結合音長與發音特徵於 GTB 之腔調化語音辨識

Accented Speech Recognition based on Gradient Tree Boosting with Duration and Articulation Features

顏明祺　Ming-chin Yen
國立嘉義大學資訊工程學系
Department of Computer Science and Information Engineering
National Chia-Yi University
s0942795@mail.ncyu.edu.tw

賴柏森　Po-San Lai
國立嘉義大學資訊工程學系
Department of Computer Science and Information Engineering
National Chia-Yi University
s0960404@mail.ncyu.edu.tw

葉瑞峰　Jui-Feng Yeh
國立嘉義大學資訊工程學系
Department of Computer Science and Information Engineering
National Chia-Yi University
ralph@mail.ncyu.edu.tw

摘要

現在利用語音作為輸入的人機介面上，須要考慮到發音變異對語音辨識(speech recognition)所造成的影響。通常是以語者或語音調適(speaker adaptation and speech adaptation)的技術來處理。但是對有腔調(accent)之語者的語音辨識效果較不理想。本論文提出將兩個聲學模型(phone acoustic model)進行合併，來重建聲學模型，來降低腔調化語音的錯誤率。首先建立帶有腔調化特徵之聲學模型，合併到傳統聲學模型中，合併時，使用狀態聯繫(state tying)與梯度漸近樹(gradient tree boosting, GTB)演算法，來調整傳統聲學模型，以重建聲學模型。以便提升腔調話語音的辨識能力。

Abstract

Using speech as the input of the human-machine interface, we need to consider the effect of the pronunciation variations on speech recognition. We usually use the speaker adaptation and speech adaptation technique to solve above problem, but the result of accent speech recognition is not good enough. This paper presents a framework to combine two phone models to reconstruct acoustic model. First we build accented acoustic model and unaccented acoustic model, than we combine the accented acoustic model into the unaccented acoustic model with the state tying and Gradient tree boosting algorithm. We can adjust the unaccented model by this framework, to reconstruct acoustic model and robust the accented speech recognition performance.

關鍵詞：語音辨識，狀態連繫，梯度漸近樹

Keywords: Speech Recognition , State Tying, Gradient Tree Boosting

一、緒論

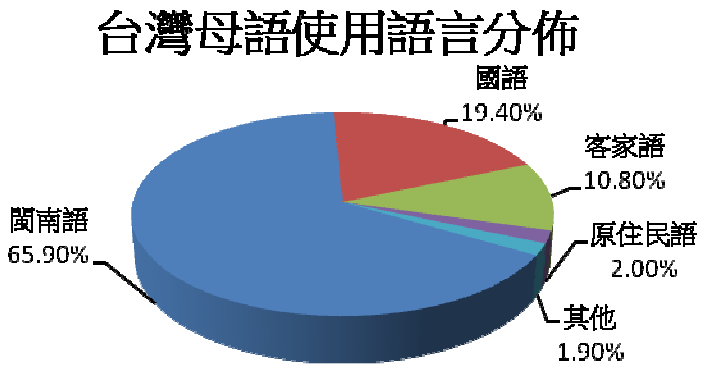
近代的語音辨識系統仍必須針對特定語者(speaker)、特定語言(language)、相同說話風格(speaking style)、特定環境(environment)和領域(domain)的應用，可以擁有高辨識正確率。但語音辨識經常被應用於語音預約系統中，無法在特定語者、相同說話風格及特定環境的情況下建立語音辨識系統，主要造成辨識率不好的原因有下列幾點[1][2]：

1. 前後文相關變異(context variability)：語音會受到字義、前後文以及常識的影響，造成同一段語音可能會有不同的意義或使用方式。
2. 風格變異(style variability)：在連續語音辨識中，有些錯誤是不自覺地發生在日常生活的對話中，與朗誦文章的情況相比，出現錯誤頻率較高。
3. 語者變異(speaker variability)：語音不只帶有語言的資訊，還帶有很多語者本身的資訊，像是年紀、性別、身份背景、健康與情緒，受到前述因素的影響，造成在相同句子下，所發出語音卻有所差異。
4. 說話速度(speaking rate)：說話速度是指每秒的音節數量，當說話速度很快的時候，會造成音節之間無法清楚區分，使得辨識效率變差。
5. 外語和地區方言(foreign and regional accent)：外語和地區方言的影響是造成語者之間差異的要素之一，比較本地語音辨識(native speech recognition)與非本地語音辨識(non-native speech recognition)的效能，就有明顯的差異。
6. 環境變異(environment variability)：由於語者身處的地點不一樣，周遭環境的聲音也會有所不同，像是在車中會有引擎聲、在室內可能會開門聲等，也會影響到辨識的效率。

由此可見，現行語音辨識系統的缺失，唯有從不限定語者、語言、主題和環境的辨識系統著手，方能克竟其功，而這也說明了要建立一個不限定語者、語言、主題和環境的辨識器，是當前重要的課題之一。由於不限定語者、語言、主題和環境辨識器的開發牽涉範圍頗廣，所需時間較長，因此本研究擬將重點放在腔調化語音(accented speech)的辨識上。

二、相關研究

腔調化語音是由於語音受到語者本身母語的影響，造成發音或是文法的錯誤。目前台灣以國語做為官方語言，根據 1993 年國際夏日語言學院(summer institute for linguistics, sil)對台灣母語語言的調查統計資料中[3]，以閩南語為母語有 65.9%，客家語有 10.8%，國語有 19.4%，原住民語的有 2%，如圖一所示。其中以閩南語做為母語有 65.9%，使國語發音會經常受到閩南語發音的影響，這是造成台灣國語的主要原因。駱嘉鵬(2005)則將台灣國語做分析[4]，分別將國語受到閩南語與客家語的影響歸納出發音變異規則。



圖一、母語使用語言分佈圖

過去關於腔調化語音的問題，通常以語者或語音調適(speaker adaptation and speech adaptation)的技術來處理，常見用於發音變異的調適聲學模型參數估測方法有最大後驗法(maximum a posteriori, MAP)以及最大可能性線性迴歸(maximum likelihood linear regression, MLLR)，雖然上述方法在正常語者間的調適已經可以達到一定的成效，然而在具有發音差異(pronunciation variance)，或帶有方言腔調之語者其語音辨認效果較為不理想，導致調適結果不彰[5][6]。

IBM 是最早使用發音字典(pronouncing dictionary)來改善大辭彙連續語音辨識器(large vocabulary continuous speech recognition, LVCSR)的效能[7]，是透過人力與專業知識來找尋變異的規則，來建立發音字典來修正辨識結果。Mirjam Wester 使用決策樹來分析變異的規則[8]，來改善發音變異情形，提出了兩種方法：專家知識的方法(knowledge-based approach)是以各種音韻特徵為基礎，來分析發音變異是由哪些特徵所組成；資料驅動的方法(Data-derived Approach)是透過辨識結果來分析變異的原因。Toshiaki Fukada(1998)發現建立發音字典需要花費大量時間與人力去分析變異規則[9]，提出一個自動產生發音字典的方法，透過類神經網路(neural network)來預測候選發音(alternative pronunciation)，並根據這些候選發音來建立發音字典。

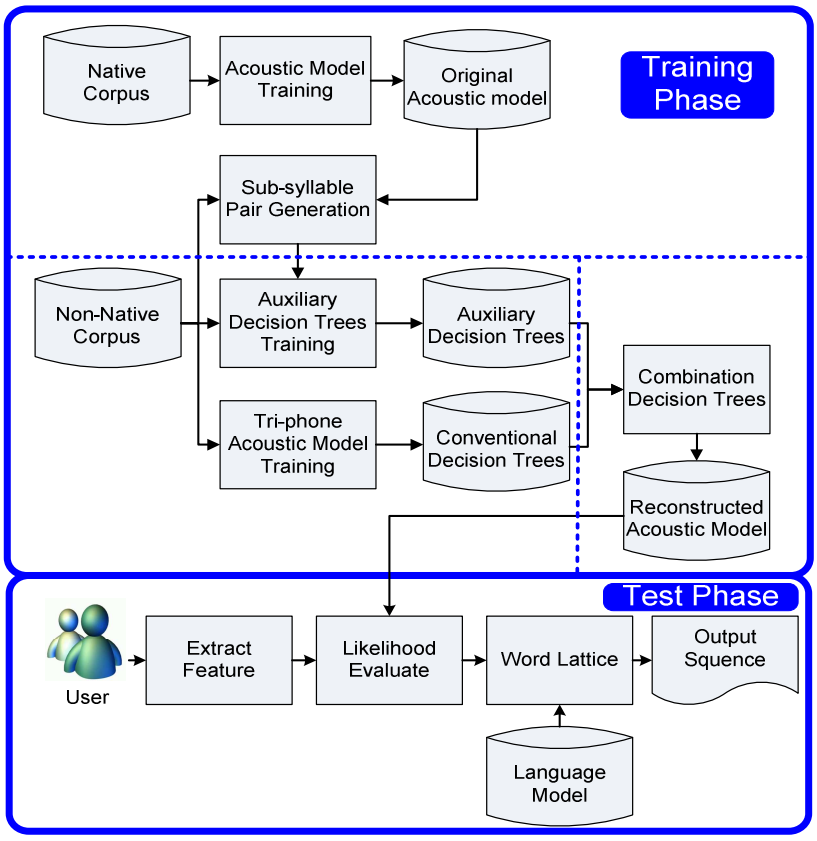
Yuya Akita 和 Tatsuya Kawahara(2005)提到發音模型通常是使用決策樹、類神經網路或是混淆矩陣的架構[10]，但是上述方法未必有使用到上下文相關音素特徵和適當的發音機率，這篇論文提出了重寫變異規則之機率的演算法與可變長度的上下文相關音素特徵的方法，提出了一個利用統計的方式來建立語言模型。呂仁園透過聲學距離(acoustic distance)和音素距離(phonemic distance)來分析發音變異之特性[11]，呂仁園的聲學距離是用馬氏距離(mahalanobis distance)來評估兩音素的聲學模型之聲學特性；音素距離是發音錯誤之機率，根據聲學距離與音素距離的高低將發音變異分成四類：在聲學距離與音素距離皆高的情況下，代表模型之間差距大，不容易辨識錯誤；在高聲學距離低音素距離的情況下，代表這個發音常被唸成其他不同的發音，通常是發生在語者本身的習慣或是協同發音(co-articulation)；在聲學距離與音素距離皆低，此兩音素的發音性質相似，易混淆；最後是低聲學距離高音素距離，是發音特性相近卻不會讓人唸錯的情況。

Liu Yi(2003)將發音變異分成兩類[12]：完全(complete)變異跟部分(partial)變異。完全變異是指整個音節(syllable)都辨識錯誤，通常是發生在語者本身發音錯誤的情況下，並使用發音字典的方式來處理。部分變異是只有聲母(initial)或是韻母(final)錯誤的情況，在國語語音辨識中聲母比韻母更容易辨識錯誤，建立具有發音變異特徵三連音聲學模型(tri-phone acoustic model)，與一般聲學模型合併來重建聲學模型。蔡沛任(2007)針對使用個別建立

混淆矩陣[13]，將目標音與辨識結果做混淆分析，構成發音變異的錯誤規則，用以預測使用者不同音節可能的發音情形，建立音節絡(syllable lattic)，並利用維特比演算法找出最佳的預測路徑，做為語者調適的標記檔，並利用最大後驗法以及最大可能性線性迴歸來調適聲學模型。Stephane Dupont(2006)認為發音變異經常受到鄰近音素的影響所造成的[14]，分別建立前後文相關與前後文無關的聲學模型，整合成發聲模型來處理連音所造成的發音變異。Annika(2006)利用音素(phone)資訊來建立多路徑之音節模型來改善發音變異的情況[15]；呂道誠(2004)藉由觀察台語與國語之間發音的差異，找出其中的規則來改善台灣國語的發音變異[16]。

本篇論文將分成數個部分來探討上面所提到的問題，首先我們會介紹本文提出的系統架構，將腔調化語音利用幾個部分來描述並且解決問題。接著是方法的部分，這裡我們主要是使用了三連音聲學模型梯度漸進樹做為腔調化辨識的模型調整，最後是實驗。我們會歸納實驗結果，探討本論文的發現和未來工作，給予結論。

三、系統架構



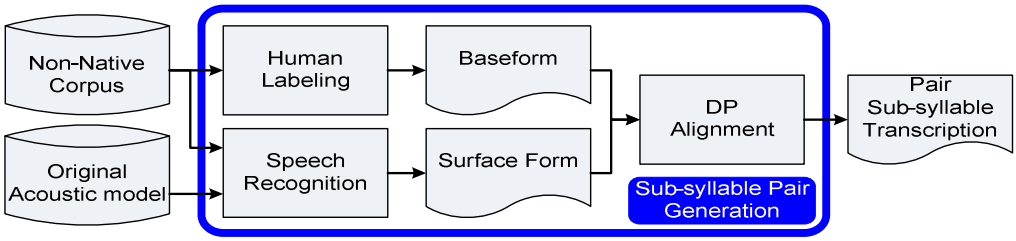
圖二、系統流程圖

（一）、訓練階段

在訓練階段首先使用本地語料庫(native corpus)來訓練出沒有腔調化的初始聲學模型，重建聲學模型主要分成三個步驟：

1. 產生次音節配對 (sub-syllable pair generation)：透過動態規劃批配(dynamic programming alignment)來對應基本結構(baseform)與表層結構(surface form)之間的關

係，產生出帶有腔調化語音特徵的標記檔，稱為次音節配對標記檔(pair sub-syllable transcription)。其中基本結構是透過人工標記(human labeling)的方式而成，是聽每一個音檔內容並將其正確發音標記上去;表層結構則是使用母語語料庫訓練的初始聲學模型去辨識非母語語料庫所得到的辨識結果，如圖四所示。

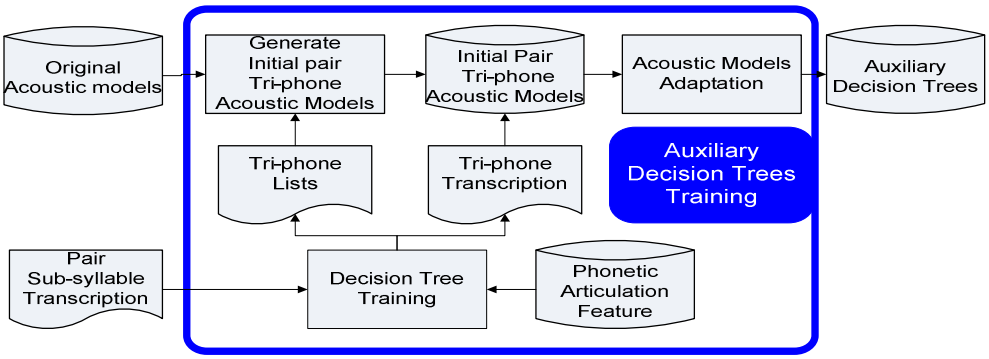


圖三、次音節產生流程圖

2. 聲學模型訓練(tri-phone acoustic model training)：將初始聲學模型透過三連音聲學模型訓練後，分別產生傳統決策樹和腔調化發音特徵的附屬決策樹。以下是分別對兩種決策樹的訓練作說明。

傳統決策樹訓練(decision trees training)：將人工標記檔轉化成帶有前後次音節的三連音標記檔，並將其種類統計成三連音目錄(tri-phone lists)，根據三連音目錄將其相對映的模型從初始聲學模型中複製出來，作為初始三連音聲學模型(initial tri-phone acoustic model)。再經由非母語語料庫與三連音標記檔作聲學模型調適(acoustic model adaptation)後，由於三連音聲學模型中模型的數量過於龐大，因此利用決策樹訓練(decision tree training)來減少模型的數量，使用語音發音特徵作為決策樹訓練的問題集，訓練出傳統決策樹。

附屬決策樹訓練(auxiliary decision trees training)：考慮到附屬決策樹的組合會較傳統決策樹的組合還要增加數倍，所以一開始就使用決策樹從音節配對標記檔中找出常見的錯誤組合，減少模型的數量，並產生出帶有腔調化發音特徵的三連音配對標記檔與目錄，藉由目錄從初始聲學模型中複製出初始三連音配對聲學模型，最後經過調適之後產生附屬決策樹。



圖四、附屬決策樹訓練流程

3. 合併決策樹(combination decision trees)：透過分析模型之分佈來決定是否要將附屬決策樹合併到傳統回歸樹中，或是將附屬決策樹透過梯度漸近樹來調適傳統決策樹，用來重建聲學模型。

（二）、測試階段

1. 首先要擷取語音特徵參數，由於語音信號的資料量過於龐大，必須將信號中擷取帶

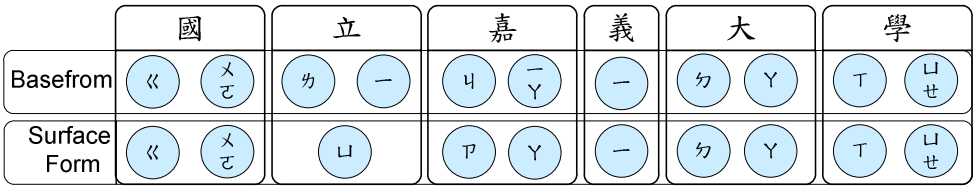
有語音特性的特徵參數，並且將語音信號分割成連續音框(frame)，分別對每一個音框擷取特徵參數，本研究使用 39 維梅爾倒頻譜係數做為語音辨識的特徵參數。

2. 使用這些特徵參數與聲學模型內所有的模型進行相似度評估，並取出前幾名的結果形成字詞網絡(word lattice)。
3. 最後利用語言模型來從字詞網絡中選出一條最佳路徑，輸出為辨識結果。

（三）、匹配

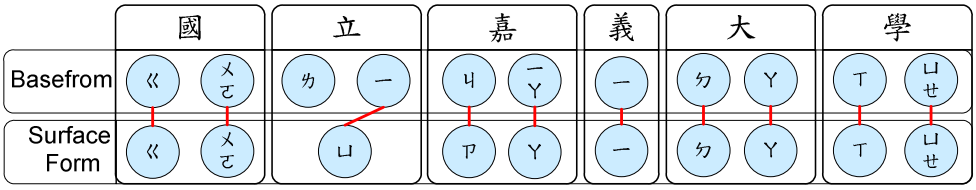
在建立分類與回歸樹之前，將基本結構跟表層結構做匹配是重要的步驟。 匹配通常是用動態規劃來處理，以編輯距離(edit distance)來當作成本函數(cost function)，取最小成本(cost)之路徑做為基本結構跟表層結構之間對應關係，其中基本結構代表正確發音，表層結構代表辨識結果。

一般以次音節(sub-syllable)為單位的動態規劃配置中的編輯距離是不足以當作成本函數，可能會導致錯誤對應關係。以「國立嘉義大學」為例子，如圖六所示。

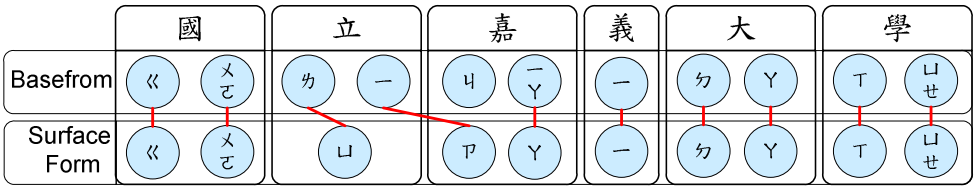


圖五、「國立嘉義大學」之基本結構與表層結構

以子音節為單位做一對一匹配，則基本結構與表層結構可能的對應關係如圖六所示。雖然這兩組對應關係的成本相同，但圖七的對應關係中「ㄌ→ㄩ」和「ㄧ→ㄗ」都是很明顯的錯誤對應關係，所以第二組對應關係是錯的。



圖六、子音節正確對應結果



圖七、子音節錯誤對應結果

我們使用音素特徵距離(phone feature distance)來解決這個問題。首先根據音素的發音類型與發音方式，將聲母跟韻母分別分類成 9 類跟 14 類如表一所示，分類表中若兩分類如果距離越近相似度就越高，反之距離越遠則相似度越低。並由分類資訊來決定每一個音素與其他音素的特徵距離，如果兩音素的分類性質越相似則成本越低，反之分類性質越不相似則成本越高。

表一、聲母與韻母分類表

分類名稱	韻母名稱	分類名稱	聲母名稱
F_Central_a	a(ㄚ),ya(ㄧㄚ),wa(ㄨㄚ)	I_UnAspStop	b(ㄅ),d(ㄉ),g(ㄍ)
F_Vback_a	ao(ㄠ),yao(ㄧㄠ)	I_AspStop	p(ㄆ),t(ㄊ),k(ㄎ)
F_Vfront_a	ai(ㄞ),yai(ㄧㄞ),wai(ㄨㄞ)	I_Feric	h(ㄏ),f(ㄈ)
F_Front_a+n	an(ㄢ),yan(ㄧㄢ),wan(ㄨㄢ),yuan(ㄩㄢ)	I_Nasal	m(ㄇ),n(ㄋ)
F_Vowel_a+ng	ang(ㄤ),yang(ㄧㄤ),wang(ㄨㄤ)	I_Lateral	l(ㄌ)
F_Vng	eng(ㄥ),ying(ㄧㄥ)	I_Dorsal	j(ㄐ),q(ㄑ),x(ㄒ)
F_Vn	en(ㄣ),yin(ㄧㄣ),wen(ㄨㄣ),yun(ㄩㄣ),wu(ㄨ)	I_Dentalvelar	z(ㄗ),c(ㄘ),s(ㄙ)
F_Front_e	ei(ㄟ),wei(ㄨㄟ)	I_RetroAff	Zh(ㄓ),ch(ㄔ),sh(ㄕ),r(ㄖ)
F_2V_e	e(ㄜ),ye(ㄧㄝ),yue(ㄩㄝ),er(ㄦ)	I_Zero	ini3,ini4,ini5,ini6,ini7
F_Central_o	o(ㄛ),wo(ㄨㄛ),yo(ㄧㄛ)		
F_Mid_o	ou(ㄡ),you(ㄧㄡ)		
F_2Vng	weng(ㄨㄥ),yong(ㄩㄥ)		
F_V	yi(ㄧ),yu(ㄩ)		
F_Zero	fin1,fin2		

四、重建聲學模型

（一）、分類與迴歸樹(classification and regression tree, CART)

建立分類與迴歸樹是用以預測表層結構中音素可能的發音變異，透過 yes/no 的問題將每一個音素分屬到 CART 的葉節點。其中問題集主要是有關音素之發音特徵及附近音素之種類。首先使用動態規劃匹配來找出基本結構與表層結構之間對應關係，將得對應關係透過問題集來得到相對應的答案，並利用這些答案來建立 CART。這裡將聲母跟韻母各建立一棵 CART 來當作發音模型，並用來處理發音變異，一共建立 153 棵 CART。

當建立完決策樹之後，將表層結構每一個次音節去執行個別之 CART，會根據與上下文子音節關係，得到發音變異候選音素，依照得到每個子音節的候選音素，建立候選音素網絡(lattice)圖。最後我們利用信心度量測(confidence measure)來從候選音素網絡中找出信心度最高的路徑，做為最終辨識結果。這裡所使用的信心度量測之目的是將較可靠的辨識結果保留，並用來做為推論其他不確定音段或訊框(frame)。這裡 W 為辨識結果之次音節，X 為輸入語音特徵集合。信心量測表示如下算式 1：

$$P(W \mid X) = \frac{P(W)P(X \mid W)}{\sum_{W} P(W)P(X \mid W)} \tag{1}$$

（二）、結合樹狀結構與資料驅動之分群法

我們將利用樹狀結構為主之分群法產生兩個三連音聲學模型，其中一個是一般常見的三連音聲學模型，稱為傳統決策樹(standard decision tree)；另一個則是具有基本結構(baseform)與表層結構(surface form)對應關係的三連音聲學模型，稱為附屬決策樹(auxiliary decision tree)，但是附屬決策樹無法直接使用在聲學模型的評估中，因為附屬決策樹的複雜不比傳統決策樹更加龐大，所以利用資料驅動之分群法將附屬決策樹之葉節點合併到傳統決策樹之葉節點中，並分析要合併之葉節點群集的分佈，來決定是透過聯繫狀態(tied-state)合併到傳統決策樹中，或是使用梯度漸近樹演算法來調適傳統決策樹之葉節點。

1. 產生附屬決策樹

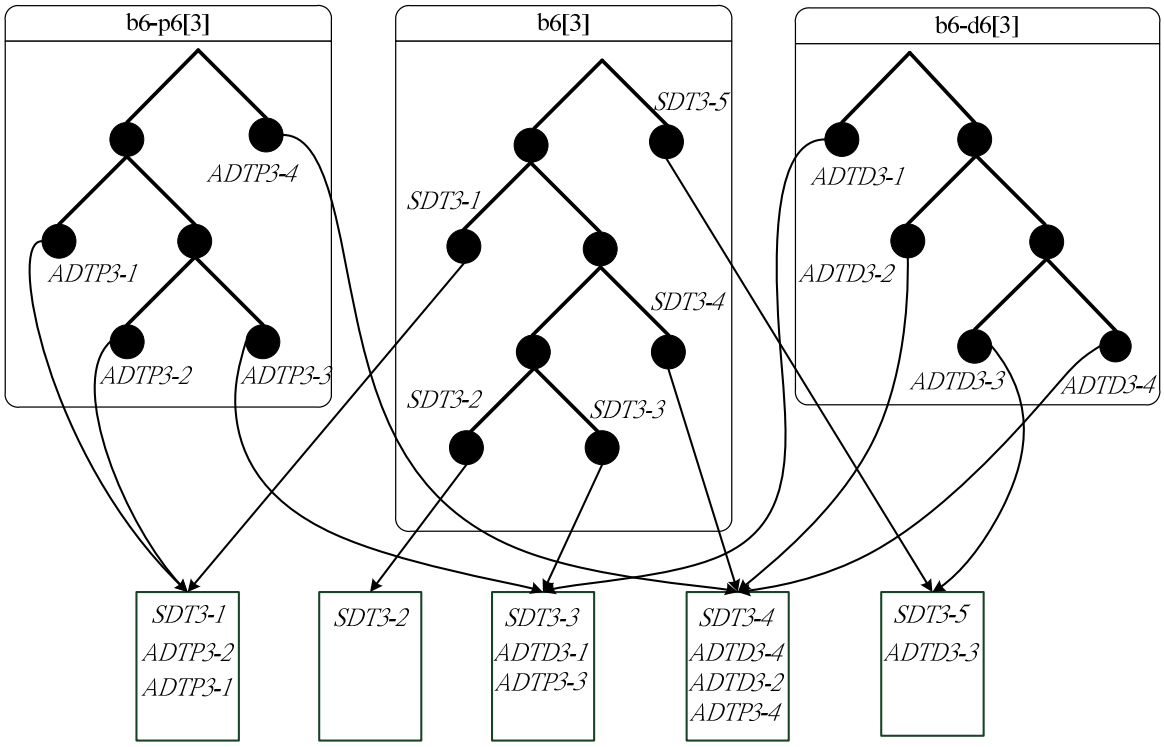
我們將語者發音特性的表層結構對應到標準發音的基本結構，而表層結構是使用原始聲學模型的辨識結果，並加上基本結構中前後音素成為三連音模型的目錄及對應標記檔。將基本結構與表層結構透過前面提到的動態規劃配置，產生出帶有腔調化發音變異的三連音對應標記檔及目錄。

根據三連音聲學模型的目錄從原始聲學模型中複製出最初的三連音聲學模型，例如：g-wa_wai+n 是複製原始聲學模型中 wa 的音素模型作為最初音素模型。並根據對應標 記檔透過波氏(Baum-Welch)演算法來調適三連音聲學模型。最後將三連音聲學模型透過以樹狀結構為主的分群法，建立成決策樹成為附屬決策樹，所使用的問題集是有關中文的發音特徵。

2. 合併決策樹

我們將附屬決策樹合併到傳統決策樹中，來提升聲學模型對腔調化語音的處理能力；首先利用資料驅動之分群法的概念，透過每個葉節點之間的距離評估，將附屬決策樹之葉結點分類到距離最近的傳統決策樹之葉節點中，之後分析要合併葉節點之分佈情況，來決定是要用狀態聯繫或是透過梯度漸近樹演算法來調適傳統決策樹，來重建聲學模型。實際上，有些傳統決策樹之葉節點會對應到數個附屬決策樹之葉節點，而有些傳統決策樹之葉節點則沒有對應到附屬決策樹，而對應數量的多寡是基於此狀態的混淆程度，換句話說，當對應到的葉節點很多時，代表此狀態混淆程度很高，圖八是合併附屬決策樹的示意圖。

如圖八所示，b[3]是傳統決策樹中 b 模型之第三狀態的決策樹，而 b_g[3]與 b_p[3]分別附屬決策樹中基本結構\b\對應到表層結構\p\、\g\第三狀態的決策樹，將每一棵附屬決策樹之葉節點合併到傳統決策樹之葉節點中，我們可以觀察到葉節點 ST3-3 同時跟兩顆不同的附屬決策樹之葉節點合併(例如：b_g[3]與 b_p[3])，而 ST3-2 並沒有與任何一個附屬決策樹之葉節點合併。透過附屬決策樹的合併，可以調整原本聲學模型內高斯混和分佈，使得部分發音錯誤的情況可以改善。



圖八、聲學模型重建

2.1 距離評估

首先，我們必須要將找出附屬決策樹之葉節點是分屬到哪傳統決策樹，評估標準根據是兩節點最小高斯距離(minimum Gaussian distance)來決定，當兩節點擁有最小距離時，此兩節點為一組。而最簡單評估兩節點距離的方式是使用歐基里德距離(Euclidean distance)，但歐基里德距離無法表現出節點整體的分布關係，所以使用馬氏距離(Mahalanobis distance)來離克服這個問題。

葉節點是由高斯混和分布所組成，假設 G 與 H 是相同傳統決策樹中的兩個葉節點，而 F 為附屬決策樹的葉節點，我們以 N(μ,σ)代表高斯分佈，其中 μ 為平均值與 σ 為變異數。因為 G 與 H 是不同的高斯分佈，所以距離 D(H,F)與 D(G,F)也應該不同，但是馬氏距離是將 G 與 H 看做單一高斯分佈，表示成 G'(μ,σ)與 H'(μ,σ)，新的平均值與變異數為原本的加權平均值，此時 G'(μ,σ)與 H'(μ,σ)可能會有相同的平均值與變異數，造成 D(H,F)與 D(G,F)距離相同，這裡我們使用 KL 距離(Kullback-Leibler distance, KLD)來解決。

我們將輸入音檔的特徵參數定義成 x，傳統決策樹的葉節點定義成 b，附屬決策樹的葉節點定義成 s，P(x|b)與 P(x|s)為輸出在狀態 b 與狀態 s 的分佈，這些狀態分佈是高斯混和分佈所組成，如式(2)所示。

$$P(x \mid b) = \sum_{k=1}^{K} w_{bk} N(x; \mu_{bk}, \Sigma_{bk}) \tag{2}$$

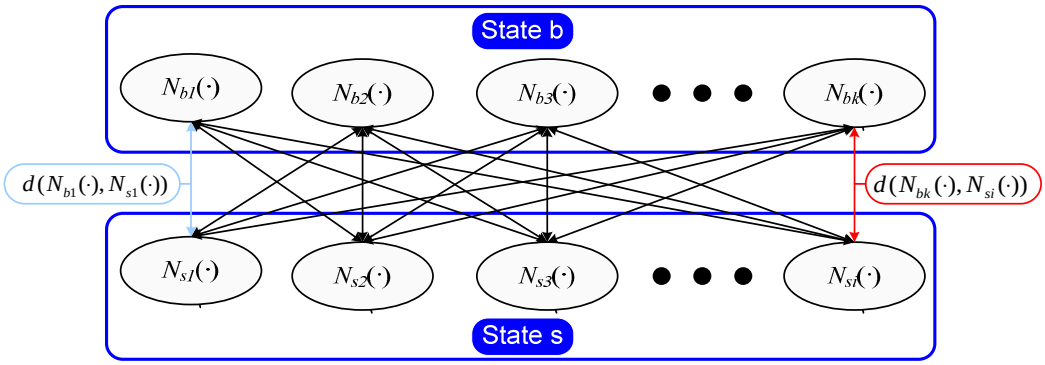
W_{bk} 與 w_{si} 是第 k 個或是第 i 個高斯分佈的權重值，其加總為 1。高斯 $N_{bk}(\cdot)$與高斯 $N_{si}(\cdot)$之間的距離則表示成 d($N_{bk}(\cdot)$, $N_{si}(\cdot)$)，最後將狀態 b 與 s 之間的距離表示成式(3)，其中權重值 w_{ik} 是 w_{bk} 與 w_{si} 的乘積。

$$D(b,s)=\min_{W=[w_{ik}]}\sum_{i=1}^{N}\sum_{k=1}^{K}w_{ik}d(N_{bk}(\cdot),N_{si}(\cdot))\quad(w_{ik}>0,1\le i\le N,1\le k\le K) \tag{3}$$

$$\sum_{i=1}^{N}w_{ik}=w_{bk},1\le k\le K \tag{4}$$

$$\sum_{i=1}^{K}w_{ik}=w_{si},1\le i\le N \tag{5}$$

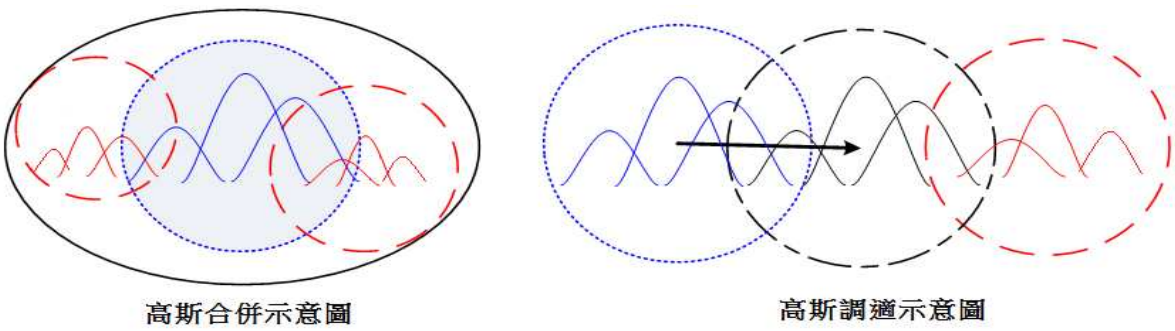
另外，我們可以將距離的量測表示成圖九，將狀態 b 全部高斯分佈與狀態 s 內全部高斯分佈分別計算的距離，將全部的距離總和即為兩狀態之間的距離。



圖九、兩狀態間之距離評估

2.2 狀態分析分佈

為了判斷是要將附屬決策樹合併到傳統決策樹中，或是用來調適傳統決策樹，是透過分析兩模型的分佈情況來決定，如果兩模型分佈有互相重疊的情況，則將附屬決策樹合併到傳統決策樹中。圖中藍色圈是表示模型「丫」的分佈，兩個紅色圈是當發音「丫」變異為「一丫」的分佈情形，可以觀察到兩分佈有重疊情況發生，將兩分佈用狀態聯繫合併起來，合併完分佈會調整成黑色圈的分佈。如果兩模型分佈沒有重疊的現象發生，此時會透梯度漸近樹演算法來調適「丫」模型的分佈。上述情形如圖十所示。



圖十、高斯合併與調適示意圖

3. 重建聲學模型

3.1 梯度漸近樹演算法

本論文使用梯度漸近樹(gradient tree boosting, GTB)演算法來調適聲學模型[17]，利用迴歸樹(regress tree)來取代一般傳統作法。GTB 演算法是一種 meta 演算法，重複地用目前預測函式之剩餘誤差來當作下一棵決策樹的訓練資料，並利用簡純的累加方式來合併迴歸樹來增強預測功能。接下來先介紹梯度漸近樹演算法。

首先是定義特徵變數 $\boldsymbol{x}=(x_1,x_2,\ldots,x_k)$，本研究使用的特徵參數為次音節的發音特徵；$y$ 是次音節的音長時間。$\{y_i, x_i\}_1^N$ 是代表共有 N 組的訓練資料，每組訓練資料包含特徵變數與音長。首先從訓練資料中建立出 M 棵不同的迴歸樹表示成 $h(x,a_1),\cdots,h(x,a_M)$，a 是迴歸樹中間節點之問題集，再從樹群中選擇出最適合的迴歸樹，依序加入預測函數 $F(x)$中，形成一個累加函數，如式(6)所示：

$$F(x)=\beta_0+\sum_{m=1}^{M}\beta_m h(x,a_m) \tag{6}$$

其中 β_m 和 a_m 分別為第 m 棵決策樹 $h(x, a_m)$之權重值與變數參數，而 β_0 是起始數值，所以與參數 a_m 是依序從 m=1 到 m=M 由損失函數(loss function)$\Psi(y, F(x))$中取最小值，損失函數是使用最小平方損失函式，如式(7)所示。

$$\Psi(y,F)=(y-F)^2/2 \tag{7}$$

接下來把從第一棵迴歸樹到第 m-1 棵迴歸樹，定義成一個累加函數 $F_{m-1}(x)$，其中第 m 棵迴歸樹的權重值 β_m 與參數 a_m 被定義成式(8)：

$$(\beta_m,a_m)=\arg\min_{\beta,a}\sum_{i=1}^{N}\Psi\left(y_i,F_{m-1}(x_i)+\beta h(x_i,a)\right) \tag{8}$$

這裡我們將初始值 $F_0(x)=\bar{y}$，$\bar{y}$ 為全部訓練資料模型之平均數。但是一般的情況下，我們不能直接的從式(8)得到權重值 β_m 與參數 α_m。所以 GTB 分別將權重值 β_m 與參數 α_m 以兩步驟方法來做近似值的評估，在評估參數 α_m 的近似值方面，我們藉由目前 $F_{m-1}(x)$的梯度值透過最小平方誤差(least-square error)來求得迴歸樹近似值，如下式(9)所示：

$$a_m=\arg\min_{a}\sum_{i=1}^{N}\left(\tilde{y}_{im}-h(x_i,a)\right)^2 \tag{9}$$

其中 $\tilde{y}_{im}$ 是梯度值表示成式(10)。

$$\tilde{y}_{im}=-\left[\frac{\partial\Psi(y_i,F(x_i))}{\partial F(x_i)}\right]_{F(x_i)=F_{m-1}(x)} \tag{10}$$

代入損失函數 $\Psi(y, F(x))$可得到式(11)。

$$\tilde{y}_{im} = y_i - F_{m-1}(x_i) \tag{11}$$

當第 m 棵迴歸樹所使用參數 α_m 後可得到葉節點 L_m，所以迴歸樹可以表示成式(12)。

$$h\left(x, \{R_{lm}\}_{l=1}^{Lm}\right) = \sum_{l=1}^{Lm} \bar{y}_{lm} \mathrm{I}(x \in R_{lm}) \tag{12}$$

R_{lm} 是表示第 m 棵迴歸樹中第 l 個葉節點所分離的區域。I(・)是布林函數，假如條件是成立的就輸出為 1。$\bar{y}_{lm}$ 是一個常數，定義成在分屬在第 m 棵迴歸樹中第 l 個葉節點的訓練資料之平均值。由於在迴歸樹中 $\bar{y}_{lm}$ 是常數，所以權重值 β_m 可以直接利用損失函式用線性搜尋來評估出來。因此我們可以得到一個新的累加函式如下式(13)：

$$F_m(x) = F_{m-1}(x) + v\sum_{l=1}^{Lm} \gamma_{lm} \mathrm{I}(x \in R_{lm}) \tag{13}$$

五、實驗結果分析

（一）、實驗語料與辨識器設定

1. 語料

語料是由國立成功大學醫院掛號語料為劇本，為了符合台灣國語之腔調化語音的需求，而重新錄製，共 4,413 句、31,029 個音節，用來訓練三連音聲學模型，取劇本中一個主題共 700 句做為訓練語料，並將全部語料做為測試語料，其相關資訊如表二所示。

表二、訓練語測試語料內容

	訓練語料	測試語料
句數	700	4,413
音節數	4,480	31,029

2. 辨識器設定

辨識器的聲學模型是由 TCC-300 麥克風語料庫訓練而成，TCC-300 是由國立台灣大學、國立成功大學與國立交通大學共 300 人錄製而成，總句數為 8,913 句，語音訊號取樣頻率與量化析度分別為 16KHz 與 16 位元，使用英國劍橋大學的 HMM Tool Kit (HTK)做為聲學模型之訓練工具，其中包含 162 個次音節模型，其中包含 38 個韻母模型。聲母模型則根據相連韻母的不同，細分成 112 個右相關聲母模型，如表三所示，還有 11 個填空詞模型與 1 個靜音模型，所使用的特徵參數是透過 39 維梅爾倒頻譜參數(mel-frequency cepstral coefficients, MFCC)，每個音框大小為 32ms，位移量為 10.625ms，音檔的取樣率為 16K，並不取倒濾波(cepstral liftering)。

表三、韻母聲母分類表

類別	聲母	搭配韻母類別	類別	韻母
1~4 類	“ㄓ”, “ㄔ”, “ㄕ”, “ㄖ”	1 類	1、2 類	“空 1”, “空 2”
5~7 類	“ㄗ”, “ㄘ”, “ㄙ”	2 類	3~7 類	“ㄚ”, “ㄞ”, “ㄠ”, “ㄢ”, “ㄤ”
8~26 類	“ㄅ”, “ㄆ”, “ㄇ”, “ㄈ”, “ㄉ”, “ㄊ”, “ㄋ”, “ㄍ”, “ㄎ”, “ㄏ”, “ㄓ”, “ㄔ”, “ㄕ”, “ㄖ”, “ㄗ”, “ㄘ”, “ㄙ”, “ㄧ”, “ㄨ”, “ㄩ”, “ ”	3~7 類	8、9 類	“ㄛ”, “ㄡ”
27~45 類	“ㄅ”, “ㄆ”, “ㄇ”, “ㄈ”, “ㄉ”, “ㄊ”, “ㄋ”, “ㄍ”, “ㄎ”, “ㄏ”, “ㄓ”, “ㄔ”, “ㄕ”, “ㄖ”, “ㄗ”, “ㄘ”, “ㄙ”, “ㄧ”, “ㄨ”, “ㄩ”, “ ”	8~9 類	11~14 類	“ㄜ”, “ㄣ”, “ㄥ”, “ㄦ”
46~57 類	“ㄅ”, “ㄆ”, “ㄇ”, “ㄈ”, “ㄉ”, “ㄋ”, “ㄍ”, “ㄎ”, “ㄏ”, “ㄓ”, “ㄔ”, “ㄕ”, “ㄖ”	10 類	15~24 類	“ㄧ”, “ㄧㄚ”, “ㄧㄝ”, “ㄧㄞ”, “ㄧㄠ”, “ㄧㄡ”, “ㄧㄢ”, “ㄧㄣ”, “ㄧㄤ”, “ㄧㄥ”, “ㄧㄦ”
58~76 類	“ㄅ”, “ㄆ”, “ㄇ”, “ㄈ”, “ㄉ”, “ㄊ”, “ㄋ”, “ㄍ”, “ㄎ”, “ㄏ”, “ㄓ”, “ㄔ”, “ㄕ”, “ㄖ”, “ㄗ”, “ㄘ”, “ㄙ”, “ㄨ”, “ ”	11~14 類	25~33 類	“ㄨ”, “ㄨㄚ”, “ㄨㄛ”, “ㄨㄞ”, “ㄨㄟ”, “ㄨㄢ”, “ㄨㄣ”, “ㄨㄤ”, “ㄨㄥ”, “ㄨㄦ”
77~87 類	“ㄅ”, “ㄆ”, “ㄇ”, “ㄉ”, “ㄊ”, “ㄋ”, “ㄌ”, “ㄍ”, “ㄏ”, “ ”,	15~24 類	34~38 類	“ㄩ”, “ㄩㄝ”, “ㄩㄢ”, “ㄩㄣ”, “ㄩㄥ”
88~106 類	“ㄅ”, “ㄆ”, “ㄇ”, “ㄈ”, “ㄉ”, “ㄊ”, “ㄋ”, “ㄍ”, “ㄎ”, “ㄏ”, “ㄓ”, “ㄔ”, “ㄕ”, “ㄖ”, “ㄗ”, “ㄘ”, “ㄙ”, “ ”	25~33 類		
107~112 類	“ㄋ”, “ㄌ”, “ㄍ”, “ㄎ”, “ㄏ”, “ ”,	34~38 類		
113~114 類	“S, “N	n/a		

（二）、實驗與評估

以下是各系統配置的實驗結果，我們使用 TCC-300 訓練出的聲學模型做為 baseline，使用重新錄製成功大學醫院掛號語料中的 700 句來建立帶有腔調之聲學模型，並且與 MLLR 之方法比較。最後使用本論文提出的架構，使用 GTB 演算法和狀態聯繫的方法來結合非腔調化與腔調化之聲學模型，用以提升對腔調化語音的辨識能力。測試語料則為重新錄製後的成功大學醫院掛號語料。其錯誤率如表四所示。

表四、各種設置對腔調話語音之錯誤率

System	Ins	Del	Sub	Error rate
Baseline	8.26%(5461)	3.75%(2360)	32.86%(21707)	44.71%
MLLR	6.35%(4050)	2.86%(1826)	27.47%(17531)	36.68%
State tying	6.47%(4203)	2.3%(1541)	26.8%(17405)	35.64%
GTB	5.91%(3756)	1.78%(1334)	23.86%(15144)	31.89%
GTB +State tying	5.57%(3512)	2.28%(1437)	22.6%(14233)	30.46%

根據實驗結果，對於腔調化語音的辨識，透過帶有腔調化語音特徵聲學模型，利用資料驅動之方式附加在無腔調化語音聲學模型的分部群集中。接著分系各群集的分布情況，若分布較為分散，則使用狀態聯繫來重建聲學模型，反之則使用 GTB 演算法來調適聲學模型。一般的辨識錯誤主要在替代部分的改善，由表四中我們可以看到，Baseline 與 MLLR 方法中的錯誤率分別為 44.71%與 36.68%。而分別對聲學模型使用狀態聯繫與狀態調適的方法，錯誤率分別為 35.64%與 31.89%。最後我們同時結合了狀態聯繫與狀態調適，將錯誤率下降至 30.46%，與 MLLR 方法比較，減少了 6.22%錯誤率，較原本提升了 16.95%。由此可以看出，對於腔調化語音辨識系統的改善，狀態調適可以提供顯著的效果。

六、結論與未來研究方向

本論文提出透過以樹狀結構為主之分群法建立兩個三連音聲學模型，這兩個聲學模型分別是傳統的三連音聲學模型，和帶有腔調化語音特徵的三連音聲學模型，我們後者透過資料驅動之分群法歸屬到前者之群集中，之後分析群集的分佈情況，如果分布較為分散，則使用狀態聯繫來重建聲學模型，如果分布有相同的趨勢，則使用裝態調適來調適聲學模型。經由分布情況來決定狀態調適還是狀態聯繫可以減少錯誤率以提升系統效能。

本研究目前只考慮到閩南語對國語所造成的腔調化語音，客家語也是母語的主流之一，可將客家語對國語所造成影響也納入特徵考量中，以增加辨識器的泛用性。另外考慮國語對閩南語的影響，並建立國台雙語辨識器。

參考文獻

[1] X. Huang, A. Acero, H. W. Hon, *Spoken Language Processing: A Guide to Theory Algorithm and System Development*, Prentice Hall, May 5, 2001.

[2] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, “Impact of variabilities on speech recognition”, SPECOM 2006, pp. 3-16.

[3] SIL International Partners in Language Development, http://www.sil.org/.

[4] Patil, A.; Gupta, C.; Rao, P., “Evaluating vowel pronunciation quality: Formant space matching versus ASR confidence”, *National Conference on Digital Object Identifier*, 2010, pp. 1-5.

[5] Lee, Chung-Han; Wu, Chung-Hsien; Guo, Jun-Cheng., “Pronunciation variation generation for spontaneous speech synthesis using state-based voice transformation”, *Acoustics Speech and Signal Processing, International Conference on Digital Object Identifier,* 2010, pp. 4826-4829.

[6] 駱嘉鵬, “閩客方言影響下的台灣國語音韻特點”, 第一屆馬來西亞漢語語言學國際學術會議, 2005.

[7] L. Lamel and G. Adda, O*n Designing Pronunciation Lexicons for Large Vocabulary Continuous Speech Recognition*, ICSLP 1996, pp.6-9.

[8] M. Wester, “Pronunciation Modeling for ASR-knowledge-based and Data-driven Methods”, *Journal of Computer Speech and Language,* 2003, pp. 69-85.

[9] Fukada, T., Yoshimura, T., *Sagisaka, Y.,”*Automatic generation of multiple pronunciations based on neural networks and language statistics*”*, In MPV-1998, pp. 41-46.

[10] K.-T. Lee, L. Melnar and I. Talley, “Symbolic Speaker Adaptation for Pronunciation Modeling”, PMLA-2002, pp. 24-29.

[11] M.-Y. Tsai and L.-S. Lee,” Pronunciation variations based on acoustic phonemic distance measures with applications examples of Mandarin Chinese*”*, *Automatic Speech Recognition and Understanding* ,2003, pp. 117-122.

[12] Y. Liu and P. Fung, “Partial change accent models for accented Mandarin speech recognition”, *Automatic Speech Recognition and Understanding,* 2003, pp.111-116.

[13] Tsai, Pei-Jen, “Error Correction and Feedback Using Phonetic Attribute Analysis for Articulation Disorders”, *NCKU*,2007, http://ir.lib.ncku.edu.tw/handle/987654321/21033.

[14] S. Dupont, C. Ris, L. Couvreur, J.M. Boite, “A study of implicit and explicit modeling of coarticulation and pronunciation variation”, *Speech Recognition and Intrinsic Variation Workshop,* 2006, pp.1353-1356.

[15] A. Hamalainen, L. ten Bosch, and L. Boves, “Pronunciation Variant–Based Multi-Path HMMs for Syllables”, *Proceedings of Interspeech,* 2006, pp. 17-21.

[16] 呂道誠, 謝鴻文, 李勇憲, 劉仲英, 許鈞南, 江永進, 呂仁園, “華台雙語發音變異性之語音辨識研究及 PDA 之應用”,*The Association for Computational Linguistics and Chinese Language Processing* , 2004, pp. O04-1024.

[17] J. Yamagishi, H. Kawai, and T. Kobayashi, “Phone duration modeling using gradient tree boosting,” Speech Communication, vol. 50, no. 5, 2008, pp. 405–415.