以狄式分佈為基礎之多語聲學模型拆分及合併

Multilingual Acoustic Model Splitting and Merging by Latent Dirichlet Allocation

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 Chiayi University
Ralph@mail.ncyu.edu.tw

李勝豐　Sheng-Feng Li
國立嘉義大學資訊工程學系
Department of Computer Science and Information Engineering
National Chiayi University
s1010431@mail.ncyu.edu.tw

許希聖　Shi-Sheng Shiu
國立嘉義大學資訊工程學系
Department of Computer Science and Information Engineering
National Chiayi University
s0990431@mail.ncyu.edu.tw

摘要

在整合型多語辨識環境下，如何避免不同語言間的音標混淆是重要的課題之一。本篇論文針對相異語言間的聲學模型混淆，提出以狄式分佈(LDA, Latent Dirichlet Allocation)為基礎的聲學模型混淆度偵測。以三連音素聲學模型為基礎將聲學模型分裂後再使用潛藏狄式分佈選擇合併的聲學模型組後並進行合併，以解決因為不同語言發音變異所產生的聲學模型混淆度。本篇論文分為三個部分，第一部分為介紹發音屬性和語音事件，其為從訊號面尋找各種特徵並與特定聲學模型之間的相關性。第二部分為介紹狄式分佈(LDA, Latent Dirichlet Allocation)以及模型間混淆度的偵測方法，狄式分佈是一個階層式的數學模型，早期是由 David M. Blei 等人提出用來做為文件分類及文件產生所使用，但其架構相當適合應用在語音辨識、自然語言處理等領域。最後部分則是對本論文所提出之方法進行實驗驗證並分析。

關鍵詞：多語辨識，狄式分佈，模型拆分，模型合併

Abstract

To avoid the confusion of phonetic acoustic models between different languages is one of the most challenges in multilingual speech recognition. We proposed the method based on Latent Dirichlet Allocation to avoid the confusion of phonetic acoustic models between different languages. We split phonetic acoustic models based on tri-phone. And merging the group that selected by Latent Dirichlet Allocation Detector to solve pronunciation variants problems between different languages. This paper has three parts. First part is introduced the Pronunciation Event and Articulatory Features. Second part is about Latent Dirichlet Allocation and the acoustic model selecting method using Latent Dirichlet Allocation. Latent Dirichlet Allocation is a Hierarchical math model that proposed by David M. Blei at 2003. It is often used on documents classification and document generation. The structure of LDA is also suitable for speech recognition and nature language processing. The final is experiment result and verification the method we proposed.

Keywords: Multilingual speech recognition, Latent Dirichlet Allocation, Acoustic Model Splitting, Acoustic Model Merging

一、緒論

(一)研究動機

由於目前網路與交通的發達，使得全球化成為必然的趨勢，再加上台灣本身就是屬於一個多族群社會，因此在日常生活環境中接觸到其他語言的機會也日進增加。除了平常所聽到和看到的資訊含有其他語言外，現在連平時對話也會經常含有英文、台語甚至是日語的情況產生。也因此語音辨識成為近年來熱門的科學研究項目之一，而且市面上也有許多相關應用類產品，例如：Google Android 平台的 Google Voice Search、Apple 的 Siri 語音助理…等，但目前這些平台都只能對單一種語言進行辨識，因此難以應付日進增加的多語環境，因此需要一種可以辨識多種語言的辨識器。

早期的多語辨識器第一步是先將語言類型辨識出來後，再將輸入的語音訊號送進對應的辨識器辨識。但是由於不同語言的音標集合並不完全相同，且多數的語音辨識器是針對特定語言的音標集合進行模型之訓練，因此若是在第一階段的語言種類辨識錯誤，而將輸入語句送入所對應的錯誤語言辨識器中，則所產生的結果將會是幾乎完全不符預期。而其綜合辨識正確率也會受到兩個元件的錯誤疊加而降低。為了解決此問題，將前後整合在一起設計一個辨識器可以直接辨識多國語言的架構則被採用，以減少錯誤率的累加。要將多個語言整合在同一辨識器內大致上可分為三種方法，第一種是將各個單一語言的音標集合併成共同音標集合，並以此為依據來建立多語的辨識器。第二種是使用專家知識所建立的跨語言音標集合來合併不同語言的音標，目前國際音標集合有國際音標(The International Phonetic Alphabet, IPA)、字母音標評估法(Speech Assessment Methods Phonetic Alphabet, SAMPA)、Worldbet 等。第三種是計算不同音標之間的相似性，並由估算出來的相似性來建立共同的音標集。

本文主要注重於使用專家知識所建立的跨語言音標集合之方法，在此主要是使用

國際音標(The International Phonetic Alphabet, IPA)，雖然 IPA 提供了一個通用的音標符號集，但在不同語言的情況下，會出現雖然屬於同一個音標，但是其發音模式仍然會有些不同的情況。或者是不同語言的不同的音標，但是其發音模式卻極為相似。為了解決上述之問題，因此將同音標但發音不同的聲學模型進行拆分，而音標不同發音類似的聲學模型進行合併，以減少混淆的情況。

(二)相關研究

為了解決發音變異而造成語音模型混淆的情形，國外學者根據不同的發音變異法進行了許多的研究。關於個人化發音變異的研究，1993 年 Hamada 和 Miki 等人提出，運用動態規劃(dynamic programming)和向量量化(vector quantization)的方式比較 Native Speaker 與 Non-Native Speaker 在同一個字發音上的差異[2]。1996 年 Neumeyer 和 Franc 等人則定義了 HMM Log-Likelihood、Segment duration 和 Timing 等特徵參數，針對整個句子做發音評估，而且實驗的結果發現 normalized segment duration scores 與專家給予的分數中有最高的相關性[3]。1997 年 Ronen 等人提出 MisPronunciation network 的概念。考慮每個音素在 native speaker 與 non-native speaker 的發音情形，建立對應的 HMM Model，辨識的時候發音網路同時考慮 native speaker 與 non-native speaker 的發音情形[4]。1999 年 Franco 等則是使用兩個分別由 native speaker 與 non-native speaker 所訓練的聲學模型，利用 log-likelihood ratio 來評估發音錯誤，同時也證明了這樣的方法比利用 a posterior score 的方式與專家所給予的分數有更高的相關性[5]。在模型方面，則是麻省理工學院的 Final State Transducer (FST) [6]。

Haizhou Li, Bin Ma, and Chin-Hui Lee 於期刊上發表的研究多語辨識[7]，提出了新的辨識單元構想，不再以音標為單元而是用人類實際發音的方法來做為單位。此方法是從訊號面尋找各種特徵，並以這些特徵建立出各種發音事件(Pronunciation Event)的辨識器，並將所有發音事件辨識器結果進行交叉比對出所要辨識單元之音標，此方法可以有效減低聲學模型數量，因此具有較佳的抗噪能力[8]。

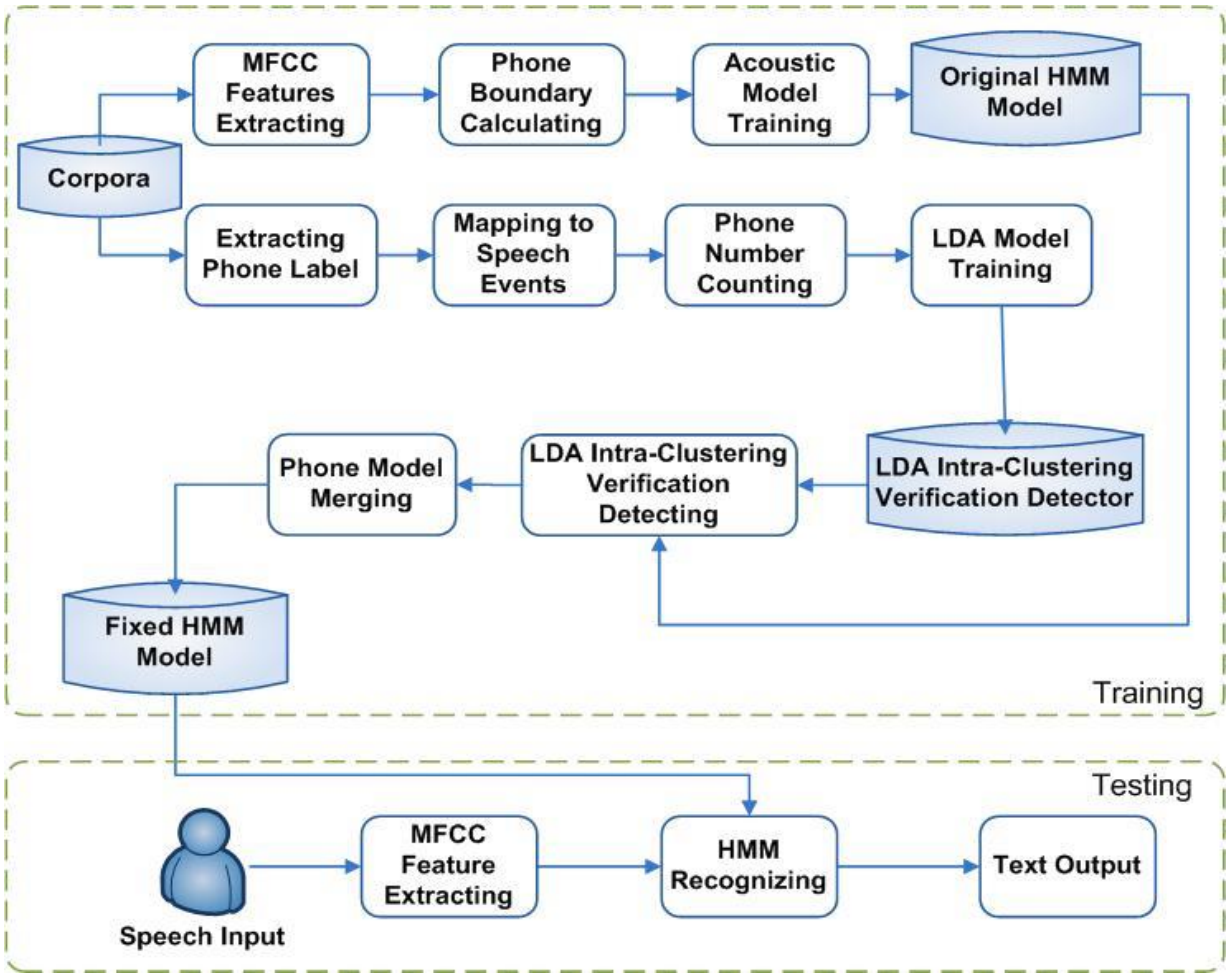
王小川[9]教授在”語音信號處理”一書中對於語音辨識的基礎知識背景和流程有相當詳細的敘述，對於想要學習語音辨識的人是一本很好的入門書籍。長庚大學則是長年來進行台灣本土語言的研究，梁敏雄[10]建立以 IPA 為基礎的台語音標集合 ForPA、文字轉語音(TTS)和發音偵錯應用於語言教學與台語語音語料庫設計與收集。陳志宇[11]則探討了同時對國台雙語的大詞彙連續語音辨識。楊永泰[12]將音標替換改變成中文而將隱藏式馬可夫模型運用在中文辨識上。

非本國母語人士所講的語言會產生發音變異，例如台灣人講英文，此種發音變異會使系統產生極大的誤判。國立成功大學的蔡佩珊、沈涵平、吳宗憲[13]提出以更小的單位 ─ senone 作為基本的辨識單元，以更加詳細的模擬發音變異，並建立包含發音變異之英文聲學模型。

二、系統架構

(一)系統框架

圖一為本論文之系統架構。根據其處理流程可以分為訓練階段與測試階段。訓練階段為使用狄式分佈進行對音素聲學模型的混淆偵測，測試階段則是將混淆發生的模型進行分割或合併後進行效能測試。



圖一、系統架構

(二)訓練部分

在訓練時的初始階段為使用 39 維 MFCC 參數和 HTK 來建立起以隱藏式馬可夫模型(HMM, Hidden Markov Model)為基礎的多語言 Tri-phone 聲學模型。此階段將不同語言相同音標視為不同音標，並且根據 Tri-phone 定義分裂模型，但是不進行 State-Tying 和模型合併。第二階段為使用語料標記建立以狄式分佈(LDA, Latent Dirichlet Allocation)為基礎的偵測器：群內驗證偵測器(Intra-clustering verification Detector)。過來根據以狄式分佈為基礎的群內驗證偵測器分別根據發音部位和發音方法進行分群，由於每一個音素都可以對應的單一的發音部位和發音方法，因此將發音方法與發音部位皆分類在同一群之音素定為合併目標，而將沒有兩者皆分類在同一群的音標示為不同之分群。最後再根據此分群對第一階段所建立的聲學模型進行合併，最後即可得到修正過之聲學模型。

(三)測試部分

測試部分最主要的是對訓練階段所修正的聲學模型進行效能與正確性測試。從使用者輸入的語句抽取 39 維 MFCC 參數和語音屬性並使用修正過的聲學模型進行辨識，並根據辨識結果來判斷是否有減少聲學模型混淆程度。

三、發音事件

為了解決語音辨識的瓶頸，美國喬治亞理工學院的李錦輝(C.-H. Lee)教授提出了偵測式(Detection-based)的方法，其主要概念為人類再發音的時候會有發音部位(Place)與發音方法(Manner)，藉由發音語言學來描述語音。而發音部位與發音方法則統稱為發音事件。而為了要偵測發音事件，則藉由直接觀察語音訊號來尋找出特定發音事件下的有效特徵來偵測，而這些特徵則稱為語音屬性(Articulatory Features)。其特性是藉由多層化架構而縮減了模型數量，也因為與人類發音的方法有所關連，因此其具有較高的強健性。因此對於不同語言相同音標也較不容易產生混淆。為了處理在多語辨識的環境下的音標混淆，本文使用了發音事件來進行偵測。

發音事件有許多不同的分類法，但其共通特性則為可以跨語言的對應到特定的音標，因此為了建立起跨語言的音標集合，本文主要使用了國際音標集合(Internal Phone Alphabet ,IPA)之定義來進行發音事件分類。經過對台語音標(ForPA)、國語音標和英文(KK 音標)對應後，使用到的發音方法(Manner)總共有六類：鼻音(Nasal)、塞音(Stop)、摩擦音(Fricative)、近音(Approximant)、塞擦音(Affricate)和顫音(Trill)。而使用的發音部位(Place)總共有十三類：雙唇音(Bilabial)、唇齒音(Labio-dental)、唇軟顎(Labio-velar)、齒音(Dental)、齦音(Alveolar)、齦後音(Post-alv)、捲舌(Retroflex)、齦顎音(Alveolo-palatal)、齦硬顎(Palato-alveolar)、硬顎音(Patatal)、軟顎(Velar)、小舌音(Uvular)和聲門音(Glottal)。

四、潛藏狄式分佈偵測器

(一)概述

潛藏狄式分佈是一種階層化的生成機率模型，是由 David M. Blei[1]於 2003 年所提出的，最初是用於文章和文本的主題偵測。狄式分佈模型有一個先決條件：詞袋假設(Bag of Words Assumption)，也就是不考慮詞彙(Word)在文章中的出現順序和文法關係，只考慮單一詞彙在特定文章中之出現次數。其所使用之原理為某些詞彙在特定主題下出現之機率和次數較高，因此狄式分佈的模型建立方式為統計各個詞彙(Word)在文本中出現的次數來計算。由於此方法在文本和文章分類可以取得相當好的成效，因此被廣泛的使用於主題偵測和主題分類的應用上。

在多語環境下，因為不同語言的語者間之發音變化幅度會相當的大，也因此若是根據以往的聲學模型只使用短期內(Short-term)的資料來辨識所能提升的效能幅度有限，而且在發音變異產生的狀況下也難以只使用訊號上短期的資料來加以識別。但語音訊號的時變性相當的大，因此在訊號面一次使用較長資料的方法難度也高。所以本文根據語料標記使用狄式分佈將整句的資訊也一併考慮以提升對發音變異之強健性。

(二)潛藏狄式分佈模型(Latent Dirichlet Allocation Model)

1. 潛藏狄式分佈(Latent Dirichlet Allocation)

潛藏狄式分佈(Latent Dirichlet Allocation, LDA)是由機率式潛藏語意分析(Probabilistic Latent Semantic Analysis, PLSA)延伸發展而來，而機率式潛藏語意分析又是由潛藏語意分析(Latent Semantic Analysis, LSA)發展而來。上述三個模型都是屬於潛藏主題模型(Latent Topic Model)，由於 N 連模型所面臨的缺乏長距離資訊和資料稀疏問題，而潛藏主題模型則是解決問題多種方法的其中之一。其概念為使用非監督式學習方法(Unsupervised Training)來找出隱含於文件或文章中的最主要的主題語意資訊。

潛藏狄式分佈不同於機率式潛藏語意分析的地方在於從文件到主題之間多了一層的狄式分布(Dirichlet Distribution)，使得模型參數數量不會因為語料增加而大幅度的增加，同時對於出現在語料庫外的文件也可以從狄式分佈中取出一個最適合此文件的潛藏主題機率分佈。由於潛藏狄式分佈是藉由逆向通過文件建立生成模型，因此要先理解潛藏狄式分佈是如何產生一篇文章的。

假設語料庫 M 中有 K 個主題，$T_1,T_2,T_3,\ldots T_k$，並且有 V 個字彙，當隨機選取一個主題 T_i 的時候，以 T_i 為主題的文章則有一的序列文字，而這些文字與 T_i 有關連，並且有一個機率值代表在主題 T_i 下時每個文字所出現之機率。而選擇其他主題時後也會有相同的參數來描述該主題。此時限定文檔長度為 N，不停的挑選文字直到數量到 N 後，便可以藉由對應的參數得到該建立文件對於各主題的相關性。

2.潛藏狄式分佈偵測器

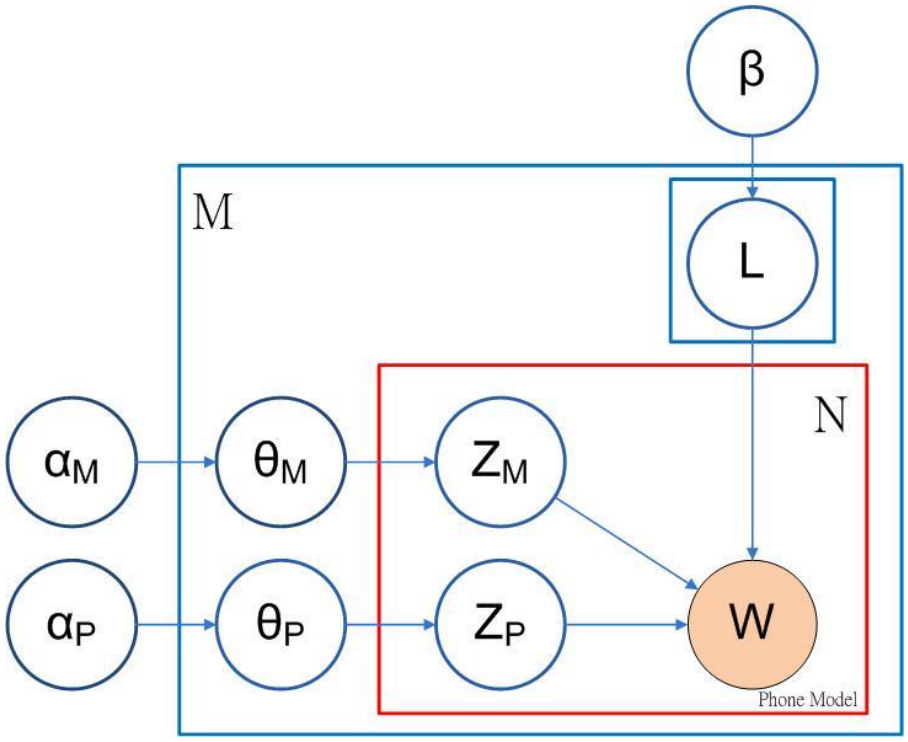
雖然偵測式方法(Detection Based)在多語環境下對於不同語言同音標之發音仍然具有高強健性不至於混淆，但是往往不同語言的語者對於相同音標之發音變異仍然會使得發音事件產生混淆，有鑒於此現象難以使用短距離的訊號來解決，因此我們使用長距離詞彙語意資訊來協助偵測出混淆的問題。並將會發聲混淆的聲學模型以及性質相同的聲學模型進行合併。不同於三連音素模型(Tri-Phone)的狀態聚類(State-Tying)根據短距離的資料相似度合併，我們藉由潛藏狄式分佈擷取長距離詞彙語意資訊，例如觸發詞對，以及使用了發音語言學的定義對分裂的三連音素模型進行聲學模型的合併。

由於潛藏狄式分佈具有詞袋假設的前提下，其所觀察到的資料長度足夠包含長距離的詞彙語意資訊。因為語音事件和語音屬性的相互關係也是屬於階層化架構，因此根據其聲學模型的相互關係對潛藏狄式分佈之架構重新建構後之圖型表示法如圖二 ，圖形中參數所代表之意義如表一。

在這裡將每一段語句(Utterance)視為一個文件(Document)，因此根據潛藏式狄式分佈定義則可以得到特定發音事件與語意上的相對關係。而聯合機率分佈為

$$p(M|\alpha_M,\alpha_P,\beta) \quad \text{(式 4.1)}$$

$$= \iint p(\theta_M|\alpha_M)\, p(\theta_P|\alpha_P) \left\{\prod_{n=1}^{N} p(Z_M|\theta_M)p(Z_P|\theta_P)p(w|Z_M,Z_P,L,\beta)\right\} d\theta_M d\theta_P$$



圖二、潛藏狄式分佈偵測器之圖型表示法

表一、潛藏狄式分佈偵測器之參數代表意義

符號	描述
α_M	K 向量，發音方法的 dirichlet 分布
α_P	K 向量，發音部位的 dirichlet 分布
β	所有語言下之聲學模型機率
θ_M	某段語句的發音方法發生之機率
θ_P	某段語句的發音部位發生之機率
W	聲學模型(phone model)
M	語料庫
N	某段語句的聲學模型集合
L	語言
Z_P	發音部位(Place)
Z_M	發音方法(Manner)

根據式 4.1 要計算的參數為 α_M, α_P ,β，但由於用來估算潛藏狄式分佈參數的最大期望演算法(Expectation-maximization algorithm, EM algorithm)一次只能估測兩個參數，而根據發音語言學之定義發音方法和發音部位兩事件為互相獨立(independent)，因此我們可以將其拆成 α_M、β 和 α_P 、β 兩個部分分開估測。

而拆開後分別的聯合機率則分別為為式 4.2 與式 4.3 所示

$$p(M|\alpha_M,\beta)=\int p(\theta_M|\alpha_M)\left\{\prod_{n=1}^{N}p(Z_M|\theta_M)p(w|Z_M,L,\beta)\right\}d\theta_M \quad \text{(式 4.2)}$$

$$p(M|\alpha_P,\beta)=\int p(\theta_P|\alpha_P)\left\{\prod_{n=1}^{N}p(Z_P|\theta_P)p(w|Z_P,L,\beta)\right\}d\theta_P \quad \text{(式 4.3)}$$

之後我們可以由 β 得到每個主題下所有音素的出現機率，而進而得到分類過後的音素集合。在原始的狄式分佈中會發生若單一文件過短會因為有效資訊過少而使得模型訓練過程無法收斂或者是影響到結果。但在我們的研究中，語句中的每一個音素對應到潛藏狄式分佈都可以視為具有資訊的單詞，因此即使只有 2-4 個字詞的短句的語句也具有足夠的資訊來加以判定其組成。

(三)語音事件降維與合併聲學模型之選擇

1.語音事件降維

在本文中將一段語句視為一個文件，因此根據潛藏狄式分佈之物理意義：若某個字詞(Word)從屬於某個主題，則當某文件屬於某個特定主題的時候，則從屬於某個主題的字詞出現次數會較高。將主題對應到的則是字詞(Word)。因為在特定的語言下，若以字詞為單位，有些音素經常性的會出現在一起，因此本文將原狄式分佈之主題數設定為字詞數量。

但由於字詞數量過多，而且潛藏狄式分佈之運算時間與主題數 N 成正比，若將所有可能出現的字詞數量設定為主題數，其主題數過多不僅在運算時間上不允許，也因為目標之集合過大，語料會面臨嚴重不足之情況，因此實際上並不可行。因此為了解決此問題，我們將字的組成音素根據國際音標集合之定義，將每個音素拆解成發音方法(Manner)與發音位置(Place)，而發音方法和發音位置在這裡即是語音事件。

2.合併聲學模型之選擇

本文根據國際音標集合定義將發音部位(Place)分為 13 類，而發音方法(Manner)定為 6 類，而母音則由舌面前後共 5 類、舌面高低共 7 類、唇形共 2 類。舌面前後為前(Front)、次前(Near-front)、央(Central)、次後(Near-back)、後(Back)。舌面高低為閉(Close)、次閉(Near-close)、半閉(Close-mid)、中(Mid)、半開(Open-mid)、次開(Near-open)、開(Open)。脣形為圓唇(Rounded)與非圓唇(Unrounded)。

我們會先將發音方法和發音部位使用不同的潛藏狄式分佈偵測器分別偵測並分群，最後再將發音部位相同但發音方法不同以及發音部位不同但發音方法相同的音素視為不同分群，換句話說則是只留下發音部位與發音方法皆分類再一起的音素進行合併。

五、實驗設計與分析

(一) 實驗語料與工具

台語實驗語料使用良敏雄博士所錄製之語料，語料為 16kHz, 16bit 之麥克風語料共 126000 句，標記為 ForPA，其中共有韻母 9 種、聲母 18 種和鼻音尾聲母 5 種。英文使

用語料為成大錄製的麥克風語料共 808 句。國語實驗語料從 TCC300 中選出 16kHz, 16bit 之麥克風語料共 2676 句。英文語料使用 TIMIT，語料為 16kHz, 16bit 支麥克風語料共 4620 句。特徵使用梅爾倒頻譜系數(Mel-scale Frequency Cepstral Coefficients, MFCC)、隱藏馬可夫(HMM)聲學模型訓練和聲學模型合併部分則是使用英國劍橋大學 HTK toolkit 來建立。潛藏式狄式分佈模型訓練則是以原作者 Blei et al.的 ToolKit 為基礎來進行修改。

隱藏馬可夫聲學模型訓練使用 39 維的梅爾倒頻譜系數，音框(frame)大小為 20ms，每次位移單位(Shift)為 10ms。隱藏式馬可夫模型每個狀態(State)留下 16 個路徑，最後再取前 10 名。所有語料皆無時間標記，時間斷點使用 Baum-Welch algorithms 計算。

(二)實驗項目

1.評估方式

要分析語音辨識的正確率必須準確的辨識出正確的字詞。辨識結果與語料標記互相比較後，結果與標記完全相符為辨識正確(Correct)，與標記不同的錯誤則根據定義分類成下列幾種錯誤：

(1)取代錯誤(Substitution Errors)：將正確的音素替換成其他音素。

(2)刪除錯誤(Deletion Errors)：沒有將該辨識的音素辨識出來。

(3)插入錯誤(Insertion Errors)：本來沒有的音素卻額外辨識出來。

為了評估本系統之效能，我們將分開分析取代錯誤率、刪除錯誤率、插入錯誤率以及字錯誤率(Word Error Rate, WER)和準確度(Accuracy)。計算公式如下式：

$$Substitution\ Errors\ Rate = \frac{S}{N} \quad \text{(式 5.1)}$$

$$Deletion\ Errors\ Rate = \frac{D}{N} \quad \text{(式 5.2)}$$

$$Insertion\ Errors\ Rate = \frac{I}{N} \quad \text{(式 5.3)}$$

$$Word\ Error\ Rate = \frac{S+D+I}{N} \quad \text{(式 5.4)}$$

$$Accuracy = \frac{C-I}{N} \quad \text{(式 5.5)}$$

其中 N 為語料標記內所有的音素總數，S 為取代錯誤數量，D 為刪除錯誤數量，I 為插入錯誤數量，C 為辨識正確字詞數量，C=N-D-S。

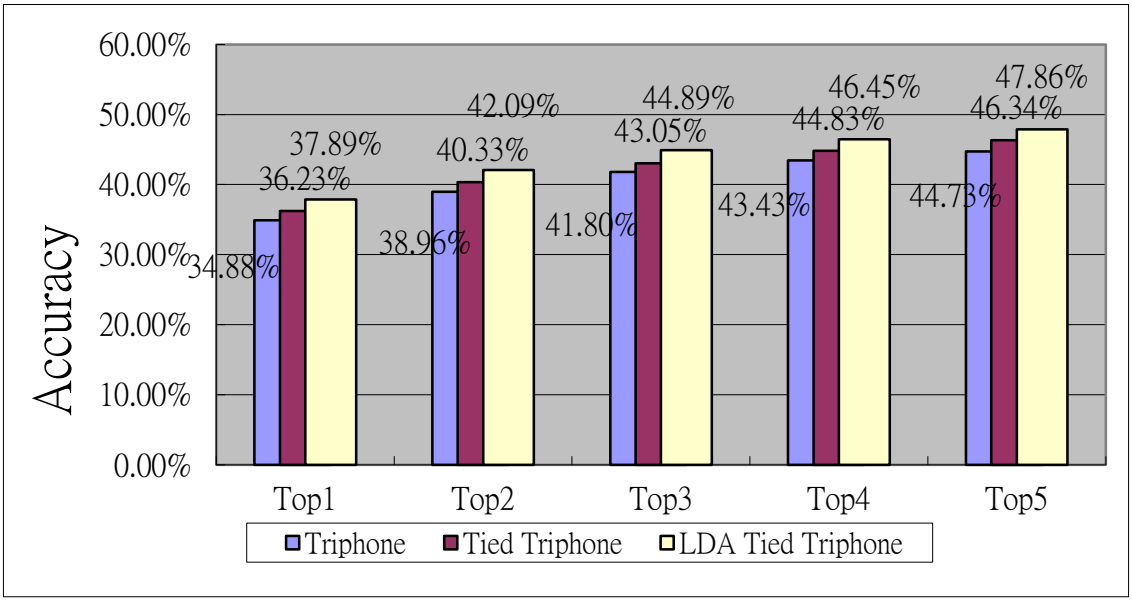
2.多語環境下潛藏狄式分佈合併聲學模型選擇之驗證

為了驗證潛藏狄式分佈合併聲學模型在多語環境下之效能，因此我們同樣分別對未聚類三連音素模型、已聚類三連音素模型和由潛藏狄式分佈合併聲學模型在國台英三語環境下進行驗證。我們將會分別使用已多語聲學模型辨識三語混合、國語、台語和英語實驗資料。以驗證在沒有語言辨識系統下多語聲學模型對於每一種語言之效能。

實驗環境和工具與單一語言進行之實驗相同，訓練語料和三語混合之測試資料為將 TCC300、TIMIT 和梁敏雄博士所錄製之台語語料混合使用，其實單語測試資料則和上述評估方式所進行的實驗使用相同的測試語料。語料內單一語句只有一種語言，也就是不做單一語句多種語言混合之實驗。所有語言的音素皆以 IPA 來表示。

(1)國台英三語混合實驗

結果如圖三所示，在多語環境下潛藏狄式分佈合併的聲學模型在 Top5 時候有 47.86%的最高準確率，且整體準確率皆高於其他兩種聲學模型。在 Top5 時的插入錯誤，本文提出之方法較已聚合三連音素模型多 0.24%，但取代錯誤和刪除錯誤則較已聚合三連音素模型分別少了 1.02%和 0.74%，因此整體的準確率較已三連音素模型多了 1.52%。因此可以驗證本文所提出之方法在多語環境下仍然有較佳的效能。

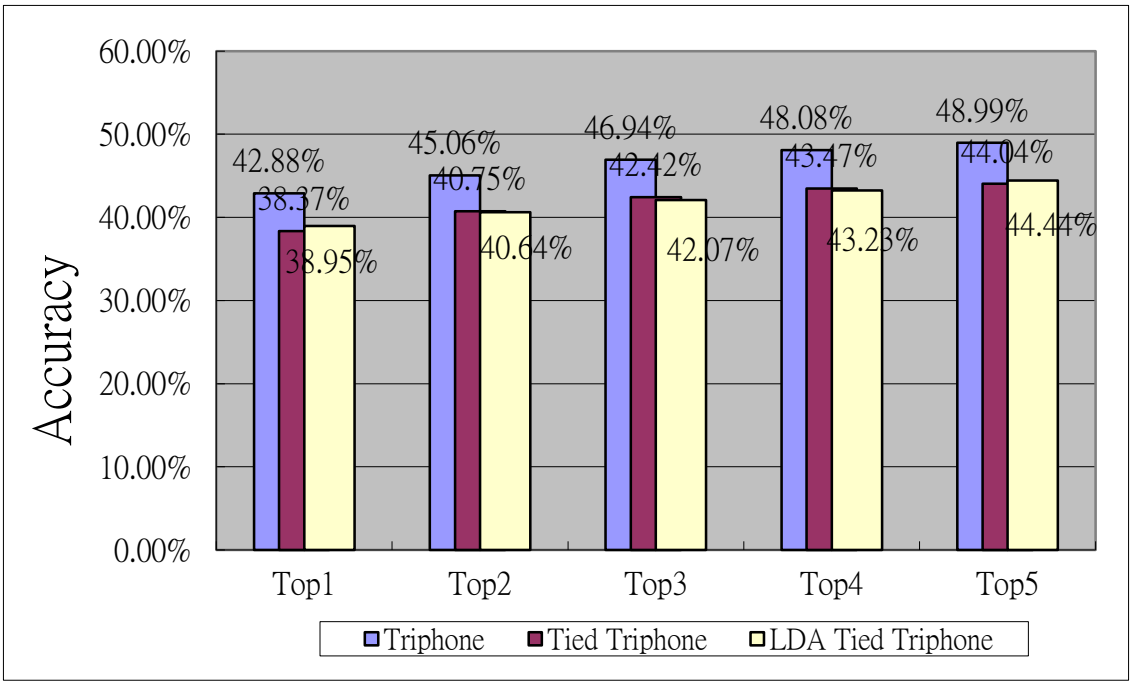


圖三、三種聲學模型在國台英混合下之辨識結果

(2)國語實驗

如圖四所示，未聚類三連音素模型在多語環境下辨識國語有最高的準確率。而有經過模型合併的聚類三連音素模型和潛藏狄式分佈選擇聲學模型擇是正確率差距不大，但兩者皆低於未聚類三連音素模型。而未聚類三連音素模型錯誤率改善主要集中在取代錯誤。潛在狄式分佈和聚類三連音素模型的比較在插入錯誤略高而刪除錯誤低則是與本文前列單語環境下之實驗相同。因此推測有可能是在進行多語模型合併時後國語的音標合併上出現問題或者是因為訓練語料過少而導致模型描述不夠全面，因此在多語環境下產

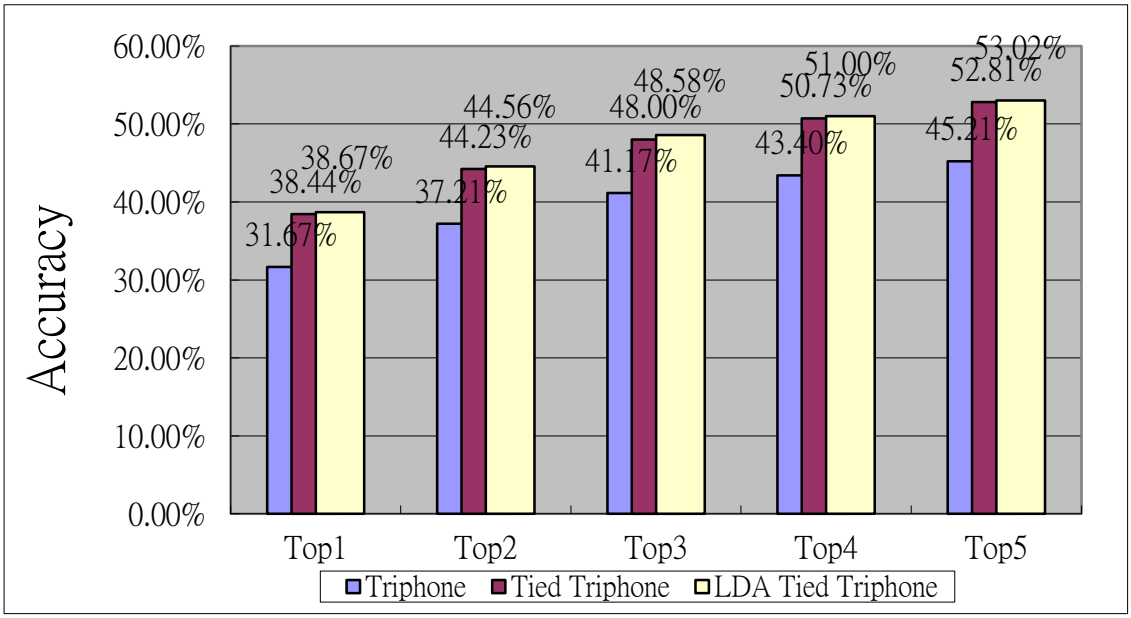
生了嚴重的模型混淆。



圖四、多語聲學模型辨識國語之辨識結果

(3)台語實驗

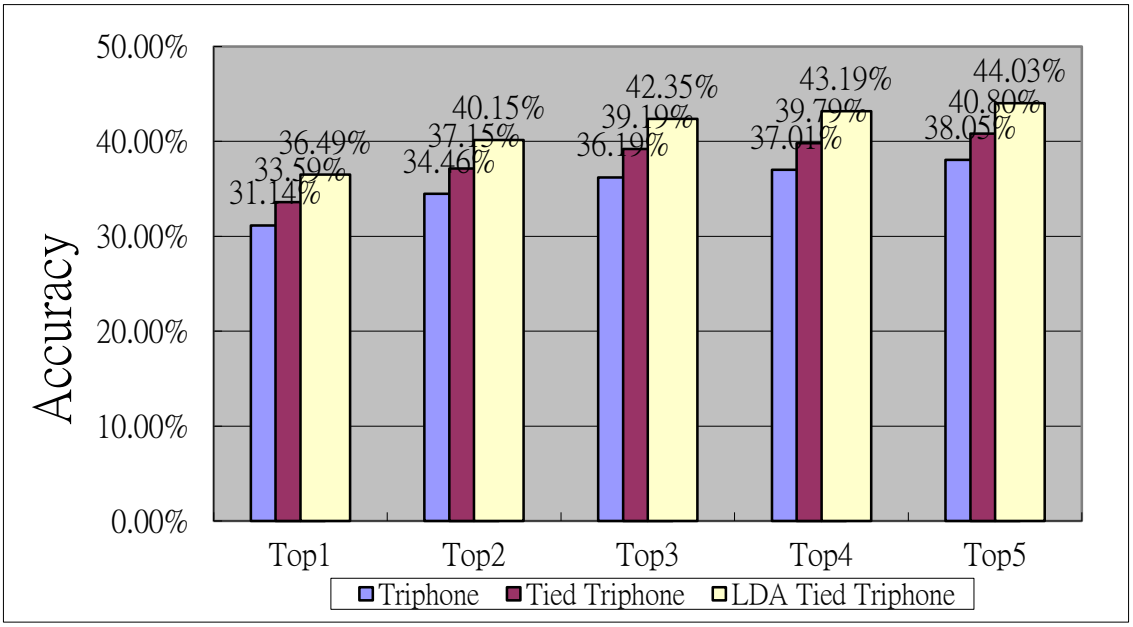
多語聲學模型辨識台語的環境下，潛藏狄式分佈選擇之聲學模型準確度在 Top5 時略高聚類三連音素模型 0.21%，而高未聚類三連音素模型 7.81%，如圖五所示。在 Top5 時取代錯誤較聚類三連音素模型低 0.23%、刪除錯誤多 0.02%而插入錯誤則是相同，整體看來仍略優於聚類三連音素模型。在 5.2.2 的單語環境下台語聲學模型實驗之結果顯示潛藏狄式分佈選擇的聲學模型準確率略低於聚類三連音素模型，而在多語環境下之準確率雖然略有改善，但依然與聚類三連音素模型差距不大。因此我們推測本文所提出的方法可能在我們所使用的台語語料環境下效能改善有限，但在多語的情況下仍然可以略高於聚類三連音素模型。



圖五、多語聲學模型辨識台語之辨識結果

(4)英語實驗

本文所提出之方法在多語環境下以辨識英語準確度提升率遠高於國語和台語，如圖六所示，在 Top5 時準確度較聚類三連音素模型高 3.16%，與未聚類三連音素模型相比則是高 5.98%。如表 5.9 所示，在 Top5 情況下，取代錯誤相較於聚類三連音素模型有 3.2%的改善，刪除錯誤則是有 0.63%的改善，但插入錯誤則是較聚類三連音素模型高 0.61%。因此可以看出本文提出之方法在多語環境下辨識英文可以減少取代錯誤和刪除錯誤，而整體的準確率也有所提升。



圖六、多語聲學模型辨識英語之辨識結果

六、結論與未來研究發展方向

由於目前網路與交通的發達，使得全球化成為必然的趨勢，而多語辨識在這個社會也愈來愈顯得重要。但有語言辨識的多語辨識存在著錯誤疊加的問題。不使用語言辨識的多語辨識方法被提出用來解決此問題，但不使用語言辨識的方法有不同語言間的聲學模型間容易混淆的問題存在，而本文提出使用潛在狄式分佈來進行聲學模型合併之選擇。所有語言音素皆以 IPA 表示，以達到參數共通之目的。並且將音素對應到發音部位和發音方法以減少數量，並且達到減少運算量之目的，使得潛藏狄式分佈偵測器得以運用長距離詞彙語意資訊，將常常先後成對出現的音素進行合併，以減少聲學模型的混淆。

實驗結果顯示，由潛藏狄式分佈所選擇合併的聲學模型和聚類三連音素模型以及未聚類三連音素模型相比在單語環境下辨識國語最高有 10.16%的準確率改善，而辨識英語則有 4.23%的準確率改善。在多語環境下混合辨識國台英三語混合的情況下有 0.24%的準確度改善，辨識英語有 3.16%的準確度改善。

本文所提出之方法雖然在多數的情況下都可以獲得準確率的改善，但是整體辨識率仍然偏低，多語辨識相較於單語辨識的聲學模型數量會隨著語言數量成倍數成長，而聲學模型數量愈多則需要更多的訓練語料來訓練。因此多語辨識經常面臨語料不足的問題，雖然經過模型合併後騎聲學模型數量仍然相當龐大。本文所提出之方法最終合併後的聲學模型數量遠大於聚類三連音素模型，因此需要更多的訓練語料。未來我們會持續的收集大量的語料，讓訓練資料更為完善，以改善語料不足的問題。

目前所提出之方法只使用長距離詞彙語意資訊來進行合併，因此當句子長度較短的時候準確度會有相當程度的下降，因此未來若可以同時考慮短距離的資料特性，例如三連音素模型聚合時所使用的馬式距離，或許可以同時在短句和長句都取得較佳的結果。

致謝

本研究承蒙中華民國國家科學委員會經費(99-2221-E-415-006)支持方得以完成，特別感謝。

參考文獻

[1] David M. Blei, Andrew Y. Ng, Michael I. Jordan, *Latent Dirichlet Allocation*, Journal of Machine Learning Research 3, pp.993-1022, 2003.

[2] Hamada, H., S. Miki, and R. Nakatsu. *Automatic Evaluation of English Pronunciation Based on Speech Recognition Techniques*, IEICE Trans. Inf. and Sys. 1993 E76-D(3):352-359.

[3] Neumeyer, L., H. Franco, M. Weintraub, and P. Price. *Pronunciation Scoring of Foreign Language Student Speech* In ICSLP' 96. Philadelphia, USA, Oct.

[4] Ronen, O., Neumeyer, L. and Franco, H. *Automatic Detection of Mispronunciation for Language Instruction*, Proceedings Eurospeech 97, Rhodes, Greece, 649-652.

[5] Franco, H., Neumeyer, L., Ramos, M., and Bratt, H. *Automatic Detection of Phone-Level Mispronunciation for Language Learning*, Proceedings Eurospeech '99, Budapest, Hungary, 851-854.

[6] H. Shu and I. L. Hetherington, *EM Training of Finite-State Transducers and its Application to Pronunciation Modeling*, Proc. ICSLP, Denver, CO, September 2002.

[7] H. Li, B. Ma, and C.H. Lee. *A Vector Space Modeling Approach to Spoken Language Identification‖, Audio, Speech, and Language Processing*, IEEE Transactions on vol. 15, NO. 1, JANUARY, pp 271-284, 2007.

[8] Sabato Marco Siniscalchi, Dau-Cheng Lyu, Torbjørn Svendsen, Chin-Hui Lee, *Experiments on Cross-Language Attribute Detection and Phone Recognition With Minimal Target-Specific Training Data*, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 3, MARCH 2012.

[9] 王小川, *語音信號處理, 二版*, 全華圖書股份有限公司, 2007。

[10] 梁敏雄, 呂仁園, *台灣多語語音資料庫之建立及應用*, 長庚大學博士文, 2008。

[11] 陳志宇, *國台雙語大詞彙與連續音辨認系統研究*, 長庚大學碩士論文, 2000。

[12] 楊永泰, *隱藏式馬可夫模型應用於中文語音辨識之研究*, 中原大學碩士論文, 2000。

[13] 蔡佩珊, 沈涵平, 吳宗憲, *發音事件驗證於多語辨識發音變異模型之產生*, ROCLING 2010, page 50-64。