利用聲學與文脈分析於多語語音辨識單元之產生

Generation of Phonetic Units for Multilingual Speech Recognition Based on Acoustic and Contextual Analysis

王士豪 [1]　　黃建霖 [2]　　吳宗憲 [2]

[1] 財團法人資訊工業策進會
Email: shwang@iii.org.tw

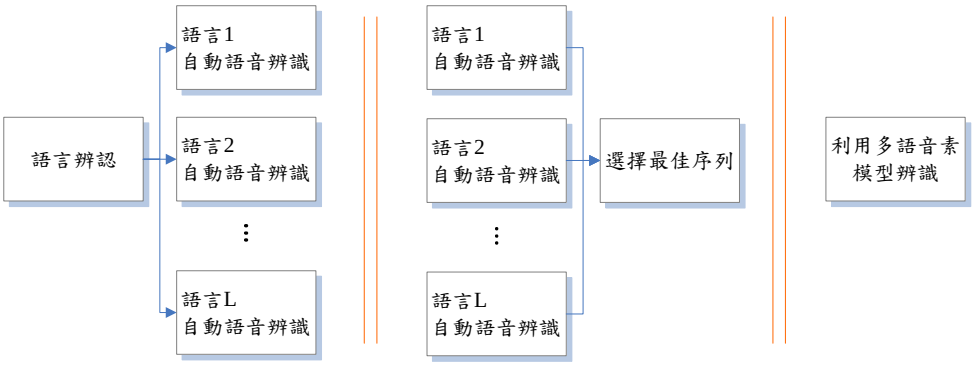
[2] 國立成功大學資訊工程系
Email: {chicco, chwu}@csie.ncku.edu.tw

摘要

由於全球化趨勢之盛行，多語語音常出現於會議紀錄及一般對話等方面。對於會議紀錄及對話系統而言，多語語音自動辨識日顯重要。在多語語音自動辨識中，辨識單元集之定義及選取，將影響辨識之效率及效能。本論文針對中英文利用 IPA 定義之多語語音辨識單元集，考慮前後文相關之三連音模型，並進一步透過對聲學相似度與前後文脈分析，決定一組精簡有效的多語辨識單元。在相似度矩陣分析中，首先我們利用事後機率統計，建立聲學相似度矩陣，然後，基於發音共聲現象的考量，分析語音發音上之相似度。本論文更引入語言超空間相似度之觀念，計算三連音辨識單元前後文脈之關係，建立語言超空間相似度矩陣。最後利用資料融合技術，合併聲學相似度矩陣和語言超空間相似度矩陣，以計算三連音辨識單元間之距離，而後利用向量量化群集方法合併相似性高之三連音辨識單元，建立一個有效的多語語音辨識單元集。本論文以 EAT 中英雙語語料庫作實驗評量，比較所提方法與之前研究方法上的差異與改進。由實驗結果得知，本論文所提出利用聲學相似度與前後文脈分析於多語語音辨識單元集之產生，可提高其辨識效能。

1. 簡介

語音是人類溝通最自然方便的方式，近年來電腦網路多媒體普及，語音在人機互動的介面更是扮演重要角色。自動語音辨識是語音應用技術的重要一環，目前有許多有關語音辨識之應用，如：自動聽寫機、語音文件摘要、語音文件檢索、口述語言對話系統以及語音命令控制等。並且隨著全球化趨勢的來臨，文化交流，商業活動和網路資訊都充斥著多語(multilinguality)的環境及各式各樣的應用，多語自動語音辨識(multilingual speech recognition)顯得相形重要。一般而言，多語自動語音辨識作法上可以歸納成三大類，如(圖 1)所示：



(a) 前端語言辨識方法　(b) 後端最佳序列選取方法　(c) 建立多語音素方法

圖 1. 多語語音辨識三類作法之流程

第一類作法：為在前端處理先利用一個語言辨認 (language identification, LID)方法[1][2]，判斷輸入語音訊號屬於哪種語言，再分別透過單一語言的自動語音辨識器進行轉譯。對於多語語音辨識演算法，前端語言辨認的成效決定了多語語音辨識的好壞，有效的語言辨認方法將使得多語語音辨識的正確率達到和個別單一語言的獨立語音辨識相當之效果。因此，利用先前的語言辨認方法，缺點在於辨識效果會受到 LID 表現影響。

第二類作法：利用個別單一語言的語音辨識器分別產生辨識結果，再經由後端處理選擇最大似然的方

法(maximum likelihood, ML)，決定最佳的多語辨識序列。作法上類似對語音辨識序列做驗證(verification)之處理 [3]；多語語音辨識的表現取決於後端最佳序列選擇之效果。利用選擇最大似然的方法缺點在於，多語辨識的效果會受到 ML 方法的限制，且辨識的多語句型需要另外考慮，切割出語句內不同語言的段落。

第三類作法：藉由定義多語語音辨識單元集[4]，合併個別單一語言之音素模型，來進行多語語音辨識。本論文乃基於此方法，探討如何定義出有效的多語語音辨識單元模型。

在多語音素模型之建立可以歸納為三種方式。首先，我們可以直接合併個別單一語言之音素集，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的分析，而是由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。第三，估計多語語音音素間相似程度，由下而上階層式進行多語音素合併，以定義多語音素集。多語語音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。但是缺點在於計算模型參數間的距離，與實際辨識演算法在執行時，所考慮的聲學相似度(acoustic likelihood)不符。

本論文探討中英文之多語語音辨識之研究，從中英文基本音素作分析。中文可以分為 37 個音素，英文可分為 39 個音素。考慮語音發音共聲的現象(co-articulation)，本論文定義前後文相關之三連音素模型(contextual tri-phone models)，進一步對語音發音相似度作聲學相似度(acoustic likelihood)分析。此外更導入語言超空間相似度分析(hyperspace analog to language, HAL)，考量三連音辨識單元前後文脈之關係，以改善過去單純考量模型參數聲學相似度來量測語音音素間相似度之方式，以決定多語音素模型，符合語音發音中受前後文影響之特性。最後，以資料融合的技術合併定義發音相似的音素。實驗評估，利用自行開發的多語語音辨識系統，使用隱藏式馬可夫模型(hidden Markov model, HMM)，建立以音素為基礎的聲學模型，並配合多語語言模型和多語發音辭典文法樹，進行連續多語語音辨識。

接下來的文章結構將分別探討如下：第二節，探討過去對多語語音辨識之研究。第三節，說明論文方法建立精簡有效的多語音素模型於自動語音辨識之應用。第四節，針對本論文所提方法建立之多語音素模型進行辨識結果評估，實驗並與之前方法比較。第五節是討論說明與結論。

2. 多語語音辨識之音素定義相關研究

多語語音辨識音素定義的方法，主要可分為三種方式：(一)直接結合個別單一語言之音素定義；(二)依據國際音素標準定義，找出個別單一語言之音素聯集；(三)從資料分析的角度(data-driven)，合併個別單一語言之相似音素。現分別介紹如下：

2.1. 直接結合個別單一語言之音素

如(表 1)所示，比照中文和英文單一語言音素的定義。

表 1. 結合中英文音素定義

音素類別	中文	英文
有聲破裂音	b_M, d_M, g_M	b, d, g
無聲破裂音	p_M, t_M, k_M	p, t, k
摩擦音	f_M, s_M, sh_M, h_M, x_M	f, v, th, dh, s, sh, hh
塞擦音	c_M, ch_M, j_M, q_M, z_M, zh_M	ch, jh, z, zh
鼻音	m_M, n_M	m, n, ng
流音	r_M, l_M	r, l
滑音		w, y
前部母音	i_M, v_M, ei_M, er_M	ih, eh, ae, iy, ey
中部母音	an_M, ang_M, en_M, eng_M	ah, uh, er
背部圓唇母音	o_M	ao
背部非圓唇母音	a_M, u_M, ou_M, e_M, ee_M, ai_M, ao_M	aa, uw, ow, ay, oy, aw

將各目標語言的音素合併成一個多語語音辨識音素集合，此方式是較為直覺的多語音素定義方式。(表 1) 內

之音素類別參考 Chomsky 定義[10]，本論文對中文音素記號多以"_M"標籤區隔，分別表示 37 個中文音素定義和 39 個英文音素定義。此方法結合中英兩種語言之音素，建立多語語音辨識之聲學模型。作法上的缺點，在於各目標語言中相似之音素，模型參數無法分享，而且當需要結合的目標語言變多的時候，所需要定義的音素模型會大量隨之增加。

2.2. 以 IPA 為基準定義多語音素

第二種多語音素定義方式是基於專家的知識，將個別獨立的單一語言對應到 IPA 標準的符號定義，藉此各語言間可以分享相同的音素定義。如(表 2)所示是以 IPA 為標準之中英多語音素的定義。

表 2. 以 IPA 為標準之中英多語音素定義

音素類別	IPA 為標準之中英多語音素
有聲破裂音	B, D, G
無聲破裂音	P, T, K
摩擦音	F, S, SH, H, X, V, TH, DH
塞擦音	Z, ZH, C, CH, J, Q, CH, JH
鼻音	M, N, NG
流音	R, L
滑音	W, Y
前部母音	I, ER, V, EI, IH, EH, AE
中部母音	ENG, AN, ANG, EN, AH, UH
背部圓唇母音	O
背部非圓唇母音	A, U, OU, AI, AO, E, EE, OY, AW

(表 2) 內之音素類別參考 Chomsky 定義[10]。如此規則地將中英兩種語言的音素結合，共計有 52 個中英雙語音素定義。作法上可以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。但此作法的缺點是建構在專家知識的分析，而非從資料特性統計的角度定義。也就是說，直接對照 IPA 定義產生的多語音素集，並沒有考慮到音素模型間頻譜特性。專家知識分析的多語音素集，與最後進行語音辨識，從資料分析角度建立的統計模型計算不一致。因此，採用直接對照 IPA 定義之多語音素模型並不能確實地呈現統計訓練資料上的分佈。

2.3. 量測音素相似度定義多語音素集

除了直接混合多語音素定義，以及利用 IPA 國際標準定義的多語音素，過去研究也曾利用估測三連音素模型間的相似度，以 HMM 模型參數距離計算，利用遞迴方法合併三連音素模型 (triphone)，建構出多語語音辨識的音素集[8][9]。兩個高斯分佈的相似可以利用平均值和變異數函數，來描述彼此的相似程度。利用 Bhattacharyya distance [8]來計算音素模型間的距離 D_{bha} ，表示如下：

$$D_{bha}=\frac{1}{8}(\mu_1-\mu_2)^T\left[\frac{\Sigma_1+\Sigma_2}{2}\right]^{-1}(\mu_1-\mu_2)+\frac{1}{2}\ln\frac{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}} \qquad \text{(式 1)}$$

其中，μ 和 Σ 分別表示音素模型的平均值和變異數向量，T 是轉置矩陣。另外，可以利用 Kullback-Leibler (KL) divergence [9]來決定兩個機率分佈的相似度 D_{KL} 。以 KL-divergence 估算兩個高斯分佈 $N(\mu_1,\Sigma_1)$ 和 $N(\mu_2,\Sigma_2)$ 的相似度，表示如下：

$$D_{KL}=\frac{1}{2}\left(\ln\frac{|\Sigma_1|}{|\Sigma_2|}+\mathrm{tr}\left(\Sigma_1^{-1}\Sigma_2\right)+\left(\mu_1-\mu_2\right)^T\Sigma_1^{-1}\left(\mu_1-\mu_2\right)-d\right) \qquad \text{(式 2)}$$

不論是以 D_{bha} 或者 D_{KL} 的評估方式，最後利用階層式由下而上進行音素相似度比較，依據音素模型的參數距離來考慮是否合併。其詳細的演算法表示如下：

1. 初始 Initialization:
建立各別單一獨立語言的音素模型。
2. 迴圈 Loop:
計算各音素模型間兩兩彼此的距離，並將最小距離的音素作合併。
3. 結束 Termination:
確認是否達到期望的音素定義個數，或者各音素之間的距離皆大於合併的最小標準。如果是，則結束迴圈；否則繼續進行第二步驟。

利用計算 HMM 模型參數(μ 和 $\sum$)的距離，來決定三連音素的合併，以建立多語音素模型。此方法的缺點與實際辨識演算法在執行時，所考慮的聲學相似度(acoustic likelihood)不符。為了改進上述方法的缺失，本論文提出利用聲學相似度及前後文脈分析，定義有效的多語辨識音素模型，詳細作法將於下節說明。

3. 利用聲學及前後文脈分析於多語音素模型定義

過去直接結合個別單一獨立語言之音素，建立多語音素集合作法上的缺點，在於沒有考量多語音素之間可能的共同音素分享。而直接利用 IPA 標準對照出多語音素集合，則缺乏實際資料統計和訊號特性上的分析。以 Bhattacharyya distance 或 KL-divergence 的作法，在於只有考慮音素模型的參數距離，並沒有針對實際辨識的應用做分析。因此，本論文提出聲學相似度及前後文脈分析自動建立多語音素模型定義，作法如(圖 2)所示。

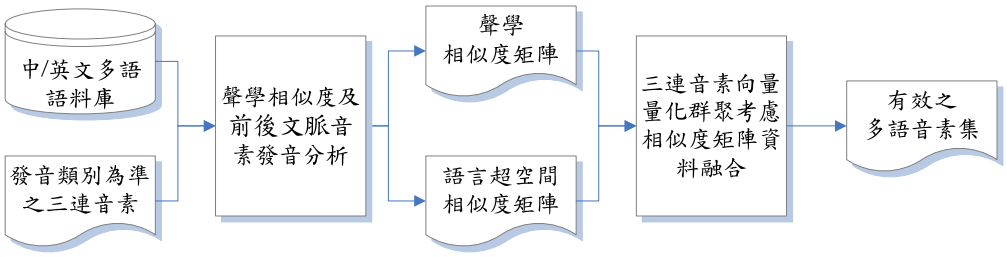


圖 2. 利用聲學相似度及前後文脈分析有效之多語音素集

對於目標語言以中文和英文為例，首先我們先對應 IPA 標準定義出中英文音素之定義，考慮前後音發聲的影響(left and right context dependent)，建立以 IPA 中英標準定義為基礎之三連音素集(triphone)。此外，從資料分析的角度(data-driven)針對聲學相似度及前後文脈做分析。對於聲學音素發音分析，計算各音素之 HMM 模型的聲學相似度(acoustic likelihood, ACL)，並建立聲學相似度矩陣。比較先前計算各音素模型之間參數距離的作法 [8] [9]，利用聲學相似度的做法更符合實際語音辨識上的考量。另外，配合語言超空間相似度分析(hyperspace analog to language, HAL) [11]，從前後文脈的分析，歸納出在連續發音中音素隨著前後文脈產生之發音特性，建立語言超空間相似度矩陣。聲學和語言超空間之相似度矩陣內，每一個音素可以利用向量表示，並計算出音素在空間中彼此相似度關係。最後，利用向量量化(vector quantization, VQ)的方法群聚相似的音素[12]，配合資料融合(fusion)的技術我們可以同時考慮聲學和前後文脈發音的影響，建立有效的多語音素集。

3.1. 聲學相似度分析

對應 IPA 音素標準定義，找出多語音素之聯集。考慮音素發音的位置和發音方法，判斷音素前後文連接的發聲情形，可以定義出 N 個三連音素模型。收集多語語音訓練語料以資料分析的角度，計算各音素在聲學上的相似度，建立聲學相似度矩陣。在聲學相似度矩陣的建立上，論文採用直接校準(forced alignment)的方法，與建立的 HMM 模型進行音素的辨識，利用直接校準方法可以確保參照出一樣的音素個數序列，避免辨識發生插入(insertion)和刪除(deletion)等錯誤的情形。統計語料內第 l 個音素模型與第 k 個音素模型間之相似度，其計算方式為將第 l 個音素所有訓練語音對第 k 個音素模型 ω_k 估算觀測事後機率值 $P(x_l^i \mid \omega_k)$，其中 x_l^i 表示第 l 個音素中之第 i 個訓練資料計算音素之間取對數用距離的方式呈現音素間彼此的關係 $\log(P(x_l^i \mid \omega_k))$，以建立聲學相似度矩陣 $\mathbf{A} = (a_{kl})_{N \times N}$。為建立一個對稱聲學相似度矩陣，我們對其計算對角平均值。

$$a_{kl} = \frac{\frac{1}{I}\sum_{i=1}^{I}\log(P(x_l^i \mid \omega_k)) + \frac{1}{J}\sum_{j=1}^{J}\log(P(x_k^j \mid \omega_l))}{2} \qquad \text{(式 3)}$$

其中，I 和 J 分別為第 l 個音素與第 k 個音素訓練語音之個數。

3.2. 前後文脈之語言超空間分析

本論文針對語言發音上相似度分析，進一步引入文件探勘(text mining)的觀念，模擬在一視窗長度 ℓ 內音素變調的情形。由數個相連音素所呈現語音發聲上的變化，研究引入語言超空間相似度分析(hyperspace analog to language, HAL) [11]，基於發音共聲的行為，如果兩個相同音素在前後文脈類似的情況下，這兩個音素在發音特性上則存在有高的相似性。藉由 HAL 的方法可以從提供音素發聲上潛藏相似度分析，作為判斷音素相似及合併的依據。在 HAL 空間中，與前後文脈相關的三連音素模型，可以利用一個向量來描述與其他音素發音共聲(co-articulation)的現象。每一個向量維度表示目標音素與其他發聲於前後文中音素關聯性的強度，關聯性高則有較高的權重。權重的計算藉由一個長度 ℓ 的觀測視窗，統計語料內各種發聲的情形，所有在觀測視窗內的音素被視為一起出現的共生單元。因此，在視窗內任意兩個音素間的距離 d ，則其權重的計算為 $w=\ell-d+1$。如(圖 3)所示，為 HAL 視窗長度與權重之結構圖。

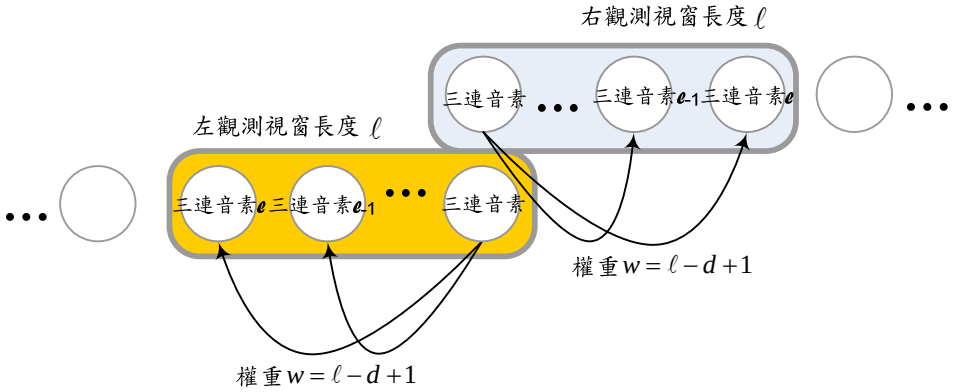


圖 3. HAL 視窗長度與權重之結構圖

在 HAL 空間建立步驟中，多語語音語料被視為一連串的音素發音序列。每串音素序列，藉由移動觀測視窗來計算語料內發音共聲的情形，每次移動一個音素。所建構出來的 HAL 空間是一個 $n\times n$ 大小的矩陣，其中 表示與前後文脈相關的三連多語音素個數。(表 3)是以“**Frank** (sil_F_R, F_R_AE, R_AE_NG, AE_NG_K, NG_K_sil) **早** (sil_Z_AO, Z_AO_sil)”的多語句子為例說明 HAL 空間矩陣 的計算方法，設定觀測視窗為 $\ell=3$ 。

表 3. HAL 空間矩陣

	sil_F_R	F_R_AE	R_AE_NG	AE_NG_K	NG_K_sil	sil_Z_AO	Z_AO_sil
sil_F_R							
F_R_AE	3						
R_AE_NG	2	3					
AE_NG_K	1	2	3				
NG_K_sil		1	2	3			
sil_Z_AO			1	2	3		
Z_AO_sil				1	2	3	

概念上，(表 3)的列向量(raw vector)表示音素與左邊文脈資訊的關聯；另外，(表 3)的行向量(column vector)則表示音素與右邊文脈資訊的關連。因此每一個音素 將由兩個向量維度呈現：

$$h_{l,k}=(v_l,v_k)=\left(\left\langle w_1^l,\ w_2^l,\ \ldots,\ w_N^l\right\rangle,\left\langle w_1^k,\ w_2^k,\ \ldots,\ w_N^k\right\rangle\right) \quad \text{(式 4)}$$

考慮音素發聲受到相鄰音素的影響，以三連音素 $h_{l,k}$ 為中心之空間相似度可以利用與右邊文脈相關之向量 v_l 及與左邊文脈相關之向量 v_k 之描述； w_N^l 和 w_N^k 分別表示利用觀測視窗於 HAL 空間內統計之音素相關權重，l 和 k 分別表示行與列之索引。

在 HAL 空間中，權重之計算需考慮正規化(normalization)因素，本論文利用在資訊檢索中相當重要之參數 $tf\times idf$ (term frequency and inverse document frequency) [13]，重新估計每個向量維度之權重，表示如下：

$$\bar{w}_i = w_i \times \log \frac{N}{C_i} \quad \text{(式 5)}$$

其中，w_i 指在向量 v_l 或向量 v_k 中第 i 個維度之權重；C_i 指在所有向量中，第 i 個維度之權重不為零的向量個數；N 為向量總個數或辨識單元個數。

3.3. 三連音素向量量化群聚考慮相似度矩陣資料融合

經過前兩小節分析，三連音素可以在聲學空間和語言超空間中，用向量的方式表示在空間中的相似程度。本論文分析聲學相似度矩陣 $\mathbf{A} = (a_{kl})_{N\times N}$ 和語言超空間相似度矩陣 $\mathbf{H} = (h_{kl})_{N\times N}$，同時考慮聲學相似度和前後文脈發音的特性，基於前後文脈相關之三連音素模型，合併相似之發音找出最為精簡有效的多語音素模型定義。作法上，參考資料融合的方法[14]，本論文利用加法融合的技術(sum rule)，結合兩相似度矩陣 $\mathbf{A}$ 和 $\mathbf{H}$，將聲學相似度和前後文脈的音素特徵作整合，表示如下：

$$\mathbf{S} = \alpha\mathbf{A} + (1-\alpha)\mathbf{H} = \sum_{l}^{N}\sum_{k}^{N}(\alpha \times a_{l,k} + (1-\alpha)\times h_{l,k}) \quad \text{(式 6)}$$

其中，α 是一個權重因子，負責融合聲學相似度和前後文脈的關聯。針對相似度矩陣 $\mathbf{A}$ 和 $\mathbf{H}$，論文中對其數值正規劃，將聲學和語言超空間相似度矩陣的分數結合，稱知識融合相似度矩陣 $\mathbf{S} = (s_{kl})_{N\times N}$ 為一個對稱矩陣，行 l 與列 k 均表示某一音素與其他音素相似度之向量。為了建立有效精簡的三連音素模型於多語語音辨識之應用，本論文利用向量量化(vector quantization, VQ)的方法[12]，從資料分析的角度(data-driven)，將原本三連音素自動地依據音素相似度分析，合併多語音素定義。向量量化為是一種非監督式的群集分析方法，可以將分散的資料群集成有意義的類別。三連音素在相似度矩陣分析後，可用向量方式表示其空間座標，論文引用[15]在矩陣中，兩向量夾角的計算方法，因此兩音素的相似度計算為 $c(s_l, s_k)$，計算如下：

$$c(s_l, s_k) = \frac{\overline{s_l} \bullet \overline{s_k}}{\left\|\overline{s_l}\right\| \cdot \left\|\overline{s_k}\right\|} = \frac{\sum_{i=1}^{N} s_i^l \times s_i^k}{\sqrt{\sum_{l=1}^{N} s_l^2} \times \sqrt{\sum_{k=1}^{N} s_k^2}} \quad \text{(式 7)}$$

其中，向量 s_l 表示目前相似度矩陣在行索引 l 的音素，向量 s_k 表示相似度矩陣在列索引 k 的音素，全部音素總共有 n 名。本研究利用調整性 k 群聚(modified k-means, MKM)分類方法[16]，定義收斂條件為分群內的資料變異度低於定義之門檻值，則達成分群終止，最後完成論文所提之有效多語音素集，其中收斂條件為：

$$(\sum_{y=1}^{Y}\Delta_y^t - \sum_{y=1}^{Y}\Delta_y^{t-1}) / \sum_{y=1}^{Y}\Delta_y^{t-1} < \theta \quad \text{(式 8)}$$

其中，Δ_y^t 表示在第 t 次遞迴中，Y 群集中第 y 群之集合內個數分數值 $\Delta_y^t = \sum c(s_l, s_k)$，$t = 1,...,t_{\max}$ 表示運算遞迴次數，$t_{\max}$ 指設定之最大遞迴次數，θ 為收斂之門檻值。

4. 實驗評估

為了評估研究方法，論文提出幾項實驗驗證：首先，實驗單獨考慮聲學相似度、語言超空間相似度與本論文所提結合聲學與語言超空間相似度分析之方法，比較其辨識結果。再者，比較與前後文脈獨立之音素集和論文所提與前後文脈相關之音素集在多語辨識準確率的差別。

4.1. 多語語音語料分析

本論文使用的多語語音辨識訓練語料，台灣腔英文(English Across Taiwan, EAT)語料庫，其中包含英文長句,英文短句,英文單詞及中英夾雜句等[17]。從 2004 年 5 月開始收集，至 2005 年 1 月初步完成收集，由師大、交大、清大、成大和台大等五所學校參與語料之錄製收集，經工研院電通所彙整。分別由英語系及非英語系學生錄製，語料依性別做分類，錄製有麥克風語料及電話語料，歸納如下表所列：

表 4. EAT 語料麥克風音檔資料統計

MIC 16khz 16bits 語料				
	英語系		非英語系	
	男性	女性	男性	女性
句數	11,977	30,094	25,432	15,540
人數	166	406	368	224

麥克風語料錄製 16KHz 取樣頻率 16bits 的取樣點音檔，電話語料錄製 8KHz 取樣頻率 16bits 的取樣點音檔，

其中電話語料又可細分為固定式電話(PSTN)語料及行動電話(GSM)語料，電話語料部份是透過 Dialogic 電話語音介面卡，錄得的 8KHz，8Bits，Mulaw 格式的取樣點,經程式轉成 8KHz，16bits，pcm 格式的取樣點；麥克風語料是由個人電腦及麥克風，直接從個人電腦的音效卡錄製 16KHz，16bits 的聲音訊號。最後將所有取樣點以 wav 格式音檔儲存。本論文研究採用麥克風語料部分。

每位語者收錄 80 句語音語料，語料內容設計有英文數字連續語音、英文字母連續語音、中英文混合句、英文單字、片語或句子等，如(表 5)所示。論文主要探討中英夾雜的多語應用，實驗抽取語料內中英文混合句型 (表 5 之 6 和 7)，語料編號#58 至#70 的音檔資訊。

表 5. EAT 語料中多語句型範例

	EAT 語料句型
1	four eight three zero one two nine
2	for instance
3	Safe
4	Silicon Graphics
5	R. S. R. T. E. K.
6	**冠軍家庭 T.V.秀入圍金鐘獎**
7	**幫我查一下 Bryan 的分機**
8	The vote at the September meeting was eleven zero

原本音檔內容皆屬於 raw 格式，因此我們事先對音檔作 dc-offset 及 silence removal 的處理。並且根據英文發音辭典與中文發音辭典，將文字註解轉成音素標記。由於語料內有部份音檔及錄音品質不良，實驗以人工的方式先行校對。最後，論文所採用之實驗語料包含有訓練用中英文混合句型共有 2,018 句，實驗評估測試共有 100 句。

4.2. 音素為基準之自動語音辨識架構

為了評估音素定義的好壞，本論文使用自行開發的多語音素辨識系統，探討多語語音辨識。採用上述之實驗語料中，我們利用 IPA 音素標準定義，找出多語音素之聯集。定義三連音素模型共 N=997 個，訓練語料少於 5 次的三連音素不予考慮。在語音參數擷取的部份，對於輸入的語音訊號計算 26 維的梅爾倒頻譜參數(mel-frequency ceptral coefficient, MFCC)，其中包含 12 階的梅爾倒頻譜參數，加上 12 階的一次微分梅爾倒頻譜參數，以及一階的能量和其一次微分參數，並且對參數做 MVA [18]處理以增加辨識的強健性。

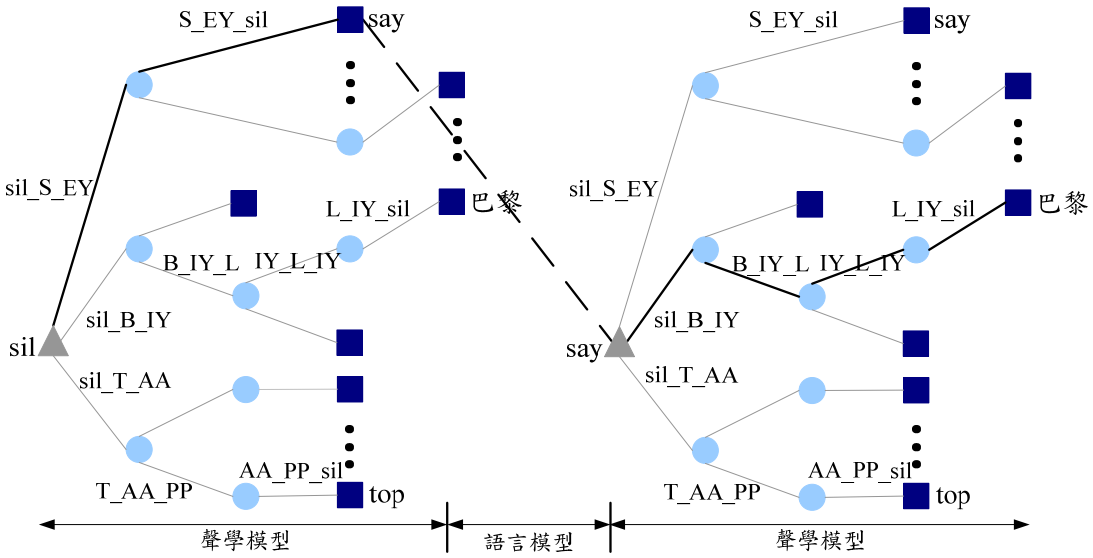


圖 4. 利用樹狀結構發音辭典文法樹於多語語音辨識之架構圖

本論文調整一般語音辨識使用的語言模型(language model)，在計算上利用均等機率(equal probability)的方法[19]，確保可以真正呈現不同多語音素定義的聲學模型(acoustic model)，對多語語音辨識的影響。在多語語音音素辨識，需要依據定義的多語音素結合各個目標語言的發音辭典，建構出一個多語發音辭典。

透過多語發音辭典，可以建構出多語發音之文法樹(grammar tree) [20]。如下(圖 4)所示。在辨識的流程上，每一個分支(arc)表示多語音素定義的 HMM 模型，研究上應用 3 個狀態(state)來描述每一個 HMM 模型，每一個狀態包含有 16 個高斯(mixture)。此多語之樹狀結構發音辭典舉例共有：say(sil_S_EY, S_EY_sil)、巴黎(sil_B_IY, B_IY_L, B_L_IY, L_IY_sil)、top(sil_T_AA, T_AA_P, AA_P_sil)等詞組。本實驗合併英文發音辭典與中文發音辭典，建立包含 29,104 個中英文詞之多語發音辭典。本圖示舉例說明由靜音(silence, sil)為起點，辨識多語語句"**say** (sil_S_EY, S_EY_sil) **巴黎** (sil_B_IY, B_IY_L, IY_L_IY, L_IY_sil)"為例，▲為樹的根節點；──線條表示多語音素也就是訓練的聲學模型，●指音素的節點；■表示葉結點，指出從根節點到此葉結點之發聲音素可能構成的所有多語詞彙；— —表示樹與樹之間連結的語言模型。

4.3. 利用聲學與語言超空間相似度分析群聚三連音素模型

語音辨識可能發生的錯誤有三種型態，分別是插入錯誤(insertion)、刪除錯誤(deletion)以及替換錯誤(substitution)。實驗中音素正確率(accuracy)的計算[21]，方式如下：

$$Accuracy = \frac{len - ins - del - sub}{len} \times 100\% \qquad \text{(式 9)}$$

其中，*len* 為辨識結果，音素序列的長度。*ins* 為比較較正確結果多辨識出的音素，屬於插入錯誤，*del* 為比較正確結果少辨識到的音素，屬於刪除錯誤。*sub* 為比較正確結果辨識錯誤的音素，屬於替換錯誤。分析不同群集條件下的群聚音素個數，利用調整 k 群聚(modified k-means, MKM)分類方法[16]，群聚三連音素模型為有效多語辨識模型。實驗聲學相似度計算(ACL)、語言超空間相似度計算(HAL)及資料融合技術(FUN)等不同方法，在收斂門檻值為 $\theta = 0.01$ 的情況下，實驗不同最大群集數 Y 。(表 6)實驗分析各種不同方法群集之多語音素個數及音素辨識的正確率，如下所示：

表 6. 不同群集數目限制條件下群聚音素個數及辨識正確率 (Y :最大群集數目, $\theta = 0.01$)

	$Y=8$		$Y=16$		$Y=32$	
	正確率	音素個數	正確率	音素個數	正確率	音素個數
ACL	62.22%	161	63.12%	288	64.37%	531
HAL	62.52%	159	64.23%	286	64.57%	530
FUN	64.44%	119	66.07%	260	64.74%	515

實驗考慮聲學相似度矩陣分數計算，利用聲學相似度群聚方法 ACL，在 $Y = 8, 16, 32$ 的情況下，分別可以群聚為 161，288 及 531 個多語音素模型，其音素辨識正確率分別為 62.22%，63.12%及 64.37%。利用語言超空間分析方法 HAL，在 $Y = 8, 16, 32$ 的情況下，分別可以群聚為 159，286 及 530 個多語音素模型，其音素辨識正確率分別為 62.52%，64.23%及 64.57%。利用資料融合方法 FUN，在 $Y = 8, 16, 32$ 的情況下，分別可以群聚為 159，286 及 530 個多語音素模型，其音素辨識正確率分別為 64.44%，66.07%及 64.74%。

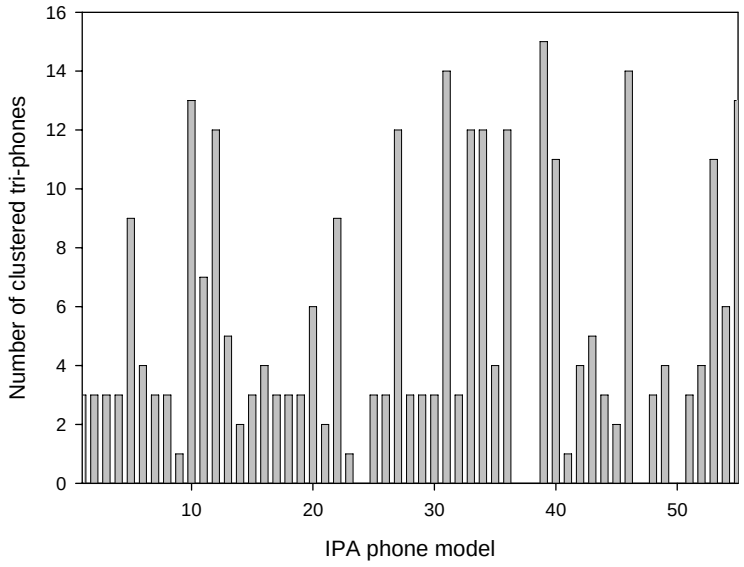


圖 5. 利用聲學相似度及前後文脈分析群集三連音素模型分佈圖，$Y = 16$

利用前後文脈分析方法 HAL 比聲學相似度方法 ACL 有較高的準確率，而同時結合聲學相似度與前後文脈分析 FUN 可以有最佳的辨識效果。當群集數 $Y=16$ 時，論文所提之方法(FUN)可以有最好的辨識效果。因此，論文設定群集分析 $Y=16$，群集三連音素模型，分析如(圖 5)所示。經過分類完後，各個 IPA 定義之音素中，所包含之三連音個數。由上圖可知，以 55 個 IPA 標準定義所產生之 997 個三連音素模型，利用資料融合方法可以合併為 260 個多語音素模型。

4.4. 聲學與語言超空間相似度分析於多語語音辨識

本論文研究探討中文和英文的多語語音辨識應用，實驗首先測試使用單音素模型(monophone)的定義，依據(表 1)和(表 2)等不同標記方法的內容，分別可以定義： (一) 直接結合個別單一語言之音素(MIX)；(二) 以 IPA 為基準定義多語音素之方法 (IPA)。實驗結果如(表 7)所示：

表 7. 單音素模型之多語辨識音素正確率 (括弧內表辨識單元之個數)

========== English Across Taiwan, EAT ==========				
----------Monophone Tree-Search Results----------				
	ACCURACY	INSERTION	DELETION	SUBSTITUTION
IPA phone sets (78)	**55.35%**	13.54%	5.27%	28.92%
Mix phone sets (55)	**55.49%**	21.94%	4.98%	18.06%
========== English Across Taiwan, EAT ==========				

實驗 EAT 語料中，有關中英文混合句型的連續語音分析。定義中文基本音素共有 37 個，英文基本的音素定義有 39 個。由此實驗結果可知，使用直接合併多語音素(MIX) 的辨識正確率為 52.35%，而採用 IPA 標準定義的中英文音素集，正確率為 55.49%。因此具有多語模型參數分享特性的 IPA 標準定義，辨識結果比較直接合併多語音素的方法，在辨識效果上來得顯著。再者，實驗比較原本以 IPA 為基準之三連音素模型定義(triphone sets)，利用語言超空間相似度矩陣群集音素定義(HAL phone sets)，利用聲學相似度矩陣群集音素定義(ACL phone sets)，以及利用資料融合方法於聲學及語言超空間相似度矩陣分析之音素定義(FUN phone sets)，如(表 8)所示：

表 8. 三連音素模型之多語辨識音素正確率 (括弧內表辨識單元之個數)

========== English Across Taiwan, EAT ==========				
----------Triphone Tree-Search Results----------				
	ACCURACY	INSERTION	DELETION	SUBSTITUTION
Triphone sets (997)	**68.07%**	15.87%	4.43%	11.63%
ACL phone sets (288)	**63.12%**	19.73%	4.88%	12.32%
HAL phone sets (286)	**64.23%**	20.67%	4.75%	10.48%
FUN phone sets (260)	**66.07%**	16.94%	4.41%	12.71%
========== English Across Taiwan, EAT ==========				

由實驗結果可知，合併前的三連音素模型(Triphone sets)的多語辨識效果可達 68.07%的正確率。利用聲學相似度矩陣群集音素定義(ACL phone sets) ，在多語辨識效果上可達 63.12%的正確率，而利用語言超空間相似度矩陣群集音素定義(HAL phone sets) ，在中英文多語辨識效果上可達 64.23%的正確率。進一步利用資料融合方法於聲學及語言超空間相似度矩陣群集分析之音素定義(FUN phone sets) ，在多語辨識效果可以提升至 66.07%的正確率。整體而言，採用三連音素的辨識效果比單音素(IPA 或 MIX)定義好。又從語言分析(HAL)效果會較聲學分析(ACL)效果來得顯著，且利用資料融合方法結合聲學相似度及前後文脈分析，對於多語語音辨識可以有明顯的提升。

5. 結論及未來展望

本論文提出應用聲學相似度及前後文脈分析於多語語音辨識之有效音素定義，以 EAT 中英文雙語語料為例。基於 IPA 標準定義之多語單音素集，本研究考慮以發音前後文相依三連音素模型。以此定義，我們分別以聲學相似度及前後文脈分析，音素間相似度高的音素合併，期望找出精簡有效的多語語音辨識音素集。利用音素 HMM 模型，以直接校準方法切音並計算事後機率值，建立聲學相似度矩陣。利用語言超空間相似度分析(hyperspace analog to language, HAL)，找出音素前後發音特性所造成的變音影響，建立語言發音上相似度矩陣。之後，以資料融合方法，同時考慮聲學和語言超空間相似度矩陣。利用向量量化群集分析，找出同一類別之音素定義，建立有效而精簡的多語音素集。實驗證明利用結合聲學和語言超空間相似度矩陣分析方法，可以達到良好的多語連續語音辨識的效果。未來可以將方法應用在單一語言語音辨

識之音素定義研究，也可以進一步分析在更多目標語言下，此方法對於多語語音辨識的效果表現。

參考文獻

[1] Chung-Hsien Wu, Yu-Hsien Chiu, Chi-Jiun Shia, and Chun-Yu Lin, 2006. Automatic Segmentation and Identification of Mixed-Language Speech Using Delta-BIC and LSA-Based GMMs. IEEE Transactions on audio, speech, and language processing, vol. 14, no. 1, pp. 266-276.

[2] Alex Waibel, Hagen Soltau, Tanja Schultz, Thomas Schaaf, and Florian Metze, 2000. Multilingual Speech Recognition. Chapter in Verbmobil: Foundations of Speech-to-Speech Translation, Springer-Verlag.

[3] Rafid A. Sukkar and Chin-Hui Lee, 1996. Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Subword based Speech Recognition. IEEE Transactions on Speech and Audio Processing, vol. 4, no. 6, pp. 420-429.

[4] Yeou-Jiunn Chen, Chung-Hsien Wu, Yu-Hsien Chiu, and Hsiang-Chuan Liao, 2002. Generation of robust phonetic set and decision tree for Mandarin using chi-square testing. Speech Communication, vol. 38(3-4), pp. 349-364.

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

[6] J. C. Wells, 1989. Computer-Coded Phonemic Notation of Individual Languages of the European Community. J. IPA, 19, pp. 32-54.

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

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

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

[10] Chomsky, N. and Halle, M., 1968. The Sound Pattern of English. New York: Harper & Row.

[11] Burgess, C. and Lund, K., 1997. Modelling parsing constraints with high-dimensional context space. Language and Cognitive Processes, 12:177–210.

[12] Robert M. Gray and David L. Neuhoff, 1998. Quantization. IEEE Transactions on Information Theory, vol. 44, no. 6, pp. 2325-2383.

[13] G. Salton and C. Buckley, 1988. Term-weighting Approaches in Automatic Text Retrieval. Information Processing Management, vol. 24, no. 5, pp. 513-523.

[14] Josef Kittler, Mohamad Hatef, Robert P.W. Duin, and Jiri MatasOn, 1998. On Combining Classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 3, pp. 226-239.

[15] Jerome R. Bellegarda, 2000. Exploiting latent semantic information in statistical language modeling. Proc. IEEE, vol. 88, no. 8, pp. 1279-1296.

[16] Jay G. Wilpon and Lawrence R. Rabiner, 1985. A modified K-means clustering algorithm for use in isolated work recognition. IEEE Transactions on Acoustics, Speech, and Signal Proc., vol. 33, no. 3, pp. 587-594.

[17] English Across Taiwan, EAT [online] http://www.aclclp.org.tw/

[18] Chia-Ping Chen, Jeff Bilmes and Daniel P. W. Ellis, 2005. Speech feature smoothing for robust ASR. in Proc. ICASSP, Philadelphia PA.

[19] Johnston, D., 1997. Statistical Methods for Speech Recognition. The MIT Press, Cambridge, MA.

[20] H. Ney and S. Ortmanns, 2000. Progress in dynamic programming search for LVCSR. Proceedings of the IEEE, vol. 88, no. 8, pp. 1224–1240.

[21] Steve Young, Gunnar Evermann, Mark Gales, Tomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, Phil Woodland, The HTK Book (for HTK Version 3.3).