基於 Sphinx 可快速個人化行動數字語音辨識系統

Quickly Personalizable Mobile Digit Speech Recognition System Based on Sphinx

顏宗芃 Tsung-Peng Yen　　陳嘉平 Chia-Ping Chen
國立中山大學資訊工程系
Department of Computer Science and Engineering
National Sun Yat-Sen University
m003040029@student.nsysu.edu.tw, cpchen@cse.nsysu.edu.tw

摘要

本論文建立了一個透過網路提供數字語音辨識服務的系統，除了語音辨識功能也提供線上個人化調適功能來克服在不同環境中的噪音强健性。以英文數字辨識來說使只需要經過少許的調適就能夠在少許的時間內打造出正確率達 80% 以上的個人化英文數字語音辨識系統。Sphinx-4 是專門為了研究而開發的工具，具有延展性、模組化、可插拔的架構，因為這些特性我們選擇使用 Sphinx-4 做為語音辨識系統的核心。為了讓選擇聲學模型與訓練語料及調適聲學模型上有一個依據，使用 AURORA2 語料庫訓練模型，台灣口音英語語料庫與 Android 裝置錄製的語料進行調適實驗，結果顯示使用 EAT 語者獨立的語料經 100 句調適後正確率能夠由 80% 進步到約 90% ；多環境模型經 Android 單人語料經 25 句能從平均 70% 提升到約 95% 的正確率。

關鍵詞： 行動化、語音辨識、個人化、調適、噪音强健性、Sphinx

Abstract

In this paper, we introduce a system for on-line digit speech recognition services. Besides the speech recognition service in our system, we also provide adaptation function to improve the noise-robustness between different environment. In the case of English digit recognition, our recognition system can achieve over 80% accuracy for a specific speaker by using a few adaptation data. We use Sphinx-4 as a speech recognition kernel in our system. Because Sphinx-4 is a system prepared exclusively for researchers, it is a flexible, modular and plug-gable framework. We provide our experiment results on AURORA2, EAT and Android device recording. We use AURORA2 database training models that adapt by EAT and Android device recording. The experimental results show we can get high accuracy after a few adaptation.

keywords: mobile, speech recognition, personalizable, adapt, noise-robustness, Sphinx

一、研究背景、動機

拜科技的演進及網路發展所賜，語音辨識成為了生活上日漸重要的角色如 Google voice search [1]、Iphone Siri [2] 及其它相關應用 [3] [4] [5]，衍生了許多可以連上網際網路的科技產品 (如 PDA 、智慧型手機、平版電腦)。這些科技產品都已經成為了現代人的生活必需品，但是大多數都是使用傳統的按鍵來進行操作，想要利用按鍵靈活的操作這些不同的裝置是非常困難的。但如果我們的裝置不侷限於按鍵輸入而使用語音輸入來控制這些裝置，甚至不需要把手機從包包中拿出來就能夠撥出電話與朋友交談。把語音變成隨身攜帶的萬用遙控器能夠大大的改善使用上的便利性，即使是身體有殘缺的人只需要透過口語，也能利用這個系統來操作這些現代科技的手持裝置。

現在大多數的即時語音辨識系統都是建立在網際網路上，在辨識的過程中使用者透過個人電腦或是其它裝置將語音傳至伺服器上，待伺服器辨識完成後將結果回傳，把語音辨識相關等較較耗費資源的工作都交給伺服器運算。這種架構讓使用者不需要使用高效能的裝置就能使用語辨識的服務，像雲端運算服務 [6] 多數都建立於大型的分散式伺服器上。在 [7] 中提到，人類可用語音輸入來控制瀏覽器指標以增進使用者與網頁的互動。

在本論文專注在建立一個能夠兼具服務與研究的語音辨識網路系統，研究不同口音、噪音環境、調適句數對辨識率的影響以提供給使用者在選擇聲學模型、訓練、調適上能夠有一個依據，利用網際網路結合自動語音辨識 (Automatic Speech Recognition, ASR) 系統，釋出一個網路語音辨識系統。

二、系統架構

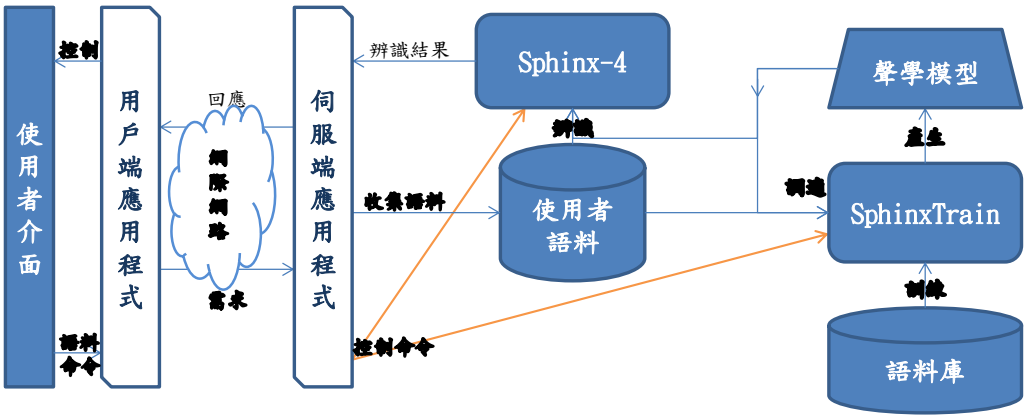

圖 1、 系統架構圖

現在常見到的主流的語音辨識研究大多是以隱藏式馬可夫模型 (Hidden Markov Model, HMM)如 [8] [9] [10]，與高斯混合模型 (Gaussian Mixture Model, GMM) [11] [12] 統計模型的方法建立的，這一類的辨識工具有 HTK [13]、CMU Sphinx 等等，在本篇論文中選擇使用 CMU Sphinx 的 Sphinx-4 [14] 作為核心辨識工具。

主從式架構是一種運用網路技術、開放的架構來降低成本的一種小型化電腦系統，用戶端可能是一台個人電腦或小型工作站，本身就具備完整獨立作業的能力；伺服端則是一台較大型的伺服器或電腦主機，而在用戶端及伺服端之間則藉著可靠的通信協定連結。

本系統以 HTTP (HyperText Transfer Protocol)的方式建立主從式架構，一個伺服端 (server) 透過網路來同時服務多個用戶端 (client)，HTTP 是網際網路應用最為廣泛的一種網路協定，它的好處在於能夠容易的使用網頁伺服器架構出用戶端給瀏覽器使用，而且在其它裝置上也很容易能夠設計出符合條件的用戶端，圖 1 表示了整個系統架構，用戶端與伺服端分別使用不同的應用程式來控制，使用者透過使用者介面 (user interface) 與用戶端應用程式溝通，用戶端應用程式將語料及命令以需求的方式送出至伺服端，伺服端應用程式收到需求後針對所需控制辨識工具做辨識或調適的動作，完成後把將辨識結果或完成訊息回應給用戶端應用程式以操控使用者介面。

三、實驗

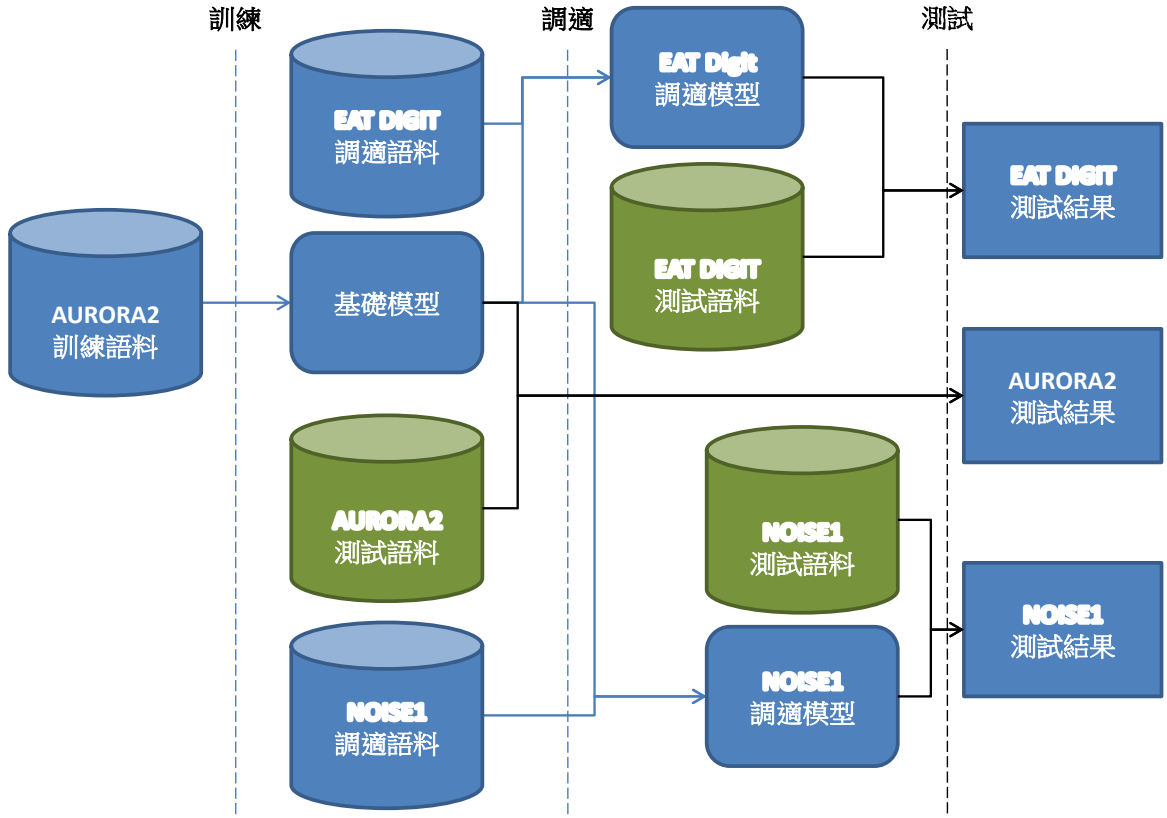

圖 2、 實驗流程

本論文中的實驗一共用到了 AURORA2、EAT DIGIT、NOISE1 三個語料庫，圖 2 簡明的表示了整個實驗的流程，第一步是使用 AURORA2 的語料進行訓練作為調適的基礎模型，再分別使用 EAT DIGIT 及 NOISE1 語料來調適進行不同的腔調、噪音、裝置實驗。

(一)、語料庫介紹

本篇論文使用 AURORA2 [15] 產生基礎聲學模型 (baseline model)，台灣口音英語語料 (English Across Taiwan, EAT) [16] 英文數字部分 (簡稱 EAT DIGIT) 及自行錄製的 NOISE1 語料做為調適語料進行一系列調適的實驗。

AURORA2 ：使用不同加成性噪音、訊噪比來測試語音辨識系統强健性的語料庫。

EAT DIGIT ：從 EAT 中的可用語料中過濾出純英文數用的部分，句數如表 1。

表 1、 台灣口音英語語料庫可用純英文數字語料句數

環境＼分類	非英語系學生			英語系學生			
	女	男	加總	女	男	加總	總和
gsm	212	334	546	386	156	542	1088
pstn	168	171	439	341	136	477	916
mic16k	421	697	1118	794	318	1112	2230

NOISE1 ：從 AURORA2 語料庫中選擇 NOISE1 的文本 (corpus) 錄製(故稱 NOISE1 語料庫)，使用裝置 DHD (HTC Desire HD)及 WFS(Wildfire S)錄製 8KHz 16bits PCM 格式語音檔案，整個語料庫如表 2 所示。

表 2、 NOISE1 語料庫錄製環境有關閉所有家電與門窗的安靜宿舍 中(dormitory) 、安靜宿舍於開啓的電風扇旁 (fan) 、下班時段西子灣捷運二號出口旁的公車等候亭 (road) 、中山大學下午五點的籃球場 (basketball) 、中山大學中午 11 點半 L 型停車場 (parkingLot) 及中山大學電資大樓 F5017b 實驗室 (laboratory)。

(a) 各環境與裝置句數

分類	環境	語者	使用裝置	
			DHD	WFS
clean	dormitory	tpyen	1001	X
noise	fan	tpyen	50	X
	road	tpyen	50	X
	basketball	tpyen	50	X
	parkingLot	tpyen	50	X
f5017b	laboratory	如表 2(b)	200	200

(b) f5017b 環境語料各語者與裝置句數

語者	性別	使用裝置	
		DHD	WFS
jcdeng	女	50	50
mkwu	男	50	50
tpyen	男	50	50
yhhuang	男	50	50

(二)、實驗設定

本實驗所使用的特徵參數為梅爾倒頻譜係數 (Mel-scale Frequency Cepstral Coefficients, MFCC) ，如表 3 所示，在擷取所有音檔特徵參數除了於 EAT 中 mic16k 所使用的取樣頻率為 16000 外其餘所使用的參數都是一致的。後端的聲學模型上使用高斯混合模型來訓練，所使用的字典 (dictionary) 與音素 (phone) 列於表 4中，隱藏式馬可夫模型每一個音素一個模型、每一個模型 3 個狀態、每個狀態包含 8 個高斯混合分佈 (Gaussian mixture distribution) 訓練上下文相關 (context dependent) 的模型。

表 3、 擷取特徵參數所使用的參數檔案

參數	說明	設定值
alpha	預强調參數	0.97
dither	增加 1/2-bit 雜訊避免零能量音框	yes
ncep	倒譜係數	13
lowerf	下截止頻率	64
upperf	上截止頻率	4000
nfft	快速傅利葉轉換大小	512
wlen	漢明窗長度	0.025
input_endian	輸入資料的位元組序，在 NIST 與 MS Wav 格式中忽略	big
samprate	取樣頻率	8000
feat	Sphinx 參數格式	1s_c_d_dd

分別對 AURORA2 的乾淨訓練語料與多環境訓練語料使用 SphinxTrain 得到乾淨語料 (clean) 及多環境語料 (multi) 兩個聲學模型，使用 AURORA2 的測試語料所得到的辨識率如表 5 所示，本論文使用字模型 (word dependant model) 作為基礎模型，把相同發音的音素模型共用所得到的結果也非常相近於現在的結果。實驗中利用 EAT DIGIT 與 NOISE1 語料做一系列的調適實驗，在這些調適實驗中所用到的調適法為最大後驗 (Maximum A Posteriori, MAP) [17] 方法來進行。

(三)、AURORA2 聲學模型使用 EAT DIGIT 語料庫語料調適實驗

為了解外國語言腔調在受過訓練後與未受過訓練的差異進行英語系與非英語系學生的腔調比較、調適效果與句數關係、跨環境調適效果三項實驗。EAT DIGIT 一共有 gsm、pstn、mic16k 三種錄音方式，再細分為 gsm 英語系學生的語料 (gsmE) 、gsm 非英語系學生的語料 (gsmN)、pstnE、pstnN、mic16kE、mic16kN，最後把這六種條件的語料分成測試語料以及調適語料，如表 6 所示。把每一種條件的語料再各分成兩半，一半做測試語料一半做調適語料。分別在後面加上 _t 與 _a 代表測試語料及調適語料，將每一種環境的語料一共分成四份 (E_ test、E_adapt、NE_test、NE_adapt) 。

1、EAT DIGIT 英語系與非英語系腔調比較

利用各環境 E_a 及 N_a 的部分調適成英語系模型及非英語系模型，再使用 E_t 與 N_t 的

表 4、 字典與音素

字典	
文字	音素
eight	EY_eight, T_eight
five	F_five, AY_five, V_five
four	F_four, OW_four, R_four
nine	N_nine, AY_nine, N_nine_2
oh	OW_oh
one	W_one, AX_one, N_one
seven	S_seven, EH_seven, V_seven, E_seven, N_seven
six	S_six, I_six, K_six, S_six_2
three	TH_three, R_three, II_three
two	T_two, OO_two
zero	Z_zero, II_zero, R_zero, OW_zero
filler	
<s>	SIL
<sil>	
</s>	

表 5、 乾淨語料與多環境語料模型辨識率(Avg.：0-20db 的平均值)

	乾淨語料模型				多環境語料模型			
dB ＼測試集	A	B	C	Avg.	A	B	C	Avg.
clean	99.5	99.5	99.3	99.5	99.0	99.0	98.8	99.0
20	95.3	96.8	95.4	95.9	98.3	98.4	97.7	98.2
15	87.1	90.2	87.1	88.3	97.6	97.5	97.0	97.4
10	63.6	71.3	64.3	66.8	95.0	95.0	94.3	94.9
5	24.6	31.6	26.7	27.8	84.5	84.5	84.4	84.5
0	2.1	4.8	3.5	3.5	47.8	48.4	50.5	48.6
-5	0.4	1.0	0.6	0.7	8.2	11.7	10.7	10.1
Avg.	54.5	58.9	55.4	56.5	84.6	84.8	84.8	84.7

表 6、 將 EAT 六種條件的語料進一步分成測試語料及調適語料

	測試語料			調適語料			
	女	男	總句數	女	男	總句數	總計
gsmN	106	167	273	106	167	273	546
gsmE	193	78	271	193	78	271	542
pstnN	84	135	219	84	136	220	439
pstnE	170	68	238	171	68	239	477
mic16kN	210	348	558	211	349	560	1118
mic16kE	397	159	556	397	159	556	1112

語料測試，得到的英語系與非英語系腔調差異做比較如表 7：

由這些數據中發現不論是使用英語系或非英語系的語料來調適，辨識率都是英語系語料優於非英語系語料。 AURORA2 的錄製語者都是以英文為母語，因此擁有較標準的口音得到較高的辨識率是可預期的現象。在以英語系語料為測試語料的條件下，不論是使用英語系或非英語系語料都能得到良好的調適效果；相對於使用非英語系語料測試時使用英語系語料調適的效果就沒那麼優異。

這個實驗的結果表示出口音對辨識率的影響很大，在訓練過與未訓練過辨識率的差距可以多達 10%，而且即使是訓練過的口音或多或少還是會受到母語口音的影響，在選擇調適語料的時後以挑選與使用者母語相同的語料為佳。

2、EAT DIGIT 調適效果與句數關係

在調適效果與句數關係的實驗中，使用不同的句數來觀察各個條件下使用不同句數調適的辨識率。在這邊調適句數單位 1 是代表一句男生語料加上一句女生語料 (5 單位就代表 5 句男生語料加 5 句女生語料，以此類推)，如單一性別語料不足則使用另一種性別的語料補足。對三種環境做句數調適效果的測試，我們讓三種環境中的英語系及非英語系的語料輪流當測試語料及訓練語料，結果如表 8 9。

由實驗結果中能夠看出調適句數與正確率成長起初正確率略微下降、後來快速成長、到最後逐漸趨緩的整個過程趨勢，由圖中也可得知在使用語者無關 (context independent) 調適語料時使用約 50 單位 (男女各 50 句) 的語料可以達到最佳的效果，而使用超過 50 單位後識率的成長便逐漸趨緩，如果我們用相同的方法直接使用 50 單位的調適語料來訓練聲模型只能得到約 60% 的正確率。

3、EAT DIGIT 跨環境調適效果

在這個實驗中我們將每一種條件的測試語料分別對六種條件的調適模型來測試在錄音裝置與錄製格式不同的條件下的差異性，所得到的結果如表 10 11所示，gsm 與 pstn 語料都是藉由電話話筒接收聲音，所錄得的 8KHz 8Bits Mulaw 格式的取樣點，經程式轉成8khz 16bits PCM 格式的取樣點，麥克風語料則是由個人電腦及麥克風經由音效卡錄製 16KHz 16bits 的聲音訊號。結果顯示這些條件下的環境是非常接近的，不論是使用

表 7、 英語系與非英語系腔調比較

乾淨語料模型			
測試語料＼調適語料	乾淨語料模型(無調適)	gsmE_a	gsmNE_a
gsmE_test	85.6	93.0	92.5
gsmN_test	73.5	86.4	89.5
測試語料＼調適語料	乾淨語料模型(無調適)	pstnE_a	pstnN_a
pstnE_test	85.2	92.5	92.7
pstnN_test	77.4	90.5	90.5
測試語料＼調適語料	乾淨語料模型(無調適)	mic16kE_a	mic16kN_a
mic16kE_t	87.1	91.5	92.7
mic16kN_t	75.5	87.6	91.2

多環境語料模型			
測試語料＼調適語料	多環境語料模型(無調適)	gsmE_a	gsmN_a
gsmE_t	85.8	92.2	91.6
gsmN_t	75.1	86.5	88.1
測試語料＼調適語料	多環境語料模型(無調適)	pstnE_a	pstnN_a
pstnE_t	84.1	92.3	91.6
pstnN_t	78.1	88.1	88.3
測試語料＼調適語料	多環境語料模型(無調適)	mic16kE_a	mic16kN_a
mic16kE_t	85.1	92.0	91.6
mic16kN_t	74.4	87.4	89.5

表 8、 AURORA2 乾淨語料模型分別以 EAT 六種條件做調適的句數與正確率

測試語料	調適語料	調適前	1	5	10	25	50	75	100	全部
gsmN_t	gsmE_a	73.5	72.1	80.1	83.8	85.5	87.2	88.4	89.0	89.5
gsmE_t	gsmE_a	85.6	83.6	87.4	89.1	89.1	91.0	91.7	92.6	93.0
pstnN_t	pstnE_a	77.4	76.9	84.5	86.3	87.2	90.2	90.5	90.9	90.5
pstnE_t	pstnE_a	85.2	84.1	87.5	88.4	90.9	92.4	91.9	92.6	92.5
mic16kN_t	mic16kN_a	75.5	77.9	82.4	83.4	85.0	87.1	88.6	89.4	91.2
mic16kE_t	mic16kE_a	87.1	86.4	87.3	88.6	89.0	91.2	92.2	92.3	91.5

表 9、 AURORA2 多環境語料模型分別以 EAT 六種條件做調適的句數與正確率

測試語料	調適語料	調適前	1	5	10	25	50	75	100	全部
gsmN_t	gsmN_a	75.1	75.1	79.9	81.6	84.7	85.7	86.3	86.7	88.1
gsmE_t	gsmE_a	85.8	84.3	86.8	88.7	89.2	91.2	91.4	91.8	92.2
pstnN_t	pstnN_a	78.1	77.6	84.0	86.4	86.0	88.7	88.1	88.5	88.3
pstnE_t	pstnE_a	84.1	81.5	87.2	89.3	89.7	90.8	91.5	91.8	92.3
mic16kN_t	mic16kN_a	74.4	75.9	79.8	81.2	84.1	84.7	86.4	87.0	89.5
mic16kE_t	mic16kE_a	85.1	84.2	86.7	88.0	88.8	90.1	90.8	90.9	92.0

電話直接錄音還是透過音效卡使用麥克風在個人電腦上錄音，在沒有其它特別噪音的情況下使用不同的取樣頻率在辨識率的差異並不大。

表 10、 AURORA2 乾淨語料模型分別使用 EAT 六種條件語料調適的辨識率

測試語料＼調適語料	gsmE_a	gsmN_a	pstnE_a	pstnN_a	mic16kE_a	mic16kN_a	Avg.
gsmE_t	93.0	92.5	92.3	91.4	92.2	91.3	92.1
gsmN_t	86.4	89.5	88.2	87.8	87.1	89.0	88.0
pstnE_t	92.1	92.1	92.5	92.7	92.9	92.7	92.5
pstnN_t	88.3	89.8	90.5	90.5	88.7	91.0	89.8
mic16kE_t	89.5	90.0	90.8	90.8	91.5	92.7	90.9
mic16kN_t	83.9	88.0	86.5	87.7	87.6	91.2	87.5
Avg.	88.9	90.3	90.1	90.2	90.0	91.3	90.1

(四)、AURORA2 聲學模型使用 NOISE1 語料調適實驗

將網路辨識系統運用在現流行的 Android 手機上面，撰寫了一個符合本論文中所提出的語音辨識系統的用戶端程式，進行一系列的辨識與調適實驗，這些實驗主要在測試手持行動裝置上使用本篇論文中的語音辨識系統的效能及實用性。

1、NOISE1 調適效果與句數

首先將 NOISE1 中的 clean 語料分兩個部分，分別為測試語料 (前500句) 及調適語料 (後501句) ，使用的調適語料由少到多，調適單位 1 代表一句調適語料，其實驗結果如表 12 所示：

由 Android 裝置所錄製的語料不論在乾淨語料模型或是多環境語料模型在未調適的情況下與誇環境實驗得到相近的結果，調適過程中所使用的都是使用同一個人的語料來進行，在句數相同的情況下明顯地勝過先前使用不同語者語料所調適的模型，另外由此表中能觀察到約在 25 到 50 句時調適效果逐漸趨緩，因此假設以 AURORA2 的模

表 11、 AURORA2 多環境語料模型分別使用 EAT 六種條件語料調適的辨識率

測試語料＼調適語料	gsmE_a	gsmN_a	pstnE_a	pstnN_a	mic16kE_a	mic16kN_a	Avg.
gsmE_t	92.2	91.6	92.0	91.3	92.7	91.6	91.9
gsmN_t	86.5	88.1	87.4	88.8	86.2	87.3	87.4
pstnE_t	90.9	89.6	92.3	91.6	92.4	91.2	91.3
pstnN_t	86.4	86.4	88.1	88.3	88.6	88.4	87.7
mic16kE_t	89.1	88.9	91.0	90.0	92.0	91.6	90.4
mic16kN_t	83.3	86.4	86.3	87.3	87.4	89.5	86.7
Avg.	88.1	88.5	89.5	89.6	89.9	89.9	89.2

表 12、 AURORA2 模型以 NOISE1 clean 語料調適句數與正確率

調適模型	調適前	2	5	10	20	25	50	100	150	200	501
乾淨語料模型	80.0	81.3	83.6	90.8	93.5	95.1	95.4	97.8	98.3	98.4	98.8
多環境語料模型	82.5	84.0	88.2	91.1	93.2	95.0	95.1	96.2	96.2	97.3	98.9

型使用 NOISE1 語料調適 25 句能得到最大的投資報酬率的結果來進行 NOISE1 之後的調適實驗。

2、NOISE1 不同噪音環境的調適效果

在前面 NOISE1 與 EAT DIGIT 調適實驗中使用乾淨語料模型與多環境語料模型的實驗結果沒有什麼差別，造成這個現象的主因就是因為所使用的測試語料幾乎都是沒有噪音的語料，而手持式裝置最方便的一處就是走到哪就能帶到哪，不論是要坐車、運動、郊遊或是參加一些其它的社交活動這些裝置幾乎是寸步不離身，但這些環境中並不會每一個地方都能跟 NOISE1 clean 的環境一樣幾乎沒有噪音，可以說是每一個環境中都難免會有一些噪音，嚴重的話甚至聽不清楚語者所說的話。撇開這些無噪音或噪音極大的極端情況找尋生活上常常會遇到的幾種噪音來進行實驗，一共選擇了 basketball、road、fan、parkingLot 四個環境噪音，每一種噪音環境下含有 50 句均使用前 25 句為測試資料後 25 句為調適語料進行調適實驗，其實驗結果如表 13 所示：

在這四種環境中只有 road 是屬於被較强的噪音所污染，其餘三種環境都是屬於輕微的噪音干擾可以從乾淨語料模型的辨識率中明顯的分辨出來，即使是在未針對新的環境來進行調適的情況下多環境語料模型仍然顯現了他在噪音環境下擁有較好辨識率的優勢。

為了進一步了解在噪音環境之下需要多少調適語料才能讓達到一般能接受的正確率進一步對這些語料進行句數與正確率的實驗，其結果如表 14 所示，就平均情況來而言針對環境進行調適 5 句之後能夠得到 80% 左右的辨識率，進行完 25 句調適之後就能得到約 90% 的正確辨識率。這張表格顯示使用乾淨語料模型噪音環境的情況下調適過程反覆不斷的上升下降，造成這種情形應該是因為有些調適語料噪音較大而有些則較

表 13、 AURORA2 模型以 NOISE1 noise 語料測試在不同噪音環境與調適後的正確率

噪音環境＼聲學模型	clean	clean_adapt	multi	multi_adapt
basketball	65.6	94.6	76.3	96.8
road	43.0	72.0	64.5	91.4
fan	62.4	88.2	76.3	97.9
parkingLot	65.6	95.7	66.7	100.0
Avg.	59.2	87.1	71.0	96.5

小，在較小噪音調適下能夠正常的對語者的腔調口音及環境調適，在較大的噪音下就會完全被噪音影響讓轉移機率產生較大幅的變動，但使用多環境語料模型的這種情況較不明顯，這証明了乾淨語料模型在並不適合在少量且噪音大的情況下進行調適。

表 14、 AURORA2 模型以 NOISE1 noise 語料測試在不同噪音環境與調適後的正確率與調適句數關係

模型	clean					multi				
句數	basketball	road	fan	parkingLot	Avg.	basketball	road	fan	parkingLot	Avg.
0	65.6	43.0	62.4	65.6	59.2	76.3	64.5	76.3	66.7	71.0
...										
4	82.8	59.1	78.5	74.2	73.7	81.7	75.3	81.7	74.2	78.2
5	87.1	63.4	80.6	83.9	78.8	82.8	76.3	88.2	82.8	82.5
6	88.2	65.6	81.7	83.9	79.9	84.9	77.4	87.1	82.8	83.1
...										
23	92.5	79.6	92.5	92.5	89.3	95.7	88.2	93.5	93.5	92.7
24	92.5	79.6	92.5	92.5	89.3	95.7	88.2	93.5	94.6	93.0
25	94.6	72.0	88.2	95.7	87.6	96.8	91.4	97.9	100.0	96.5

3、NOISE1 不同裝置的調適效果比較

除了環境噪音對辨識率的影響以外還要考慮到的就是裝置上的差異性，畢竟每個裝置上的麥克風品質不盡相同。造成辨識率差異的不僅僅只會有麥克風，現在有一些裝置還會自動將輸入音源做降噪處理，功能非常人性化也非常的好用，但礙於手邊沒有這麼多裝置可以做辨識率測試的實驗，我們只取得 DHD 及 WFS 兩個裝置來進行實驗，使用 NOISE1 中分類為 f5017b 的語料每一個語者在相同裝置之下均使用前 25 句為測試資料後 25 句為調適語料其實驗結果如表 15 所示。從表中我們不僅能觀察到英文發音造成的差異也能看到裝置不同所帶來的影響，在四位語者中以 yhhuang 英文發音最為標準，所實驗出來的辨識率果然也是最好的。而不論是使用乾淨語料模型或多環境

語料模型以 DHD 裝置的語料在調適前的辨識率明顯低於 WFS 裝置，即使如此在經過 25 句的環境調適以後就能達到平均 95% 以上的辨識率，藉由這個實驗我們可以了解到使用現成的聲學模型於腔調、使用裝置不同的情況也不需要經過大量的調適就能達到良好的辨識率。

表 15、 AURORA2 模型以 NOISE1 f5017b 使用不同裝置語料調適句數與正確率

DHD 裝置				
語者＼聲學模型	clean	clean_adapt	multi	multi_adapt
cjdeng	54.8	97.9	54.8	97.9
mkwu	51.6	95.7	50.5	96.8
tpyen	53.8	96.8	54.8	97.6
yhhuang	63.4	100.0	71.0	100.0
Avg.	55.9	97.6	57.8	98.1
WFS 裝置				
語者＼聲學模型	clean	clean_adapt	multi	multi_adapt
cjdeng	80.6	88.2	75.3	95.7
mkwu	78.5	98.9	80.6	100.0
tpyen	79.6	100.0	83.9	98.9
yhhuang	86.0	100.0	93.5	100.0
Avg.	81.2	96.8	83.3	98.7

五、結論與未來展望

本論文利用現有的語音辨識工具 Sphinx-4 整合出一個網路語音辨識服務系統，這個系統透過網路提供了英文數字語音辨識的服務並支援快速個人化功能，可以在不同環境中快速的達到理想的辨識率，系統內所使用的核心辨識核心 Sphinx-4 是由 JAVA 語言編寫而成的，擁有極具延展性、模組化、可插拔的架構並且有良好跨平台能力的優點，本身也提供了許多的應用程式介面，可以追蹤解碼器、運行速度、記憶體使用量等等，非常適合用於研究。因為 Sphinx-4 的特性使伺服端可以在任何支援 JAVA 的作業系統上運行，而用戶端可以是電腦、手機或其它可上網的裝置。

此系統透過網路提供即時的語音辨識，並且可以將使用者及研究人員將使用期間所辨識過的語料收集起來，使用上非常容易且方便，再透各種語料的調適實驗讓使用者在挑選語言模型、訓練語料及測試語料時有個依據。對於這個平台我們跨出的第一步是將這個系統整合出來，提供原始碼讓任何有興趣的人使用。

這個系統擁有網路語音辨識、調適及語料收集的功能，並能夠在使用的過程中將語料收集至伺服端。透過網路語音辨識的功能若能加上其它的技術就能衍生新的應用。如加入人工智慧應用在智慧型手機上，就能展現出更完善的功能。而在語料收集這個區塊目前只是單純的把音檔儲存在伺服端，沒有執行分類或是過濾的動作，其它功能

也還尚有不足的部分。例如可以利用可插拔的特性加入對傳輸檔案進行編碼壓縮來節省網路頻寬、線上即時更換聲學模型解決不同語言問題、對聲學模型調適克服不同使用環境等等功能。針對上述幾點情況進行擴充，這個系統就能夠吸引更多人使用，以促進語音辨識相關應用研究的發展。

參考文獻

[1] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, ""Your word is my command": Google search by voice: A case study," in *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*, 2010, ch. 4, pp. 61–90.

[2] I. VanDuyn, "Comparison of voice search applications on ios," http://www.isaacvanduyn.com/downloads/research-proposal.pdf, [Online]. Available.

[3] M. Kamvar and S. Baluja, "A large scale study of wireless search behavior: Google mobile search," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '06. New York, NY, USA: ACM, 2006, pp. 701–709.

[4] T. X. He and J.-J. Liou, "Cyberon voice commander 多國語言語音命令系統 (Cyberon Voice Commander - a Multilingual Voice Command System) [In Chinese]," in *ROCLING*, 2007.

[5] Y. Lu, L. Liu, S. Chen, and Q. Huang, "Voice based control for humanoid teleoperation," in *Intelligent System Design and Engineering Application (ISDEA), 2010 International Conference on*, vol. 2, 2010, pp. 814–818.

[6] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the clouds: A berkeley view of cloud computing," EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009-28, Feb 2009.

[7] J. Borges, J. Jimenez, and N. Rodriquez, "Speech browsing the world wide web," in *Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on*, vol. 4, 1999, pp. 80–86 vol.4.

[8] L. Rabiner and B.-H. Juang, "An introduction to hidden markov models," *ASSP Magazine, IEEE*, vol. 3, no. 1, pp. 4–16, 1986.

[9] Y. Zhao and B.-H. Juang, "Stranded gaussian mixture hidden markov models for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4301–4304.

[10] ——, "Exploiting sparsity in stranded hidden markov models for automatic speech recognition," in *Signals, Systems and Computers (ASILOMAR), 2012 Conference Record of the Forty Sixth Asilomar Conference on*, 2012, pp. 1623–1625.

[11] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, D. Povey, A. Rastrow, R. Rose, and S. Thomas, "Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 4334–4337.

[12] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, A. Rastrow, R. Rose, P. Schwarz, and S. Thomas, "Subspace gaussian mixture models for speech recognition," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 4330–4333.

[13] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*. Cambridge University Press, 2006.

[14] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: a flexible open source framework for speech recognition," Mountain View, CA, USA, Tech. Rep., 2004.

[15] D. Pearce, H. günter Hirsch, and E. E. D. Gmbh, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *in ISCA ITRW ASR2000*, 2000, pp. 29–32.

[16] 中華民國計算語言學學會, "台灣口音英語語料庫說明 English Across Taiwan (EAT)," http://www.aclclp.org.tw/doc/eat_brief.pdf, [Online]. Available.

[17] C.-H. Lee and J.-L. Gauvain, "Speaker adaptation based on map estimation of hmm parameters," in *Proceedings of the 1993 IEEE international conference on Acoustics, speech, and signal processing: speech processing - Volume II*, ser. ICASSP'93. Washington, DC, USA: IEEE Computer Society, 1993, pp. 558–561.