發音事件驗證於多語辨識發音變異模型之產生

Pronunciation Variation Model Generation based on Pronunciation Event Verification for Multi-Lingual Speech Recognition

蔡佩珊，沈涵平，吳宗憲

國立成功大學 資訊工程學系

pshan08@gmail.com, hanpinsheen@yahoo.com.tw, chunghsienwu@gmail.com

摘要

基於聲學模型的多語辨識器對於辨識標準語音已達相當程度的正確性，但對於發音變異語音辨識的正確性仍嫌不足。由於產生發音變異現象的情況具有相當的多樣性，且偵測變異之聲學模型辨識器的準確度仍存有誤差，這些現象皆影響發音變異模型之產生。本研究基於聚類狀態模型(Senone Model, SEM)，在狀態(State)的單位下觀察聲音訊號的變化，藉以較準確的模擬細微發音變異現象。另外也提出了基於發音事件的非監督式發音事件模型建立，透過此模型進一步驗證基於聲學模型的偵測變異驗證結果，以改善聲學特性所造成的錯誤偵測。對於發音變異模型進行之實驗結果顯示，本論文所提出之方法，對發音變異語句之辨識正確率上，具有相當程度的改進。

關鍵詞：發音事件，多語語音辨識，發音變異

一、緒論

全球化趨勢來臨，文化交流，商業活動和網路資訊充斥著多語(Multilinguality)的環境及各式各樣的應用。因此，單一語言的語音辨識器，已無法滿足需求，多語自動語音辨識顯得相形重要。此外，在全球化的影響之下，學習母語以外的語言已經成爲全球的趨勢，當下尤以英語爲全球民眾爭相學習的第二語言，而民眾在口說非母語之語言時往往會受自身母語影響，挾帶母語之腔調，產生發音變異。在台灣，「中式英文」即爲語調產生發音變異之實例。就人機互動而言，如何去做到互動的親切與便利是非常重要的，倘若因爲發音變異的影響導致語音辨識器不斷的出現錯誤，原本期望其所帶來的便利性反而會因辨識率的降低成爲人機互動的阻礙。

在多語語音辨識中可能面臨的其中一個問題，即非本國(Non-native)人所造成的發音變異嚴重影響多語語音辨識的效果，因此本論文主旨即在於多語語音辨識之發音變異(Pronunciation Variation, PV)的研究，從而改善因非本國人口說所造成之發音變異對於語音辨識器的影響。因此，希望可以建立一個中英多語辨識器，此辨識器可以正確辨識標準發音，亦能夠提升對於發音變異語句之辨識率。所以，在本論文方法中藉由擴增發音變異聲學模型的方式，並透過較小的單位即聚類狀態(Senone)，建立以 senone 爲基礎的音素聲學模型，且希望利用 senone 能較準確的描述模擬發音變異，產生帶有發音變異資訊之英文聲學模型，將其加入標準發音聲學模型，改善中英多語語音辨識器受英語發音變異造成正確率下降的影響。此外，亦希望可使標準發音語句之辨識能夠不受影響。

多語音素模型之建立可歸納爲三種方式。首先，我們可以直接合併個別單一語言

Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING 2010), Pages 50-64, Puli, Nantou,Taiwan, September 2010.

之音素集，建立多語音素模型，但是這種方法沒有考慮多語音素間參數分享的特性。

第二，藉由對照國際音素標準定義，考慮個別單一語言之音素，達到多語音素間參數共用的特性，但是此作法上缺乏資料統計分析，而是專家知識決定各音素定義。國際音素標準定義包含有：International Phonetic Alphabet (IPA)[5]、Speech Assessment Methods Phonetic Alphabet (SAMPA)[6]和 Worldbet[7]等。此種方式是基於專家知識，將個別獨立的單一語言對應到標準的符號定義，藉此各語言間可分享相同的音素定義。此作法可有效地將部分的中英文音素合併，共享語言間彼此的共同音素，減少語音音素模型的定義和訓練，且適用於任何語言(Language Independent)。但此作法缺點是建構在專家知識分析，而非從資料特性統計的角度定義。也就是說，直接對照標準定義產生的多語音素集，並沒考慮到音素模型間頻譜特性。

第三，估計多語語音音素間相似程度，除了利用 IPA 國際標準定義的多語音素，過去研究也曾利用建立混淆矩陣(Confusion Matrix)以估測音素模型間的相似度，以 HMM 模型參數距離計算，利用遞迴方法合併音素模型，建構出多語辨識的音素集。多語音素間相似度的量測，可以利用 Bhattacharyya distance[8]或者是 Kullback-Leibler (KL) divergence[9]的方法，計算多語音素模型間的距離，決定相似度以定義多語音素集。此作法上，同時考慮多語音素間參數分享的特性，並利用資料統計分析決定音素定義。但缺點在於計算模型參數間的距離，與實際辨識演算法執行時，所考慮的聲學相似度(acoustic likelihood)不符。

近年來，一些藉由解決發音變異現象來改善發音變異造成辨識率降低的研究漸漸被提出來。而在找尋發音變異規則上，大致有三種方式，第一種爲知識爲基礎的方式(Knowledge-based)，即透過專家定義，找出發音變異的規則[14][15][10][11]，但其缺點是發音變異的情況太多，無法將所有發音變異規則都觀察到。

第二種方法利用資料分析(Data-driven)，[16][12][17]使用標準辨識器辨識變異語料，建立混淆矩陣找出發音變異規則。[20]利用本國(Native)語言辨識器辨識標準的非本國語言，以得到本國語言與非本國語言之間發音的對應，再利用決策樹(Decision Tree)加以歸類，得到發音變異規則。[21]透過辨識器，以狀態爲單位，計算混淆機率(Confusion Probability)以得到發音變異程度。

第三種方式是結合知識爲基礎及資料分析兩種方式[22][13]。目前對於找出發音變異規則所使用的方式大部分是基於辨識器去偵測發音變異音段，但其缺點是無法保證辨識器是百分之百正確，因此有可能因爲辨識器的錯誤而造成發音變異的錯誤偵測。

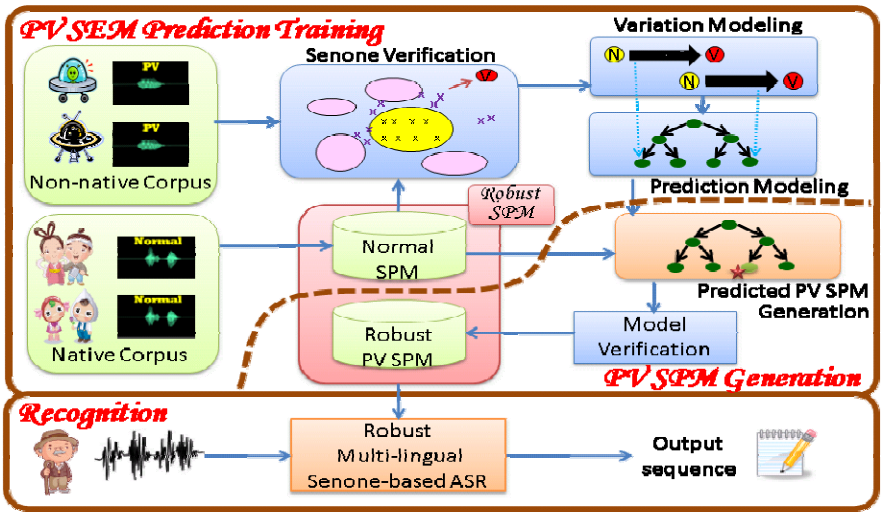
近年來，已有許多針對發音變異影響多語語音辨識效果的研究。這些研究希望在不影響原有標準的模型下，去模擬發音變異，讓辨識器除了能夠正確辨識標準語句外也能提升對發音變異之辨識正確率，因此藉由擴增發音變異模型來描述模擬發音變異。

而藉由擴增發音變異聲學模型來改善發音變異造成多語語音辨識率下降問題的方式中，若以音素模型爲單位建立發音變異模型並無法準確的描述發音變異現象，其原因在於發音變異情況大致可分兩種：完整改變之發音變異(Completely Changed Variation)及部分改變之發音變異(Partially Changed Variation)。例如：英文「Thank」的標準讀音爲/θæŋk/，假設誤念爲[θ'æŋk]，/s/亦爲一標準音素，若變異讀音[θ']之變異情況是介於標準讀音/θ/與/s/之間，則稱此種變異爲部分改變之發音變異；而若變異讀音[θ']完全誤發爲/s/，則稱此種變異爲完整改變之發音變異，對於部分改變之發音變異，以音素爲基礎之聲學模型並不適合用於模擬發音變異情況，因其無法細緻的去模擬較細微的發音變化，另外，完整改變之發音變異現象又可視爲部分改變之發音變異的一種特例。爲了能較準確的模擬細微發音變異現象，因此提出以更小單元爲基礎模擬發音變異。此較小單位即 SEM，因具發音變異的聲學模型可視爲由多個變異的 SEM 所組成，因此希望利用

SEM 可較準確的描述發音變異，進而產生以 SEM 爲基礎之具發音變異資訊的聲學模型，即發音變異 SPM。

爲解決發音變異影響多語語音辨識效果問題，首先需要找出具發音變異現象的資料，才能進一步模擬發音變異。近年來，大部分做法是基於辨識器並考慮聲學特性，來偵測找出發生發音變異的語音段，但此做法往往會因爲辨識器本身準確度不夠，而對發音變異語音段造成錯誤偵測，無法準確描述模擬發音變異。爲解決此問題，除了使用辨識器考慮聲學特性偵測發音變異現象，更進一步利用發音事件(Pronunciation Event, PE)進行發音變異驗證，希望藉此改善因辨識器造成的錯誤偵測，能較準確的驗證出發音變異語音段。

由於上述問題，論文中提出解決發音變異的方法是基於聚類狀態(Senone)模型去描述發音變異現象，因此本論文使用的兩個主要單元爲聚類後聲學狀態模型(Senone Model SEM)及基於聚類後聲學狀態之聲學模型(Senone-based Phone Model, SPM)。本論文使用的方式是將聚類後的聲學狀態模型(Senone Model SEM)做爲 HMM 聲學模型的基本單位[18][19]。決策樹可用來實現高效三連音素模型(Tri-phone)對 senone 的對應，通過回答一系列前後音所屬類別（元/輔音、清/濁音等等）的問題，以確定其 HMM 狀態應對應至哪個 senone。



圖一、系統示意圖

本研究希望從具發音變異的語料中，藉由考慮聲學以及發音事件於 SEM 層次上驗證，找出實際上具發音變異的語音段，並且利用線性發音變異預測轉換函式產生發音變異的 SPM。另外藉由語音的發音事件參數將發音變異做分類，利用決策樹歸納出不同發音方式下的變異特性，藉以預測產生訓練語料以外的發音變異 SPM。

圖一爲系統示意圖，利用標準語料先訓練一組以 SEM 爲基礎之標準音素聲學模型，再透過 SEM 對具發音變異的語料進行驗證。驗證後找出可能爲發音變異的語音段並且訓練爲發音變異 SEM，接著對成對的標準及發音變異 SEM 利用線性關係訓練發音變異預測轉換函式，並配合語言特徵參數使用決策樹來做分類，最後可藉由決策樹預測發音變異轉換函式，並產生訓練語料以外的發音變異 SPM，之後爲了希望預測產生的發音變異模型能夠具有足夠鑑別力，再進一步進行預測模型之驗證，保留富強健性的發音變異預測 SPM。而在辨識的部分，使用標準的 SPM 與預測產生的發音變異 SPM 所整合成具強健性的 SPM(Robust SPM)進行測試得到辨識結果。

二、發音變異之 SEM 驗證

(一) 標準發音 SPM 之訓練

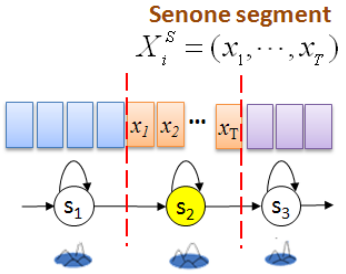
在聲學模型建立上，一般研究對於中文語言使用單一音素模型(Mono-phone)，而英文使用三連音素模型，但爲將不同語言結合建立一組中英多語辨識之聲學模型，在此初步做一個單一語言辨識的評估。對於中文及英文各自分別建立單一音素模型(M_Sub 與 E_phone)與三連音素模型(M_TRI 與 E_TRI)，初步評估結果如表一。根據結果可發現中文與英文其使用三連音素模型之辨識率較高。因此，在本研究上將使用三連音素模型來建立中文多語語音之聲學模型。而在中英語音辨識中定義的音素單元，爲結合中英文，因此在音素定義上，採用中英文對照國際音素標準定義的方式定義音素集合，其原因在於此方法能共享語言間彼此的共同音素，減少語音音素模型的定義和訓練，且適用於任何語言。接著採用標準的由上而下凝聚演算法(Bottom-Up Agglomerative Algorithm) 來建立 SEM，並產生 SPM。

表一、中英單一語言辨識評估表

Model ID	# Model	Dictionary	Inside WAR
M_Sub	151	5000	52.66%
M_TRI(50)	1491	5000	84.04%
E_phone	38	4905	59.91%
E_TRI	3635	4905	96.29%

(二) Senone 驗證

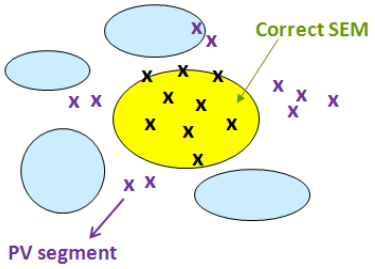
對於具發音變異的語料，當中存在標準的發音與變異的發音，而爲了描述模擬發音變異，需要偵測哪些音段可能有發音變異的現象存在，並找出發音變異的規則，因此，本論文提出透過 senone 驗證的方式，對於具發音變異的語料進行驗證並偵測發音變異現象。



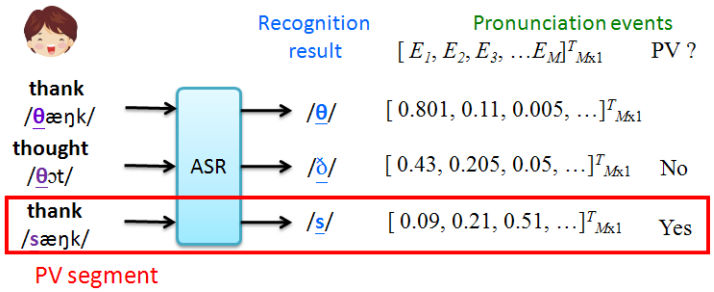
圖二、Senone 語音段示意圖

在 senone 驗證部份，主要分成兩個階段，第一階段將考慮聲學特性進行驗證：如圖二，對於一個 SEM 所對應的音框序列稱之爲 senone 語音段(Senone Segment)，一 senone 語音段若經由此階段驗證爲具發音變異，則稱爲候選之發音變異語音段(Candidate PV Segment)。第二階段則考慮發音事件進行驗證：此階段根據發音事件參數，針對前一階段驗證認爲的候選之發音變異語音段再次進行驗證以發音事件爲基礎之驗證，經過第二階段驗證認爲具有發音變異現象者，則將其視爲發音變異語音段(PV Segment)，如圖三爲 senone 驗證示意圖。本論文提出透過兩階段進行驗證方法的主要原因爲，第一階段考慮聲學特性透過辨識器進行驗證，但由於無法保證辨識器能夠具有百分之百的辨識正確率，因此，經由辨識器驗證時無法確認 senone 語音段是真的具有發音變異或者是因爲辨識器本身不夠準確而造成的錯誤偵測，如圖四所示。而造成發音變異時，其所屬的發音事件與標準發音時的有所不同，因此，在第二階段再進一步加以考慮發音事件

(Pronunciation Event, PE)透過事件偵測器(Event Detector)加以驗證是否爲發音變異語音段。若經聲學特性驗證認爲具發音變異現象，但經發音事件驗證不具發音變異，則認爲是因辨識器所造成的錯誤偵測，因此經過兩階段驗證認爲具發音變異現象之語音段，才認定爲發音變異語音段。



圖三、發音變異語音段示意圖



圖四、senone 驗證兩階段示意圖

(三) 以聲學特性爲基礎之驗證

Senone 驗證第一階段考慮聲學特性，利用標準 SEM 透過辨識器進行發音變異之偵測。在此，定義式子(1)以及式子(2)用以計算語音段之變異程度

$$G_{veri}(x) = \log g(x \mid \lambda_{correct}) - \log g_{Anti-Model}(x) \tag{1}$$

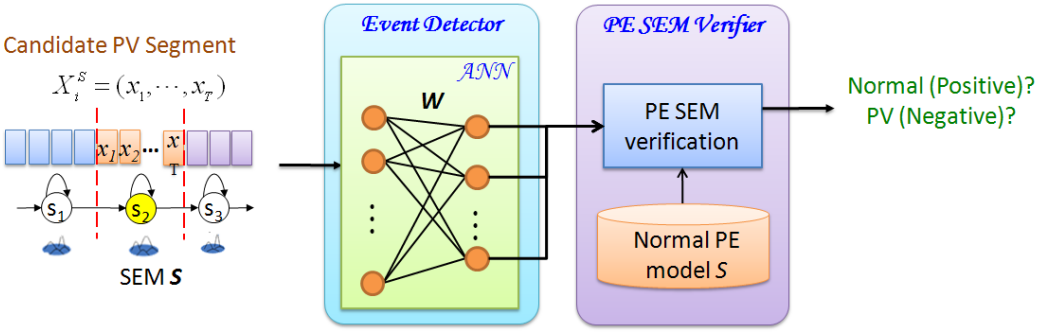
$$g_{Anti-Model}(x) = \frac{1}{N} \sum_{\substack{AAM=1, \\ \lambda_{AAM} \neq \lambda_{correct}}}^{N} g(x \mid \lambda_{AAM}) \tag{2}$$

其中，x 爲一語音段特徵參數，$g(\cdot)$ 爲辨識之記分函式，$\lambda_{correct}$ 爲正確答案之 SEM，λ_{AAM} 爲反聲學 SEM，即爲與 $\lambda_{correct}$ 較相近的 SEM，而 N 則爲反聲學 SEM 之個數。式子(1)前項即爲 x 對於 $\lambda_{correct}$ 的辨識分數，而後項爲反聲學 SEM 經辨識所得分數之平均。若一語音段爲標準發音則式子前項分數會較高，後項分數則較低，因此前後項相減之後所求得的 G_{veri} 值較高，反之則較低，所以在此設定一門檻值 $threshold_G$，若求得的驗證分數低於此門檻值，則 x 爲候選之發音變異語音段。

(四) 以發音事件爲基礎之驗證

爲降低語音辨識器所造成的錯誤偵測影響，因此本研究在 senone 驗證的第二階段再進一步考慮利用發音事件加以驗證。由於發音變異語音段所屬之發音事件不同於標準發音語音段之發音事件，因此，在此階段驗證中，透過訓練一個標準發音事件模型(Normal PE Model)來對候選發音變異語音段所偵測的發音事件進行驗證，若被標準發音事件模型所判定具發音變異現象即爲發音變異語音段。而在以發音事件爲基礎之驗證

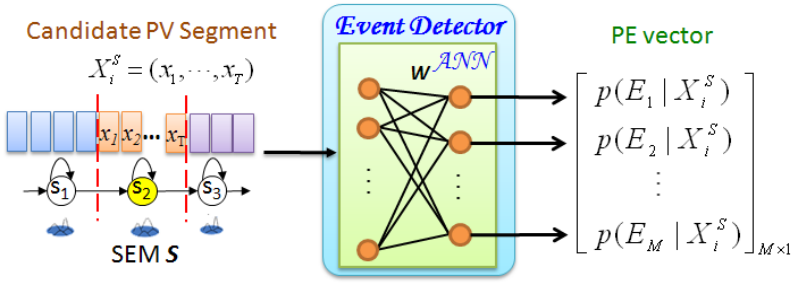
中，候選發音變異語音段可透過事件偵測器(Event Detector)偵測其發音事件並用以驗證。在此所定義的發音事件如下：Anterior, Back, Consonantal, Continuant, Coronal, High, Low, Nasal, Round, Silence, Strident, Tense, Vocalic, Voice。而候選發音變異語音段會藉由偵測出的發音事件[23]透過發音事件驗證器中的標準發音事件模型進行發音變異現象的驗證。



圖五、發音變異事件驗證器示意圖

圖五爲發音變異事件驗證器示意圖，其輸入爲經第一階段以聲學特性爲基礎之驗證所驗證出的候選發音變異語音段，此語音段會先經一事件偵測器偵測其發音事件，偵測之後可得到的輸出爲一個 $M \times 1$的發音事件向量(PE Vector)，再將此輸出做爲發音事件 SEM 驗證器(PE SEM Verifier)的輸入，當中透過標準發音事件模型進行對偵測的發音事件進行驗證，而經驗證後的輸出結果即可判斷語音段是否爲發音變異語音段。

在此用於候選發音變異語音段偵測發音事件的事件偵測器，如圖六所示，其爲以音框爲基礎之事件偵測器，若事件偵測器的輸入爲一個有 T 個音框的候選發音變異語音段，其輸出爲一個 $M \times 1$的發音事件向量(PE Vector)，此發音事件向量由 T 個音框透過事件偵測器之偵測結果的平均所得到，而發音事件向量中的每個維度爲一候選發音變異語音段針對某一發音事件的機率。



圖六、事件偵測示意圖

而針對一個音框之事件偵測是透過類神經網路(Artificial Neural Network)得到偵測結果，其輸入爲候選發音變異語音段中一個音框之參數向量 $\boldsymbol{z(t)}$，$\boldsymbol{z(t)}$ 是由此音框的特徵參數向量以及類神經網路中的目前狀態向量(Current State Vector) $\boldsymbol{u(t)}$ 所組成，透過類神經網路當中的參數 $\boldsymbol{w_E}$ 運算後，可分別得屬於某一發音事件 E 的機率。

$$p(y = E \mid X_i^S) = T^{-1} \sum_{t=1}^{T} \left(\frac{\exp\left(w_E^T z(t)\right)}{\sum_{j=1}^{M} \exp\left(w_j^T z(t)\right)} \right) \tag{3}$$

$$z(t) = \begin{bmatrix} 1 & x(t) & u(t) \end{bmatrix}^T \tag{4}$$

$$u(t+1)=\left(1+\exp\left(-v_E z(t)\right)\right)^{-1} \quad (5)$$

其中 W 與 V 分別爲針對輸出以及下一個狀態之權重矩陣(Weight Matrices)。而發音變異事件偵測器當中的 PE SEM 驗證器，當中使用二元支援向量機(Binary Support Vector Machines, Binary SVM)訓練出的標準發音事件模型(Normal PE Model)做爲驗證模型。由於基於辨識器之驗證可能造成錯誤的偵測，因此無法保證經由聲學特性驗證之後的資料一定具發音變異現象，所以無法直接拿其結果訓練標準發音事件模型。另外，屬於某一標準發音的發音特徵，其所偵測的發音事件應爲相似的，因此可依據此線索找出潛在的發音變異現象。而對於二元支援向量機來說用於訓練的正集(Positive)與負集(Negative)資料會影響其效能的好壞，所以在進行訓練時需要挑出好的訓練資料，因此在此使用迭代的方式進行標準發音事件模型的訓練，在每次迭代中選出較佳的資料訓練新的模型直到收斂。因此，在測試的部分，經由迭代訓練得到最後的 SVM 模型，可用來再次判斷經第一階段「以聲學特性爲基礎之驗證」，驗證通過的候選發音變異語音段，若候選之發音變異語音段，經事件偵測器偵測的發音事件向量在此階段驗證結果爲負(Negative)，即爲具發音變異之語音段。

三、發音變異 SEM 預測與 SPM 建立

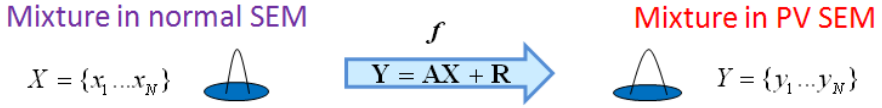
爲能夠正準確的描述模擬發音變異，因此提出在 SEM 層次上模擬發音變異，並藉由標準 SEM 以及發音變異 SEM 透過發音變異預測轉換函式，來模擬發音變異產生發音變異 SEM。且爲了對未蒐集到發音變異語料之標準聲學模型預測產生其發音變異模型，在此利用決策樹進行預測。而針對發音變異造成多語語音辨識之辨識率下降的問題，在不影響原有標準 SPM 的情況下，產生具發音變異的 SPM 方法來進行改善，並驗證其是否適合留下一同做爲多語語音辨識器之聲學模型。在此本章將介紹如何利建立轉換函式預測模型，且藉此產生預測具發音變異之 SPM 並對其做模型驗證。

(一) 發音變異 SEM 預測模型之建立

針對產生發音變異之 SEM，經由以高斯分佈爲基礎的階層式轉換函式架構可以找出成對的高斯分佈，接著採用線性的假設關係，將發生變異的高斯分佈視爲標準高斯分佈的線性組合和轉換，將成對的高斯分佈利用線性轉換的方式來描述標準與發音變異高斯混合的關係，我們定義標準高斯分佈（normal, $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2\ldots,\boldsymbol{x}_N)$）

$$\mathbf{Y}=\boldsymbol{AX}+\boldsymbol{R} \quad (6)$$

藉由式子(6)的線性轉換函式來轉換成爲目標的發音變異高斯分佈（PV, $\boldsymbol{Y}=(\boldsymbol{y}_1,\boldsymbol{y}_2\ldots,\boldsymbol{y}_N)$）。其轉換關係如下圖所示。



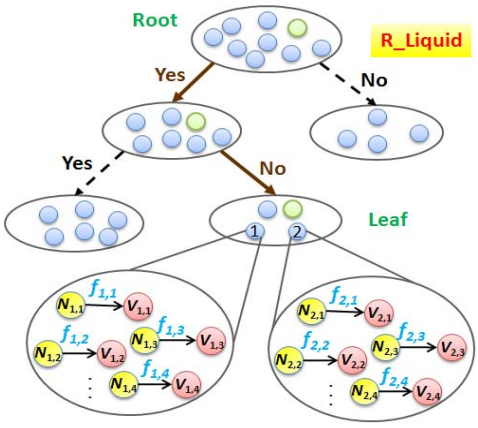
圖七、線性轉換關係示意圖

如圖七，利用正常語音資料 X 透過旋轉矩陣 A 的轉換後，將旋轉的誤差用 R 表示。

(二) 發音變異 SEM 預測模型

我們選用決策樹將發音變異特性做分類，同一類別中的資料點帶有相同的發音特性變化。使用決策樹做爲預測模型的優點在於，以樹狀的結構來表現資料的分佈，其建立出來的模型容易瞭解，並且能追蹤每節點上使用的變數進而瞭解資料真正的特性。根

據圖八所示之轉換函式預測模型示意圖，首先發音變異轉換的 SEM 模型，透過轉換函式之決策樹(Transformation Function Decision Tree)，將轉換函式依照上一節所提到的發音事件參數來做決策樹的建置和分類。最後分類得到的每一個樹葉節點，會保留所有屬於此葉節點的 SEM 其所有高斯分佈之轉換函式，並用當中的轉換函式來預測標準 SEM 與發音變異 SEM 間的變化與差異。



圖八、轉換函式預測模型示意圖

標準的 SEM 經由轉換後變成發音變異的 SEM，利用決策樹來做預測時，希望預測的結果與目標的 SEM 在聲學上的差異越小越好。也就是來源標準的 SEM 經過分類後，根據所在類別選取的轉換函式來進行轉換，其轉換後之結果(Converted)與變異的目標 SEM 越相似越好。本研究模型採用決策樹，在分裂的條件上設定爲“分裂後的轉換誤差(Generation Error)小於分裂前轉換誤差”，轉換的誤差計算方式定義如下：

$$GenErr_i = \sum_{m=1}^{M}\sum_{t=1}^{T}\left\|\mathbf{y}_m - (\mathbf{A}_{m,t}\mathbf{x}_m + \mathbf{R}_{m,t})\right\|^2 \qquad (7)$$

其中 y_m 爲存在於子節點中的第 m 個目標高斯分佈(Target Mixture)，x_m 爲對應於目標高斯分佈之標準高斯分佈(Normal Mixture)，$\mathbf{A}_{m,t}\mathbf{x}_m + \mathbf{R}$ 爲此高斯分佈鄰近的第 t 個線性轉換函式，M爲於此子節點中高斯分佈的總數。在此同時考慮鄰近的 t 個線性轉換函式，原因在於希望被分於同一群的轉換函式差異不會太大。欲得到最佳的分裂的結果，亦即欲最大化減少的轉換誤差量，減少的誤差量計算方式爲母節點的轉換誤差扣除分裂後子節點轉換誤差。其定義如式子(8)。

$$\text{Generation Error Reduction (GER)} = GenErr_p - \sum_i \frac{M_i}{M_p} GenErr_i \qquad (8)$$

其中 $GenErr_p$ 爲母節點的轉換誤差，$GenErr_i$ 爲第 i 個子節點的轉換誤差。在轉換函式預測模型示意圖中，每個資料點爲一個 SEM 中的高斯分佈，每個資料點參數包含了訓練語料中頻譜參數與發音參數 A，其中頻譜參數包含轉換函式中的來源頻譜參數 X 與目標頻譜參數 Y，皆爲 39 維的 MFCC 及能量參數，所有資料點從根節點(Root)出發，分裂的時候由節點上的問題，決定資料點要被分到左子節點或右子節點。以圖八中綠色的資料點爲例，考慮完分裂的條件後，在 Root 分裂時的問題是[R_Liquid]，與該資料點發音參數吻合，答案爲 Yes，故分到左子節點(若答案爲 No，則分到右子節點)，最後在葉節點中記錄所有在葉節點中的 X 與 Y SEM 中每個對應高斯分佈得到轉換函式的參數。

(三) 預測發音變異 SPM 之轉換函式挑選

發音變異 SPM 是由多個具變異的 SEM 所組成，因此透過訓練的轉換函式預測模型，可對標準 SPM 中的每個 SEM 預測產生具發音變異的 SEM，進而預測出具發音變異的 SPM。在訓練出轉換函式預測模型後，在對於所有標準 SPM 中的每個 SEM 進行挑選動作時，先抽取出嘴巴發音參數，並從決策樹中去對模型參數進行分類，來建立決策樹分類之預測模型。使用發音參數語文字上的資訊做爲問題集，所取出的參數從預測模型的 root 節點開始進行問題的比對，符合該節點的問題就往左邊節點移動，不符合該節點的問題就往右邊節點移動，直到挑選到最後葉節點爲止。

而在訓練轉換函式預測模型時，每個樹葉節點中記錄了所有在此葉節點成對的 SEM X 與 Y 中每個對應高斯分佈得到之轉換函式參數。因此在預測時，當 SEM 挑選到最後葉節點後，即可針對其當中每一個高斯分佈，在此葉節點中挑選一個最相近的高斯分佈之轉換函式。因此，標準 SPM 中每個 SEM 都可藉由其每一個高斯分佈的轉換，產生一個發音變異 SEM。

若考慮所有產生預測發音變異 SPM 之排列組合，即有八種可能。因此，一個標準 SPM 即可產生出八個預測發音變異 SPM(Predicted PV SPM)。至於，八個預測發音變異的 SPM 是不是皆適合做爲描述發音變異的聲學模型，則需要再做進一步的模型驗證。

(四) 發音變異 SPM 之模型驗證

對於預測產生的發音變異 SPM，必須去驗證其是否適合做爲描述發音變異的聲學模型，且希望經由驗證後，認爲適合留下的發音變異 SPM，在原有標準 SPM 以及具發音變異 SPM 中是具有足夠鑑別力(Discrimination)，因此，在此使用一模型驗證方法，做爲發音變異 SPM 是否留下的評估方式。

本論文將模型驗證一共分爲兩個階段，原因爲由於預測的發音變異 SPM 還未能確定是否有足夠的鑑別力適合保留做爲辨識的模型，因此在第一階段僅使用標準 SPM 對所有預測之發音變異 SPM 進行驗證，亦是希望能夠先驗證預測產生之發音變異 SPM 不會對標準 SPM 造成混淆，在此定義了鑑別函式(Discrimination Function)，如式(9)，計算鑑別程度(Discrimination Degree)，其定義如下：

$$d_i^{PV\_N} = -g(Y_i \left| \lambda_{Y_i} \right.) + \max_{m \in X} \{ g(Y_i \left| \lambda_m \right.) \} \tag{9}$$

其中 $d_i^{PV\_N}$ 爲第 i 個預測發音變異 SPM 之鑑別程度，Y_i 爲對於第 i 個預測發音變異 SPM 之發音參數，λ_{Y_i} 爲 Y_i 之 SPM，而在此式中的 λ_m 代表所有標準的 SPM。

另外，再定義以下式子，做爲一個門檻值

$$d_i^{normal\_N} = -g(Y_i \left| \lambda_{X_i} \right.) + \max_{m \in X, m \neq X_i} \{ g(Y_i \left| \lambda_m \right.) \} \tag{10}$$

其中 $d_i^{normal\_N}$ 爲對應於第 i 個預測發音變異 SPM 之標準 SPM 的鑑別程度，X_i 爲對應於第 i 個預測發音變異 SPM 之標準發音參數，而在此式中 λ_m 代表除了 λ_{X_i} 之外的所有標準 SPM。在第一階段驗證中，若 $d_i^{PV\_N} < d_i^{normal\_N}$，則代表第 i 個預測發音變異 SPM 被驗證通過，將其稱爲候選之發音變異 SPM。第二階段使用標準 SPM 以及經前一階段驗證的候選發音變異 SPM 再次進行模型驗證，希望最後留下的強健性發音變異 SPM(Robust PV SPM)，具有足夠的鑑別能力。在此階段一樣定義兩個鑑別函式，

$$d_i^{PV\_M} = -g(Y_i' \left| \lambda_{Y_i'} \right.) + \max_{m \in \{X \cup Y'\}, m \neq Y_i'} \{ g(Y_i' \left| \lambda_m \right.) \} \tag{11}$$

$$d_i^{normal\_M} = -g(Y_i' | \lambda_{X_i}) + \max_{\substack{m \in \{X \cup Y'\}, \\ m \neq Y_i', m \neq X_i}} \{g(Y_i' | \lambda_m)\} \quad (12)$$

其中 $d_i^{PV\_M}$ 爲第 i 個候選發音變異 SPM 之鑑別程度，Y_i' 爲對於第 i 個候選發音變異 SPM 之發音參數，$\lambda_{Y_i'}$ 爲 Y_i' 之 SPM，而在式子(11)中的 λ_m 代表除了 $\lambda_{Y_i'}$ 之外的所有標準 SPM 與候選發音變異 SPM。$d_i^{normal\_M}$ 爲對應於第 i 個候選發音變異 SPM 之標準 SPM 的鑑別程度，X_i 爲對應於第 i 個候選發音變異 SPM 之標準發音參數，而在式子(12)中 λ_m 代表除了 λ_{X_i} 與 $\lambda_{Y_i'}$ 之外的所有標準 SPM。在第二階段驗證中，若 $d_i^{PV\_M} < d_i^{normal\_M}$ ，則代表第 i 個選發音變異 SPM 被驗證通過，將其稱爲強健性發音變異 SPM，而強健性發音變異 SPM 則會一同做爲多語語音辨識器之聲學模型。

四、實驗結果與分析

(一) 實驗語料

本研究中訓練中英多語語音辨識聲學模型之語料，中文爲 TCC300 語料庫，其中男女生語料句數各爲 2500 句。英文則爲 TIMIT 語料庫，語料句數共 4300 句，男女生語料句數分別爲 3020 以及 1280 句。因希望本論文系統適用於台灣人，因此另外準備一英文語料庫：EAT(English Across Taiwan)語料庫，其爲台灣口音英語語料庫，當中又分爲主修英文之台灣人口說英語語料(Major)以及非主修英文之台灣人口說英語語料(Non-major)，在實驗中將 EAT major 語料視爲標準英語發音的語料，並用以調適由 TCC300 以及 TIMIT 語料所訓練的中英文多語語音辨識聲學模型，使此聲學模型較能描述台灣人口說英文的特性，其句數共 327 句，男女生句數分別爲 89 以及 188 句。而 EAT non-major 語料則視爲具發音變異的語料，使用於發音變異現象驗證以及線性轉換函式之訓練，句數共 382 句，男女生句數分別爲 193 以及 139 句。測試語料則針對中文標準發音語料選取 100 句做測試，其中男女生語料各 50 句，英文標準發音語料以及發音變異語料由於語料較少，因此皆選取 50 句做測試。在此，選取的測試語料皆不包含在訓練語料內。

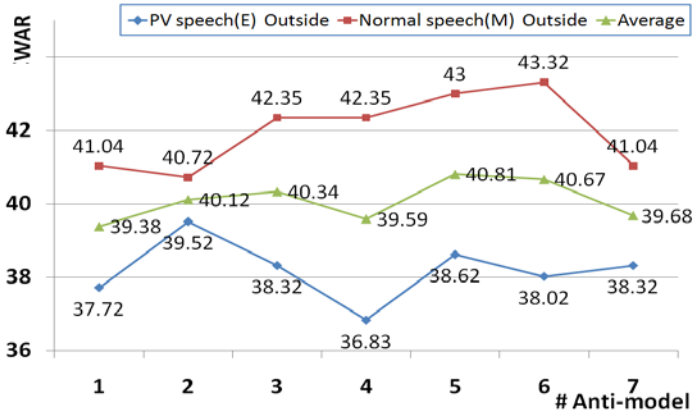
(二) 實驗環境

本論文目標爲建立一能改善發音變異造成辨識正確率下降之中英多語自動語音辨識器，本研究所使用基於隱藏式馬可夫模型之語音辨識系統 HTK(Hidden Markov Model Toolkit)[3] 做爲實驗中的模型訓練與辨識元件。語音特徵參數爲梅爾倒頻譜係數(Mel-Frequency Cepstrum Coefficient, MFCC)，其維度爲 39 維。並且根據表一採用三連音素模型來建立中英多語語音辨識之聲學模型，包含三個狀態單元(State)，分別爲 16 個高斯分佈(Mixture)。而基礎的中英多語語音辨識上，定義了 75 個音素單元，其中包含 74 個音素單元以及靜音(silence)。

根據上述實驗設定，使用蒐集的標準發音語料訓練，並根據 SPM 訓練流程訓練基礎中英多語語音辨識聲學模型，共訓練出 5489 個 SPM 以及 3075 個 SEM。且爲使所建立的中英文語音辨識器能適用於台灣人口說之中英文語句，因此基礎中英多語語音辨識器會再經過 EAT major 語料進行調適後，而基礎系統分別對中文 TCC300、英文 TIMIT 以及英文 EAT major 標準發音之測試語料的詞辨識率(Word Accuracy Rate, WAR)分別約爲 76.33%、91.69%以及 56.52%。辨識率的差異來自於不同語料庫之錄音品質與口說之清晰度有所不同所導致。

(三) 實驗與評估

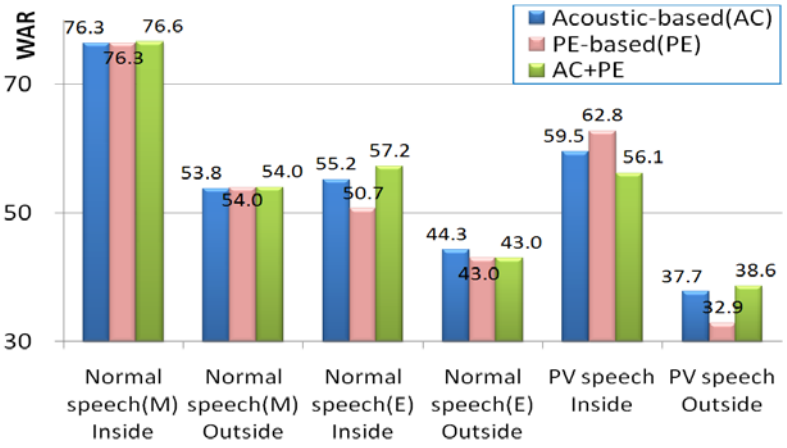
在驗證發音變異資料中，本小節討論使用不同的驗證發音變異資料方式以及比較在發音變異驗證步驟中所使用不同的設定參數對辨識率造成的影響。本論文提出的驗證發音變異方法爲在 SEM 上進行兩階段驗證：分別爲以聲學特性與發音事件爲基礎之驗證。實驗中，以聲學特性爲基礎之驗證(Acoustic- Based Verification)，簡稱爲 AC；而將以發音事件爲基礎之驗證(PE-Based Verification)簡稱爲 PE；而提出使用兩階段驗證的方式則簡稱 AC+PE。



圖九、驗證發音變異資料方法與反聲學 SEM 個數之評估結果

在此，先對兩階段發音變異驗證方法中，以聲學特性爲基礎之驗證的參數即反聲學 SEM 之個數進行實驗。在這邊挑選實驗的反聲學 SEM 個數爲 1 到 7。

如圖九，在此比較不同的反聲學 SEM 個數對於系統辨識率的辨識結果，因爲希望能夠對於發音變異辨識有較好的結果，但亦希望在標準語音辨識以及發音變異辨識中取得一個平衡，因此在此使用對於標準語句與發音變異語句之外部測試(Outside Test)結果以及兩者平均結果做評估。由兩者平均的結果可發現當反聲學 SEM 個數爲 5 時有較好的結果，而其標準語句與發音變異語句辨識之詞正確率分別爲 43%以及 38.62%。雖然對於發音變異語句之辨識結果顯示，反聲學 SEM 個數爲 2 時有最高的詞正確率，但其標準語句之辨識結果卻爲當中之最低。爲了對標準語句以及發音變異語句能都夠有不錯的辨識結果，因此在接下來的實驗，在以聲學特性爲基礎之發音變異驗證步驟中，反聲學 SEM 個數皆會設定爲 5，使用以此設定的中英文多語語音辨識器進行實驗比較。



圖十、不同的發音變異驗證方式之評估結果

另外，本論文提出使用兩階段進行發音變異資料驗證，因此，將比較僅使用其中

一種與使用兩階段驗證之間的差異，並證明提出使用兩階段進行發音變異資料驗證具有較好的效果。實驗中，分別使用 AC、PE 及 AC+PE 三種驗證方式進行發音變異資料的驗證，並建立富強健性的多語語音辨識系統，對標準語料與發音變異語料進行辨識，在此的標準語料，包含了標準的中文語料與英文標準語料，而在此亦會對內部測試(Inside Test)以及外部測試(Outside Test)進行比較。

實驗結果如圖十，使用兩階段的其中一種驗證方式與兩階段發音變異驗證，對於中文標準語料的 inside 以及 outside 測試之辨識率差異不大。而對於英文標準語料的 inside 測試，以 AC+PE 的驗證方式建立的辨識系統有最高的詞正確率約爲 57.2%，比單以 AC 或者 PE 的驗證方式高出分別約 2%與 6.5%之詞正確率，在 outside 測試的部分，以三種方式所得的辨識結果並無太大差異，而其中 AC+PE 之所以效果較使用單一 AC 來得差，推論是因爲用來訓練 PE 模型之語料並不算充足，使 PE 在此之表現反倒降低了 AC 所帶給系統之效能。因此在對於標準語料的測試結果中，發現以 AC+PE 的發音變異驗證方式有比較好的辨識效果。

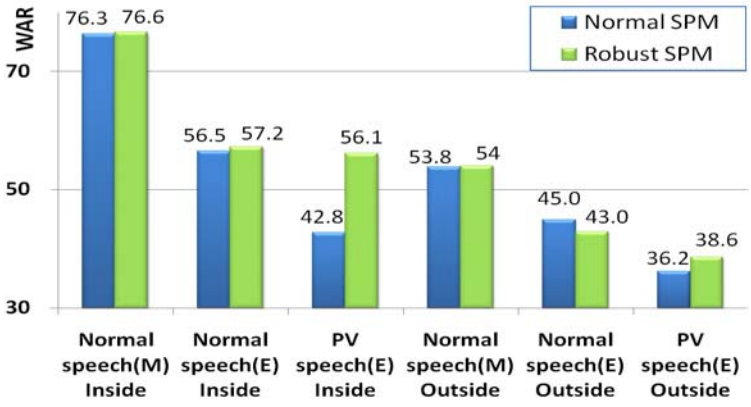
而對於發音變異語料的測試結果，單以 AC 或者 PE 的方式驗證所得的 inside 辨識結果都較 AC+PE 的 56.1%好，分別高出約 3.4%與 6.7%，其中以 PE 的方式爲最高，但其 outside 測試結果低於 AC+PE 驗證方式的 38.6%詞正確率約有 0.9%以及 5.7%，且以 PE 方式爲最低。對於發音變異語料的測試結果，比較 inside 以及 outside 測試，單以 AC 或 PE 的方式驗證，可能過度依賴其考慮的特徵參數驗證，因此對於 inside 測試的結果都較高，而對於 outside 的資料則無法有較好的辨識率，猜測此應爲過度訓練(Overtraining)所造成的結果。

因此，從實驗結果可以得知，使用本論文提出的 AC+PE 的發音變異資料驗證的方法，能夠在標準與變異語料其 inside 與 outside 的測試中，能取得一個較爲平衡的結果。

(四) 發音變異 SPM 建立之評估

在本論文中，提出改善因發音變異影響多語語音辨識正確率之方法，爲產生發音變異 SPM。因此，在此討論所提出的方法是否對上述問題有所改善。在本小節中，進行兩個多語語音辨識聲學模型的比較：分別爲標準多語語音聲學模型(Normal SPM)以及本論文提出之強健性多語語音聲學模型(Robust SPM)。

上述的兩個多語語音辨識聲學模型，會分別對標準發音語料及具發音變異之語料進行測試，希望可以證明本論文所提出加入發音變異 SPM 之方法，對於標準語音亦能夠有不錯的效果，且同時能改善因發音變異造成多語語音辨識器正確率下降的影響。



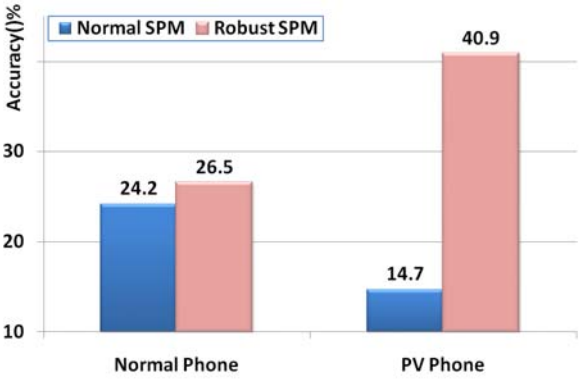
圖十一、系統評估結果

實驗結果如圖十一，對於中文標準語料之 inside 測試結果，Normal SPM 與 Robust

SPM 兩者無明顯差異，其詞正確率皆爲 76.3%，對於英文標準語料的測試，robust SPM 之詞正確率爲 56%，較 Normal SPM 之詞正確率低約 0.5%，而對於發音變異語料之辨識結果，Robust SPM 之詞正確率約爲 56.1%，且明顯高於 Normal SPM 的 42.8%詞正確率。而在 outside 測試亦可看到相同的情況，其中 robust SPM 對中文標準語料、英文標準語料以及發音變異語料的測試結果分別爲 54.0%、43.0%以及 38.6%。其中 Robust SPM 對於英文標準發音之辨識結果較 Normal ASR 的些微降低，推測其原因應爲由於 Robust SPM 是擴增了發音變異的 SPM，以改善因發音變異降低辨識系統正確率的情形，但對於原有的標準英文 SPM 可能造成混淆，因而降低了對英文標準語料之正確率。但針對英文發音變異語料之結果，可證實使用 Robust SPM 能提升對發音變異語句之辨識正確率，所以對於些微降低的標準發音語料之詞正確率，屬可接受範圍，而此實驗結果亦達到原有期望，即在不影響標準發音語句之正確率太多的情況下，改善對於發音變異語句之正確率。

接下來對 Normal SPM 與 Robust SPM 做交叉驗證(Cross Validation)評估，每次分別從發音變異語料中取出約 50 句語料做爲測試語料，其餘的則做爲訓練語料，且每次挑選的測試語料皆不重複，因此針對所蒐集到的發音變異語料共可分爲九組的測試資料，將九組的測試資料分別進行測試後，將其結果平均即可得到交叉驗證的結果。

在此實驗中，我們用 Normal SPM 和 Robust SPM 對發音變異語料計算音素的辨識正確率。針對發音變異語料對標準音素以及變異音素進行標記，再分別使用 Normal SPM 與 Robust SPM 進行辨識，此實驗欲觀察加入發音變異模型之後的 Robust SPM 對標準音素以及發音變異音素的影響。



圖十二、發音變異聲學模型之評估結果

實驗結果如圖十二，對標準音素來說，使用 Robust SPM 以及 Normal SPM 之音素正確率並無太大差異，約爲 24.2%至 26.5%，但對發音變異音素，Robust SPM 之音素正確率爲 40.9%，比使用 Normal SPM 提升了約 26.2%之辨識率，由此可見，使用 Robust SPM 可提升對發音變異音素的辨識正確率外，仍可使得對於標準音素的正確率有些微的提升。

五、結論

本研究建立發音變異模型以提升對於發音變異語句之辨識率。在驗證發音變異現象上，同時考慮聲學特性以及發音事件，利用 SEM 進行發音變異資料的驗證，且爲了較精準的模擬發音變異現象，提出於 SEM 模擬發音變異，並藉由擴增以 senone 爲基礎的英文發音變異聲學模型，改善因發音變異造成的辨識系統正確率降低的問題。另外，對於未能收集到的變異語料問題，利用決策樹並考慮語言的特徵參數進行分類，預測產

生以 senone 爲基礎之發音變異模型。

由實驗結果可以得知，在對於發音變異現象的描述並產生發音變異聲學模型，相較於利用音素建立發音變異聲學模型的方式，本論文提出以 senone 爲基礎透過線性轉換後建立的發音變異聲學模型，使辨識之詞正確率在 EAT non-major 語料的 inside 以及 outside 測試中分別提升了約 13.3%(圖十一之 PV speech(E) inside)與 2.4%(圖十一之 PV speech(E) outside)。另外，在驗證發音變異資料上，相較於只考慮聲學特性或者是發音事件的方式，本論文所提出的兩階段發音變異資料驗證，同時考慮聲學特性以及發音事件的方法，其所建立的 robust SPM 於 outside 測試中分別提升了約有 0.9%以及 5.7%(圖十之 PV speech outside)的詞正確率。而對於預測發音變異聲學模型，由本論文所提出建立利用決策樹考慮語言特性進行分類，預測產生的 robust SPM 方法之 outside 測試辨識詞正確率爲 38.6%，較 normal SPM 提升約 2.4%之詞正確率。

本研究著重於開發多語語音辨識器用於辨識中文或英文之語句，倘若可建立中英混合語言之語言模型，並與本系統進行整合，未來可再將此整合之系統應用於中英混合語言之辨識上，拓展應用之空間。

參考文獻

[1] A.-P. Breen and P. Jackson, “Non-Uniform Unit Selection and the Similarity Metric within BT’s Laureate TTS System,” The Third ESCA/COCOSDA Workshop on Speech Synthesis, pp. 201-206, 1998.

[2] P. Rubin, T. Baer, and P. Mermelstein, “An Articulatory Synthesizer for Perceptual Research, “ Journal of the Acoustical Society of America, Vol. 70, pp. 321-328, 1981.

[3] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X.-Y. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The Hidden Markov Model Toolkit (HTK) Version 3.4, 2006. http://htk.eng.cam.ac.uk/

[4] A.-P. Dempster, N.-M. Laird, and D.-B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1, pp. 1–38, 1997.

[5] International Phonetic Association (IPA), “Handbook of the International Phonetic Association : A Guide to the Use of the International Phonetic Alphabet,” Cambridge University Press, 1999.

[6] J.-C. Wells, “Computer-Coded Phonemic Notation of Individual Languages of the European Community,” Journal of the International Phonetic Association, Vol. 19, pp. 31-54, 1989.

[7] J.-L. Hieronymus, “ASCII Phonetic Symbols for the World's Languages: Worldbet,” Journal of the International Phonetic Association, 1993.

[8] B. Mak and E. Barnard, “Phone clustering using the Bhattacharyya distance, ” in Proc. of the International Conference on Spoken Language Processing (ICSLP), pp. 2005-2008, 1996.

[9] J J. Goldberger and H. Aronowitz, “A Distance Measure Between GMMs Based on the Unsented Transform and its Application to Speaker Recognition,” in Proc. of EUROSPEECH, pp. 1985-1988, 2005.

[10] A. Kipp, M.-B. Wesenick, F. Schiel, “Pronunciation Modeling Applied to Automatic.

Segmentation of Spontaneous Speech," in Proc. of Eurospeech, pp. 1023–1026, 1997.

[11] S. Stefan, "Generating Non-Native Pronunciation Lexicons by Phonological Rules," in Proc. of International Conference of Phonetic Sciences (ICPhS), pp. 2545-2548, 2003.

[12] G. Bouselmi, D. Fohr, and I. Illina, "Combined Acoustic and Pronunciation Modelling for Non-Native Speech Recognition", in Proc. of Interspeech, pp. 1449-1452, 2007.

[13] N. Cremelie, J.-P. Martens, "Automatic Rule-Based Generation of Word Pronunciation Networks," in Proc. of Eurospeech, pp 2459-2462, 1997.

[14] S. Downey and R. Wiseman, "Dynamic and Static Improvements to Lexical Basefoms," ESCA Workshop on Modeling Pronunciation Variation, pp. 157-162, 1998.

[15] A. Kipp, M.-B. Wesenick, and F. Schiel, "Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora," in Proc. of the International Conference on Spoken Language Processing (ICSLP), pp. 106-109, 1996.

[16] Y.-R. Oh, J.-S. Yoon, and H.-K Kim "Acoustic Model Adaptation based on Pronunciation Variability Analysis for Non-Native Speech Recognition," in Proc. of ICASSP, pp. 137-140, 2006.

[17] G. Bouselmi, D. Fohr, I. Illina, "Multi-Accent and Accent-Independent Non-Native Speech Recognition," in Proc. of Interspeech, 2008

[18] M.-Y. Hwang, and X. Huang, "Subphonetic Modeling with Markov States --- Senone," in Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 33-36, 1992.

[19] M.-Y. Hwang, X. Huang, and F. Alleva, "Predicting Unseen Triphones with Senones," IEEE Transactions on Speech and Audio Processing, Vol. 4, No. 6, pp. 412-419, 1996.

[20] S. Goronzy, K. Eisele, "Automatic Pronunciation Modeling for Multiple Non-Native Accents," in Proc. of Automatic Speech Recognition and Understanding (ASRU), pp. 123-128, 2003.

[21] Q. Zhang, T. Li, J. Pan, and Y. Yan, "Nonnative Speech Recognition Based on State-Level Bilingual Model Modification," in Proc. of Third International Conference on Convergence and Hybrid Information Technology (ICCIT), Vol. 2, pp.1220-1225, 2008.

[22] J. Yang, P. Wu, D. Xu, "Mandarin Speech Recognition for Nonnative Speakers Based on Pronunciation Dictionary Adaptation", in Proc. of International Symposium on Chinese Spoken Language Processing (ISCSLP), pp.1-4, 2008.

[23] S.-M. Siniscalchi, T. Svendsen, and C.-H. Lee,"Toward A Detector-Based Universal Phone Recognizer," In Proc. of ICASSP, . 4261-4264, 2008.

[24] C.-H Lee, C.-H. Wu, and J.-C. Guo "Pronunciation Variation Generation for Spontaneous Speech Synthesis Using State-Based Voice Transformation," in Proc. of ICASSP, pp. 15-19, 2010.