一般的最大概似度估計了。但一般在做模組選取時是採用 delta-BIC，因爲直接用大於或小於零就可判別，但如用一般最大概似度估計，則要對每個不同的狀況設下不同門檻值，才能作決定，因此本論文也採用 delta-BIC 來做多語聲學模型最佳化的判準。

根據之前依不同限定所產生的結構樹，我們從結構樹最下面兩兩聲學模型或群組間的 delta-BIC 值大於零的部分做合併，而合併後的聲學單位，利用[14]的技術，將原本未合併之前模型 p 和 q 所對應的訓練語料分享給新的合併之後的 r 模型，而達到聲學單位整合的目的。這樣的合併會採由下而上的方式一直進行，直到當 delta-BIC 值小

於零的時候才會停止。停止後所合併的群組則爲新的一組聲學單位，用來訓練聲學模型，做語音辨識。

四、實驗與結果

（一）實驗語料

本實驗，是爲了找出最佳的多語聲學單位，並且驗證此組聲學單位能達到最佳的語音辨識結果。因此，我們用了華台雙語語音資料庫 ForSDAT 中的 01 年麥克風語料[8]。這個語料庫裡包含了 100 人的訓練語料和另外 20 人的測試語料。在訓練語料中，每個語者都錄製了兩種語言，且男女比率平均。其相關的統計數字列在<表二>

	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

表二、ForSDAT01 年華台雙語語音資料庫的相關統計數字

（二）實驗設定

由之前的介紹，我們可將多語聲學單位的實驗依三種方法來分類，分別爲語言相關(Lang-De)、獨立語言(Lang-In)和本篇所提出的聲學單位數目最佳化，而在本論文所提出的方法中，又依不同的限定，可分爲 4 類，分別是共同聲學單位的限制(C-I)、混淆音素的限制(C-II)、音素的限制(C-III)和子母音的限制(C-IV)。

在特徵擷取上，我們採用了以梅爾倒頻係數(簡稱 MFCC)爲主的方法，而 20 毫取一個音匡，每 10 毫秒移動一個音框，每個音匡有 39 維度。每個聲學模型使用 HMM 來做訓練，而模型的單位爲左右相關音素，而每個 HMM 有 3 個狀態，每個狀態下的高斯分佈模型(簡稱 GMM)數目，則是依照每個狀態下所能對應到的訓練語料量來決定，根據[15]原則，所對應到的訓練語料量越多的狀態，其 GMM 數目也會越多，在此系列實驗中，我們設定每個 GMM 必須要有 30 個以上的音匡。所以如果每個狀態下的 GMM 數目的增加是根據訓練語料的多寡來決定，那稱之爲"*動態 GMM*"。而相反的，如果每個狀態下的 GMM 數目是以固定的倍數增加，那我們稱之爲"*固定 GMM*"。

由於 delta-BIC 的計算是依照 GMM 的單位來做的，而聲學單位又是以 HMM 爲單位，因此，在做聲學單位最佳化的過程中，除了以 HMM 爲合併單位之外，我們還用狀態爲單位來合併。以 HMM 爲單位的時候，是否合併，是觀察 HMM 下三個狀態的 delta-BIC 平均值來判斷。而根據我們觀察，有些 HMM 中前兩個狀態其 delta-BIC 都小於零，而第三個卻大於零，但平均之後還是小於零，因此不能合併。爲了讓合併更佳寬鬆，讓每個狀態都能自由的做合併與否的判斷，我們也採用了狀態爲單位的合併。

實驗的目的要觀察不同的聲學模型間的語音辨識率，因此，語言模型的機率在這一系列的實驗中，都讓每個不帶聲調的音節之間的機率設爲相同，也就是說，其爲均勻分

佈。而依此所產生的搜尋網路複雜度爲 924。

（三）結果分析

A. 以專家知識爲本的固定 GMM 與動態 GMM 辨識結果

我們將 Lang-De 與 Lang-In 的聲學模型用來做固定 GMM 與動態 GMM 的比較，結果顯示在<表三>，而相對應的 GMM 總數和平均數列在<表四>。

		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。

		8-mix	16-mix	32-mix	64-mix
動態 GMM	Lang-De (1503)	32,848(7.1)	54,824(12.1)	88,000(17.8)	109,905(24.3)
	Lang-In (1242)	26,328(7.1)	**46,360(12.4)**	69,328(18.6)	98,857(26.5)
固定 GMM	Lang-De (1503)	36,054(8.0)	72,144(16.0)	126,252(28.0)	261,522(58.7)
	Lang-In (1242)	29,787(8.0)	59,616(16.0)	115,506(31.1)	219,834(59.6)

表四、Lang-De 與 Lang-In 的聲學模型用來做固定 GMM 與動態 GMM 的 GMM 總數和平均數(括弧中的值爲平均值)

在<表三>中，可看出動態 GMM 的語音辨識結果比固定的方式的結果來的佳。這表示說，GMM 的增加，必須考慮到實際語音訓練量的狀況，而不能一昧著增加 GMM 的數目。另外，對於動態和固定的方法，其最佳辨識率都出現在每個狀態裡最高有 16 個 GMM 的設定下，之後再增加 GMM 的數目，辨識率反而會下降。而固定方法的下降率會比動態方法的下降率來的多。因此，我們在 Lang-In 用動態方式產生 GMM 於 16-mix 實有最佳的辨識率，爲 64,7%。由於我們所使用的軟體爲 HTK[14]，所以在估計 GMM 參數時，會將每個 GMM 的比重做調整，因此，有些比重較輕的 GMM 會被刪除，而造成固定方法的 GMM 平均數的下降。其中，Lang-De 的聲學數目爲 1503，而 Lang-In 的聲學數目爲 1242。

雖然出動態 GMM 的語音辨識結果比固定的方式的結果好，但<表四>裡，動態方法的總 GMM 數會比固定方法的總 GMM 數來的少，因此，這裡可看出，並不是總 GMM 多，辨識率就會跟著變好。反而倒是要顧慮每個狀態下平均 GMM 數要配合實際的訓練語音量，這樣才能得到好的辯識結果。

B. 以 HMM 爲單位的最佳化聲學模型的結果

由之前的結果得知，用動態的方式來增加 GMM 會有不錯的結果，本小節也是用此方法，但卻是使用本論文提出的結合專家知識與資料驅動的方法來產生最佳化的聲學模型單位。辨識率與 GMM 相關資訊分別列於<表五>和<表六>。第一欄位中的數字爲最

後每個不同限定下所產生的聲學單位數目。

	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

	8mix	16mix	32mix	64mix
C-I (1242)	26,328(7.1)	46,360(12.4)	69,328(18.6)	98,857(26.5)
C-II (527)	12,578(7.8)	24,671(15.6)	41,459(26.2)	55,285(36.4)
C-III (1083)	24,618(7.5)	46,256(14.3)	**78,841(24.3)**	**98,754(30.4)**
C-IV (862)	19,590(7.6)	38,498(14.9)	63,784(24.6)	84,059(32.5)

表六、不同限制下最佳化聲學模型的 GMM 總數和平均數(括弧中的值爲平均值)

C-I 爲共同聲學單位的限制，其聲學數目和 Lang-In 的數目是一樣的，這表示說，具有相同標音符號但隸屬不同語言的聲學單位，其 delta-BIC 值都大於零，因此其聲學單位都會合併。在<表五>中，我們最佳的辨識結果是採用 C-III 的方法，辨識率爲 66.1%，此方法連產生混淆音素都是用 delta-BIC 來做決定的。值得一提的是，利用此方法，從 8-mix 到 64-mix 的結果都是節節上升，和其他方法不同，有的從 16-mix 或 32-mix 之後，辨識率就開始走下坡了。與 Lang-In 或 C-I 來比較，在 32-mix 與 64-mix 時，C-III 都有較佳的辨識率。此外，觀察<表六>我們也可發覺和<表四>有相同的結論，就是 GMM 總數多並不一定代表辨識率一定高。還要配合每個狀態下 GMM 的平均值，才能使辨識率上升。如比較 C-I 和 C-III 在 64-mix 時的辨識率與 GMM 的關係，可以發現這兩個的 GMM 總數差不多，但後者的每個狀態平均 GMM 數卻高於前者，這表示說，C-III 的方法不僅可將發音相似的發音合併，但合併的同時，也把那些本來訓練語料相對少的聲學模型聚合在一起，因爲 delta-BIC 的公式中也考慮到合併後和合併前的訓練語料出現的數目，因此在這兩個因素下，才能將辨識率向上提升。而 C-II 的方法，雖然在每個狀態下有最多 GMM 平均數，但最後的聲學模型數目是原始的三分之一左右，使得聲學單位在這麼少情況下能有良好的鑑別率，而造成在<表五>裡其辨識率是進陪末座。

C. 以狀態爲單位的最佳化聲學模型的結果

以上C-I到C-IV的聲學模型都是以HMM作爲合併的單位，但delta-BIC的計算是以狀態爲單位，因此這裡我們做了比較有彈性的變動，將合併單位從HMM轉爲狀態。語音辨識結果呈現在<表七>。其中C-I與C-II的結果和<表五>相同，這表示說，這兩個的delta-BIC值，其每個狀態是否要合併和以HMM爲單位是一樣的。第一欄位中的數字爲狀態數目。

而另外一方面，在傳統聲學模型的合併裡，有相關學者是採用決策樹(decision tree)的方法，用一些語音學、或語言學上的專家知識規則，來將聲學模型做做分類，相同類的聲學模型就互相分享訓練語料[16]。因此在這裡，我們也利用這樣的技術，來和本篇

所提出的方法來做比較。

而決策樹和 C-III 的方法最大差別就在於，1.前者是使用最大相似度法則而後者是採用 delta-BIC 來做分類以及尋找混淆音素。使用最大相似度法則時，需要制訂一個門檻值，好讓決策樹在分類的時候能停止分裂，不同的門檻值會影響到最後聲學模型的狀態數目，或說是狀態的解析度。而 delta-BIC 則不用到門檻值，而是採用一個固定的懲罰值來替代，而這個值的大小，根據我們的實驗經驗，並不會對最後的狀態結果有很大的區別。所以我們用 delta-BIC 來做聲學模型最佳化的時候，考慮到其 delta-BIC 值是否小於 0，並不用再設定其他的參考值來做最佳化。可是在決定不同決策樹的門檻值時，其最後結果（狀態數目）也會跟著不相同。因此我們試著用幾組不同的門檻值來做停止分裂的條件，把最好的辨識結果呈現在表七的最後一行。

	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。

整體來說，使用狀態爲合併單位所訓練出來的聲學模型(C-III 與 C-IV)，其辨識率比以 HMM 爲合併單位所訓練出來的聲學模型來的好。另外使用決策樹方法所做出來的效果，在 8-mix 的時候有超過 C-III 的結果，但在增加 GMM 的數目之後，其效果就沒有 C-III 來的好，但也是比 C-I,C-II 和 C-IV 來的好。分析如下：決策樹裡的問題共分三大類，其分別爲：子音、母音、和語言議題，在這些議題下，我們總共產生 124 個問題來將這些聲學模型做分類。有了這些的問題，我們也就不用如 C-I 到 C-IV 的限制了。所以，在這些限定下，決策樹還是有他的優點，因爲是一次考慮到這四類的問題，而不是如 C-I 到 C-IV，是一次只考慮到單一種狀況。決策樹的樹頭是整個左右相關聲學模型(3726)，最後則剩下 3374 個狀態。但在最後結果上，還是 C-III 的 64-mix 勝出，同時這也產生了本論文的最佳辨識率 66.7%。

五、結論

本論文提出了一套整合專家背景知識與實際語音分析的方法，利用聲韻學和語音學的知識，將 AHC 的結構樹做分類的限定，再由 delta-BIC 的判斷來把同一棵樹中的聲學單位做合併，並且找出最佳數目的一組新的聲學單位，再將他們以 HMM 重新訓練，實驗在 ForSTDA01 的華台雙語語料庫。實驗結果驗證了，1.利用動態方式增加 GMM 的辨識率比固定方式的方法來的高。2.結合專家知識與資料驅動的方法所訓練的新聲學模型，其辨識率比只用專家知識所訓練的新聲學模型來的佳，在訓練的同時，能提高每個狀態下

平均 GMM 的個數。3.利用狀態爲合併單位比 HMM 爲合併單位，更能充分的反映出哪些聲學模型需要做合併，產生更有彈性的合併，而達到最佳的辨識效果。

參考文獻

[1] T. Schultz and A. Waibel, "Multilingual Cross-lingual Speech Recognition," in Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[2] Brian Mak and Etienne Barnard, "Phone clustering using the Bhattacharyya distance," in Proc. of ICSLP 1996, pp. 2005-2008.

[3] Chung-Hsien Wu, Yu-Hsien Chiu, Chi-Jiun Shia, and Chun-Yu Lin, "PHONE SET GENERATION BASED ON ACOUSTIC AND CONTEXTUAL," in Proc. of ICASSP 2006

[4] Mathews, R. H., 1975. Mathews' Chinese-English Dictionary, Caves, 13th printing.

[5] A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian information criterion," in Proc. EUROSPEECH 1999, pp. 679–682.

[6] J. C. Wells, "Computer-Coded Phonemic Notation of Individual Languages of the European Community," Journal of the International Phonetic Association, 1989, pp. 32-54.

[7] James L. Hieronymus, "ASCII Phonetic Symbols for the World's Languages: Worldbet," Journal of the International Phonetic Association, 1993.

[8] Ren-yuan Lyu, Min-siong Liang, Yuang-chin Chiang, "Toward Constructing A Multilingual Speech Corpus for Taiwanese (Min-nan), Hakka, and Mandarin," Computational Linguistics and Chinese Language Processing, Vol. 9 (2), 2004, pp. 1-12.

[9] Y. J. Chen, C-H. Wu et al. "Generation of robust phonetic set and decision tree for Mandarin using chi-square testing," Speech Communication, Vol. 38 (3-4), 2002, pp. 349-364.

[10] T.S. Chen, C.C. Lin, Y.H. Chiu and R.C. Chen "Combined Density- and Constraint-based Algorithm for Clustering," In Proceedings of 2006 ICISKE 2006.

[11] Jacob Goldberger and Hagai Aronowitz, "A Distance Measure Between GMMs Based on the Unsented Transform and its Application to Speaker Recognition," in Proc. of EUROSPEECH 2005, pp. 1985-1988.

[12] J. Han, and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000.

[13] LIU Yi and Pascale Fung, "Automatic Phone Set Extension with Confidence Measure for Spontaneous Speech," in Proc. of EuroSpeech 2005.

[14] Phil C. Woodland, Steve J. Young, "The HTK Tied-State Continuous Speech Recogniser," in Proc. of EurpSpeech 1993.

[15] X. Anguera, T. Shinozaki, C. Wooters, and J. Hernando, "Model Complexity Selection and Cross-validation EM Training for Robust Speaker Diarization," in Proc. of ICASSP

2007, Honolulu, HI. April 2007.

[16] Dau-Cheng Lyu, Bo-Hou Yang, Min-Siong Liang, Ren-Yuan Lyu and Chun-Nan Hsu, "Speaker Independent Acoustic Modeling for Large Vocabulary Bi-lingual Taiwanese/Mandarin Continuous Speech Recognition," in Proc. of SST 2002.