國立交通大學
National Chiao Tung University

出國報告（出國類別：出國短期研究）

題目：「整合發音及韻律訊息之大詞彙語音辨認研究」

服務機關：電機工程學系
姓名職稱：江振宇 博士後研究員
前往國家：美國亞特蘭大喬治亞理工學院
出國期間：2012/04/01~2012/06/01
報告日期：2012/06/29

一、摘要

江振宇博士受「美國喬治亞理工學院-電機與電腦工程學院的信號與影像處理中心」李錦輝(Prof. Chin-Hui Lee) 教授邀請，於民國 101 年 4 月 1 日~民國 101 年 6月1日這段期間，接續101年1月至3月之研究，與另一位義大利訪問學者 (Prof. Sabato Marco Siniscalchi, Department of Telematics, Kore University of Enna, Enna, Italy)開始合作「整合發音及韻律訊息之大詞彙語音辨認研究」(Knowledge Integration for Improving Performance of Large Vocabulary Continuous Speech Recognition) 之相關工作，主要工作項目有二：1)將發音及韻律訊息加入傳統馬可夫聲學/語言模型 (HMM-based AM/n-gram-based LM) 之大詞彙英語辨認系統，目前已有初步結果，預計將結果投稿於 ICASSP2013 國際會議及 Speech Communication 期刊論文、2)將發音方法訊息加入交大方面發展之「韻律訊息輔助之大詞彙語音辨認系統」，以協助驗證詞格(word lattice)上的假設音節結構(hypothesized syllable structure)，研究結果已投稿至 ISCSLP 2012。

二、目次

三、本文

（一）目的

利用發音訊息或韻律訊息於輔助語音辨認，在這幾年的語音處理研究中越來越受重視。在發音訊息部分，喬治亞理工的李錦輝教授，於 2005 年左右，提出了一個稱為「語音屬性為基礎之語音辨認系統」(Detection-Based ASR)為架構的語音辨認系統，此架構追本朔源地以人類語音辨認(Human Speech Recognition)為出發點設計，其中包含語音知識源偵測(knowledge detection)以及語音知識源整合(integration of knowledge source)這兩部分。語音知識源偵測(knowledge detection)之目的為偵測語音辨認所需要的訊息，這些訊息主要包含發音方法(Manner of articulation)、發音位置(place of articulation)。語音知識源整合(integration of knowledge source)的目的在整合語音知識源偵測(knowledge detection)得到的資訊，利用一些圖像辨識(pattern recognition)或是動態搜尋(dynamic programming)的方法引入語音知識源(knowledge source)進行整合，得到語音辨認的目標單元，如音(phone)、音節(syllable)或詞(word)，或是在音(phone)/音節(syllable)/詞(word) 格(lattice)上提供音(phone)、音節(syllable)或是詞(word)的信心分數，以協助傳統語音辨認進行重新計分(rescoring)。語言知識源(knowledge source)可以是特定語言(language dependent)或所有語言(language independent)的特性或模型，因此在「語音屬性為基礎之語音辨認系統」(Detection-based ASR)架構下，語音處理中不同領域的專家，便可以專注於設計其擅長之語音知識源偵測器(knowledge detector)，並提供其熟悉之研究領域知識(domain knowledge)，以便進行知識整合(knowledge integration)，以期達到更佳之語音辨認效果。目前，「語音屬性為基礎之語音辨認系統」(Detection-based ASR)已成功運用於英語的語音辨認，不但能以詞格重新計分(lattice rescoring)的方法提升辨認率，亦可使用由下往上(bottom-up)的整合方法進行音素構詞以及大詞彙辨認。

在韻律訊息方面，主要議題有 1.加入韻律訊息的語音辨認架構設計(Framework of automatic speech recognition assisted with prosodic information)、2.

韻律模型的設計及建立(design and construction of prosodic model)、3.韻律相關之聲學模型(prosody-dependent acoustic model)、4.韻律訊息相關之語音模型(prosody-dependent language model)、5.韻律模型對不同語者的調適(speaker adaptation for prosodic model)。我們基於先前對於「非監督式中文韻律標記及模式的研究」，已將此研究的韻律模型應用於輔助中文大詞彙連續語音辨認，並得到不錯的結果，大幅提升了詞、字以及音節辨認率，另外，在韻律相關聲學模型(prosody-dependent acoustic model)方面，我們亦利用其中的韻律斷點資訊建立韻律斷點相關(break-dependent)的聲學模型，在音節辨認的實驗裡可以得到些許的音節辨認率提升，亦可在音節格(syllable lattice)上得到較高的正確音節涵蓋率(coverage rate)，另外的好處是可以在音節格(syllable lattice)上提供韻律斷點(break)資訊，以利後級進行語言及韻律資訊整合，以得到更好的詞辨認率。

基於上述之背景， 本計畫目的在於結合 1.李錦輝教授之「語音屬性為基礎之語音辨認系統」(Detection-based ASR)架構、2.交大語音處理實驗室所建立的韻律模型，擬將原本使用於中文的韻律模式延伸至英語韻律模式，亦將原應用於英語之語音屬性偵測前級(Detection-based ASR frontend)應用於中文大詞彙變認，使中文和英文的大詞彙語音辨認系統，能因為發音及韻律訊息的加入，達到較佳之語音辨認率。

（二）過程

本人於 2012 年的 4 月 1 日到 6 月 1 日期間，於喬治亞理工電機與電腦工程學院的信號與影像處理中心，延續之前 2012 年 1 月 3 日到 3 月 31 日之研究，除了與李錦輝老師固定於每周三下午 5:00 到 6:00 進行密集討論，並開始和 Prof. Sabato Marco Siniscalchi 進行有關如何將韻律和發音訊息加入大詞彙語音辨認的討論，以下分別以時間順序對於這段時間討論及初步結果進行簡述：

(1) 2012 年 4 月 1 日~2012 年 5 月 1 日

與 Prof. Sabato Marco Siniscalchi, Department of Telematics, Kore University of Enna, Enna, Italy 討論研究細節，在韻律模型的設計考量上，如 Fig 1 所示，決定

將韻律段點(Break type)作為高階層的語言參數(如詞類、詞長)和韻律訊息來做為傳遞韻律訊息及韻律結構的介面，如此，韻律模型便可以協助語音辨認，提供有效的韻律訊息，另外，除了傳統聲學模型(AM)及因子語言模型(FLM)外，也考慮使用發音方法(manner of articulation)資訊協助語音辨認，如此，韻律訊息提供的便是由上往下(Top-down)的高階訊息(high-level information)，而發音方法訊息便是提供由下往上(bottom-up)的低階訊息 (low-level information)。

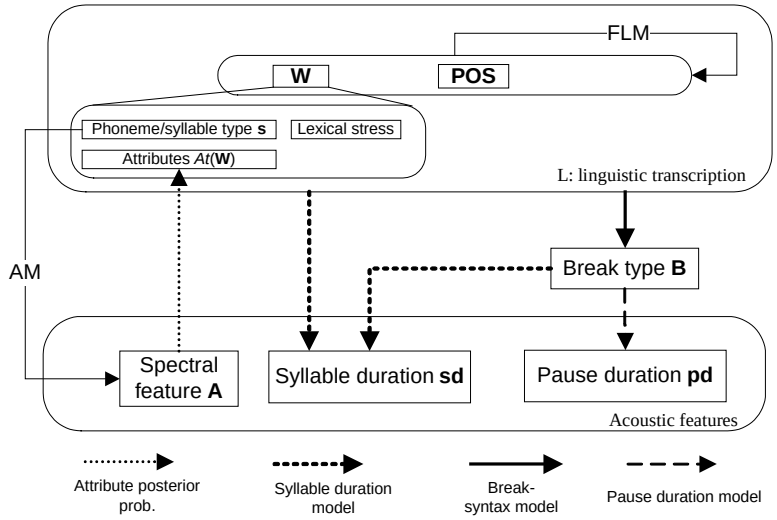

Fig. 1: 以韻律及發音方法訊息協助之大詞彙語音辨認示意圖

以 Fig. 1 所呈現的概念，我們開始設計並建立辨認器架構，辨認方法是以傳統語音辨認器先產生詞格(word lattice)，再將韻律訊息及發音方法訊息的分數/機率放入詞格(word lattice)進行重新計分，辨認器架構及各個工作方塊(building block)如Fig. 2所示。語音信號經由音框參數抽取(Frame-based feature extraction)，抽取出語音的頻譜及音高參數，接下來利用基礎語音辨認器使用光束搜尋(beam search)方法，採用傳統馬可夫聲學模型(HMM-based acoustic model)和詞雙連文語言模型(word bigram LM)產生詞格(word lattice)，此詞格(word lattice)含有音節切割資訊(syllable segmentation information)以及詞。接下來，我們使用因子語言模型(Factored LM)將原本的詞格(word lattice)以格展開器(lattice expander)將詞格(word lattice)展開成含有詞、韻律斷點訊息(Break information)、詞類訊息(POS information)以及音節切割資訊(syllable segmentation information)的格(lattice)，最

後，我們以格重新計分器(lattice re-scorer)將格(lattice)上所有的假設(hypotheses)以韻律模型 (prosodic models)和發音方法的信心分數 (a posterior probabilities of attributes) 進行重新計分，求得最佳之詞序列(word sequence)、詞類序列(POS sequence)以及韻律斷點序列(break sequence)。

我們用於英語大詞彙語音辨認的數學式如下所示：

$$\mathbf{L}^*, \mathbf{B}^*, \Upsilon^* \approx \arg\max_{\mathbf{L},\mathbf{B},\Upsilon} \left\{ \underbrace{P(\mathbf{A},\Upsilon|\mathbf{W})^{\alpha_1}}_{\text{AM}} \underbrace{P(\mathbf{W})^{\alpha_2}}_{\text{Trigram LM}} \underbrace{P(\mathbf{POS}|\mathbf{W})^{\alpha_3}}_{\text{POS model}} \underbrace{P(\mathbf{B}|\mathbf{L})^{\alpha_4}}_{\text{Break-syntax model}} \underbrace{P(\mathbf{sd}|\mathbf{B},\mathbf{L},\Upsilon)^{\alpha_5}}_{\text{Syllable duration model}} \underbrace{P(\mathbf{pd}|\mathbf{B},\Upsilon)^{\alpha_6}}_{\text{Pause duration model}} \underbrace{P(At(\mathbf{W})|\mathbf{A},\Upsilon)^{\alpha_7}}_{\text{Attribute porsterior prob.}} \right\}$$

Factored Language Model(FLM)

Prosodic Models

其中 $P(\mathbf{A},\Upsilon|\mathbf{W})$ 、 $P(\mathbf{W})$ 為聲學及語言模型； $P(\mathbf{POS}|\mathbf{W})$ 為詞類模型； $P(\mathbf{B}|\mathbf{L})$ 、 $P(\mathbf{sd}|\mathbf{B},\mathbf{L},\Upsilon)$ 及 $P(\mathbf{pd}|\mathbf{B},\Upsilon)$ 分別為韻律模型中的 1)韻律段點-語法模型、2)音節時長模型、以及 3)靜音時長模型； $P(At(\mathbf{W})|\mathbf{A},\Upsilon)$ 為發音方法模型。由數學是可知因為韻律訊息必須經由語法中的詞類(POS)來傳遞高階語言參數對韻律斷點型態(break type)的影響，所以在格(lattice)上必須展開詞類(POS)和韻律斷點(break type)的假設路徑。另外，$\{\alpha_i \mid i = 1 \sim 7\}$ 為各個子模型的分數權重，可由訓練集合(training set)或發展集合(development set)調整其值使其達到最好之辨認律。

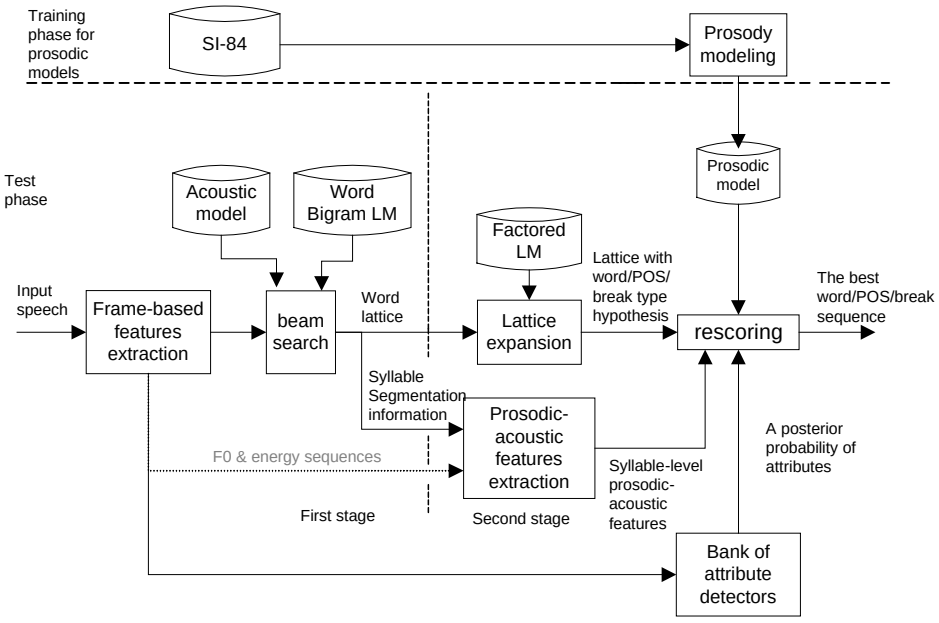

Fig. 2: 以韻律及發音方法訊息協助之英文大詞彙語音辨認架構

(2) 2012 年 5 月 1 日~2012 年 6 月 1 日

進行語音辨認的實驗，其中聲學模型其中 $P(\mathbf{A}, \Upsilon|\mathbf{W})$ 是由 WSJ SI-84 語料庫之語音訓練而成，語言模型 $P(\mathbf{W})$ 是以 WSJ 87'-91'之文字語料訓練而成的詞三連文語言模型(Trigram LM)，韻律模型 $P(\mathbf{B}|\mathbf{L})$、$P(\mathbf{sd}|\mathbf{B},\mathbf{L},\Upsilon)$ 及 $P(\mathbf{pd}|\mathbf{B},\Upsilon)$ 是以 SI-84 中非口語標點符號(Non-Verbal Punctuation)語音訓練而成，發音方法模型 $P(At(\mathbf{W})|\mathbf{A},\Upsilon)$ 也是由 WSJ SI-84 語料庫之語音訓練而成。語音辨認的實驗結果如表一所示，在加入韻律模型和發音方法分數後，的確對於語音辨認有所幫助。

表一：加入韻律(prosody)及發音方法(manner of articulation)訊息的 WSJ0 Nov92 測式語料辨識結果

	Word Corr. (%)	**Word Acc. (%)**	**Sentence Corr. (%)**
Baseline	95.89	94.60	56.97
+attribute	96.17	95.09	59.09
+prosody	96.00	94.90	58.48
+attribute+prosody	96.15	95.14	59.39

另外，我們亦將由英語訓練出來的發音方法(manner of articulation)辨認器，應用於中文大詞彙辨認器，希望能提升辨認率，並驗證語音屬性偵測前級(detection-based ASR frontend)對於不同語言的一般性 (generality) 及跨語言知可移植性 (portability)，Fig. 3 為此研究的工作方塊圖(block diagram)。首先，我們以基礎語音辨認器 (ASR system)以最大相互資訊條件馬可夫聲學模型(HMM-based MMIE AM)及詞三連文語言模型(word trigram LM)產生詞格(word lattice)，此詞格(word lattice)內包含假設音節(hypothesized syllable)的切割位置，接下來我們設計一個音節結構驗證器(syllable structure verifier)，使用跨語言之屬性偵測組(bank of attribute detectors)，對於語音辨認系統(ASR system)產生之詞格(word lattice)進行重新計分，希望能將因為過於相信語言模型(LM)分數而導致辨認錯誤的結果進行更正。值得注意的是，在這個研究裡面所使用的屬性偵測組

(bank of attribute detectors)是由英文語料訓練而成，但是辨認目標是中文語音，若此跨語言之偵測器(detector)能發揮功能，便能證明語音屬性為基礎之語音辨認系統(detection-based ASR)的 一般性(generality) 及可移植性(portability)。接下來，我們使用格展開器(lattice expander)將詞格(word lattice)展開成包含韻律斷點訊息(break information)、韻律狀態(prosodic state)、詞類(POS)、標點符號(punctuation mark)以及詞(word)的格(lattice)，最後我們利用格重新計分器(lattice re-scorer)以韻律模型將延展之詞格(word lattice)進行重新計分，得到最佳之詞序列(word sequence)、詞類序列(POS sequence)、韻律斷點序列(break type sequence)以及韻律狀態序列(prosodic state sequence)。

我們以一個 300 人錄製的中文朗讀語音資料庫作為實驗語料 (TCC300)，其中聲學模型是由 TCC300 語料庫之語音以最大相互訊息(MMI)條件訓練而成，語言模型是以 Sinorama、CIRB030 以及 Sinica Corpus 訓練而成的詞三連文語言模型(Trigram LM)，韻律模型是以 TCC300 之部分語音訓練而成，發音方法模型 $P(At(\mathbf{W})\,|\,\mathbf{A},\Upsilon)$ 是由英語 WSJ SI-84 語料庫之語音訓練而成。

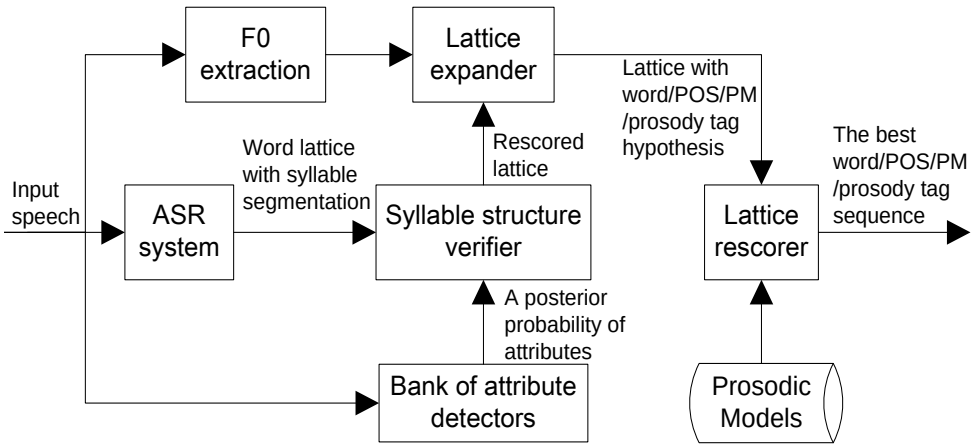

Fig. 3: 以韻律及跨語言發音方法訊息協助之中文大詞彙語音辨認架構

語音辨認的實驗結果如表二所示，可看出同時加入發音方法、韻律斷點以及韻律狀態資訊，可得到最佳之辨認結果，一般來說，只加入韻律斷點資訊可提供最佳的辨認率改善，而加入發音方法相對得到較少的改善，其原因有二：1) 此

發音方法辨認器是由英語語料訓練而得，因此應用在中文上會有限制、2)音節(syllable) 內沒有正確的音(phone)切割位置，所以我們採用動態時間對應(DTW)方法對於音節(syllable)內找到最佳之發音屬性切割，如此會對於音節結構驗證(syllable structure verification)能力會有所影響。

表二：加入發音方法(+M)、韻律斷點(+B)以及韻律狀態(+P)於 TCC300 語音資料庫上實驗結果之詞錯誤率 (WER)、字元錯誤率 (CER)以及音節錯誤率 (SER)。

	WER (%)	CER (%)	SER (%)
Baseline	13.75	10.56	7.79
+M	13.45	10.20	7.44
+B	12.57	9.36	7.13
+M+B	12.43	9.81	6.90
+B+P	12.26	8.93	6.73
+M+B+P	12.24	8.85	6.63

Fig. 4 顯示一個利用音節結構驗證器(syllable structure verifier)改正之辨認結果範例，此範例是由 TCC300 中的 "NCKU_f060602_0" 測試句子中擷取，原本正確的部分句子應該是 "(超級, chao1-ji2, super) (大, da4, large) (縣, xian4, county) (臺北縣, tai2-bei3-xian4, the Taipei County) (其, qi2, its) (縣長, xian4-zhang3, County Magistrate) (寶座, bao3-zuo4, post)"，然而基礎辨認器辨的辨認結果是 "(及, ji2, and) (市長, shi4-zhang3, Major)，當我們加入發音方法的分數時，辨認結果更正為 "(縣長, xian4-zhang3, County Magistrate)"。Fig. 4 顯示音節 'xian4' 的發音之事後機率(a posterior probabilities)，此音節在基礎系統裡會辨認成 'shi4'，請注意音節 'xian4' 的發音方法序列為 摩擦音(fricative)-母音(vowel)-鼻音(nasal) 而'shi4' 的發音方法序列為摩擦音(fricative)-母音(vowel)。由 Fig. 4 可以清楚看到發音方法母音(vowel) 的事後機率(a posterior probabilities)在音節的尾巴比發音方法鼻音(nasal) 的事後機率(a posterior probabilities) 低了許多，因此沒有鼻音(nasal)在尾巴錯誤的音節 'shi4' 可被更正成為 'xian4'。

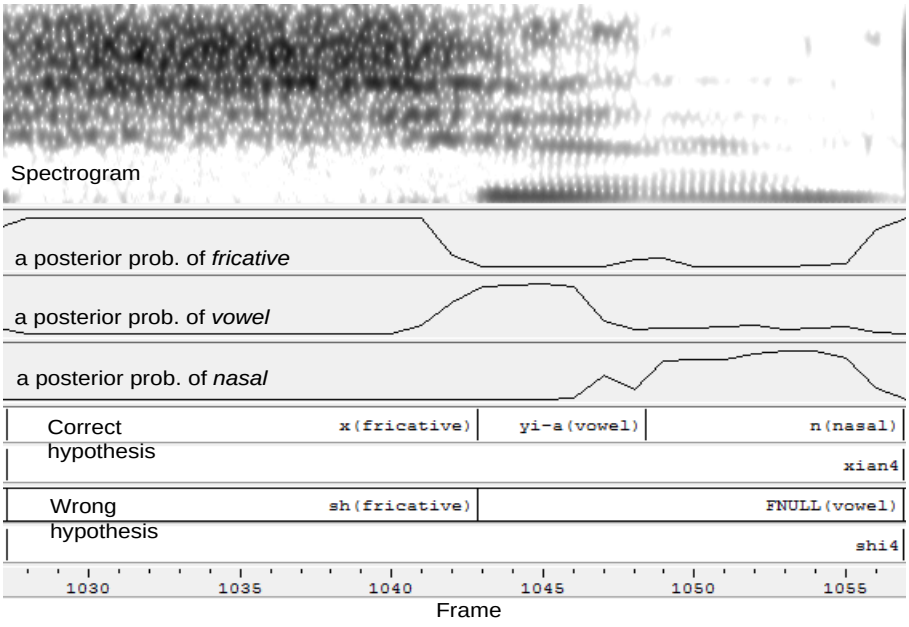

Fig. 4: 在 NCKU_f060602_0 語句中辨認錯誤區域的發音方法分數，包含 spectrogram (上)、發音方法摩擦音(fricative)、母音(vowel)以及鼻音(nasal)的事後機率(posterior probabilities) (中)、以及發音方法、音節的切割位置資訊 (下)。

（三）心得及建議

此次十分感謝教育部及校方能支持這次為期共 5 個月的訪問研究，讓本人與李錦輝老師及 Prof. Sabato Marco Siniscalchi 密切地研究及討論。在討論的過程中，重新檢視過去累積的研究經驗與研究方法，將中英兩種語言的發音及韻律特性做了較完整的比較，由研究結果可以發現，如果能夠研製出通用於各種語言間的語言屬性偵測器，除了對於各個語言有更深入的了解外，更能提升語音辨認技術，目前本研究的技術經驗包含了中、英文韻律模式、英文發音屬性以及英文跨中文之發音屬性，接下來可以延伸此研究，朝向發展為中文特別設計之發音屬性偵測，對於中文語音辨認應可進一步有所助益。另外，在研究的過程中，發現到美國在語音處理的技術能夠累積，很多原因是因為美國重視研究的建設基礎，如語音資料庫的蒐集、整理以及標準化，皆是由多所學校及美國政府 (如國防部) 合作而成，並非少數人力完成，所有的基礎系統皆有公開且標準化的的模型，讓所有進行研究的學者有所依規，讓所有比較公平，因此學術界的經驗可以系統性的累積。反觀台灣於基礎工作上就比較不受重視，一開始可能可以快速的完成特

定的研究，然而久而久之，學術上的累積便因為缺乏穩固的基礎而進展變慢，這的確是令人值得反思的地方。因此，本人在此建議台灣參與語音處理和自然語言處理的相關研究單位，應盡量推行實驗資料庫的標準化，定期更新資料庫修正的結果、建立相關的基礎實驗模型，這樣可以使各大單位的研究成果有所累積、有所延續，使得台灣的研究成果能在國際上更受注意，以語音處理研究為例，相關研究單位應重新整理重要之語音和資料庫，並讓這些資料庫更加標準化，如 TCC300 台灣口語語音資料庫、MCDC 自發性語音資料庫、中研院平衡語料庫、以及中研院語法樹語料庫等等。