室內長距離語音辨識技術挑戰與初探

Challenges and Preliminary Study on Indoor Distant Speech Recognition

廖憲正　郭志忠　林政賢
Hsien-Cheng Liao, Chih-Chung Kuo, Cheng-Hsien Lin

中文摘要

長距離語音辨識受到收音裝置、室內空間響應、語者說話位置與方位、以及環境噪音等因素影響，本文針對各個因素進行解析，並嘗試提出解決之方法以及進行初步的實驗。實驗結果顯示長距離影響了如鼻音與塞擦音等子音語音訊號，使得該類型語音之辨識與驗證正確率大幅下降。在加入長距離語音語料進行調適後，可提升語音辨識正確率約10%。而以深度神經網路為基礎之語音模型在加入長距離語料後，更可以得到約60%的音節辨識正確增加率。

Abstract

Distant speech recognition accuracy is highly correlated with types of recording devices, room acoustics, speakers' location and orientation, and environmental noises. This article analyzed causes which decrease distant speech recognition accuracy and tried to propose possible solutions with preliminary experiments. The results showed that the recognition and verification accuracy of consonants, like nasal and affricate, decreased significantly as distance increased. After model adaptation using our distant speech corpus, the recognition accuracy was improved by 10%. There was even about 60% accuracy improvement rate when we used deep neural network as acoustic models trained with the distant speech corpus.

關鍵詞(Key Words)

自動語音辨識(Automatic Speech Recognition；ASR)
深度神經網路(Deep Neural Network；DNN)
語音人機介面(Voice User Interface；VUI)

1．前言

當語音辨識已可正確地顯示你所說的辨識詞彙時，人們便期待這樣的自然人機介面（Natural User Interface; NUI）技術可以使用在其他地方，如智慧家庭、車用導航與多媒體、機器人互動等。而在這些尚未被滿足的需求中，智慧家庭因資通訊技術之發展，在硬體上已有很大的進步，然而在軟體上，仍然維持著舊有選單式的操控方式，這對成員中有老人或是小孩的家庭造成使用上極度的不方便。如能將語音人機介面（Voice User Interface; VUI）這樣的自然人機介面帶入到智慧家庭中，將會大幅提升操作上的便利性。

然而，語音辨識要在智慧家庭這樣的室內空間中使用具有相當大的挑戰性。因為每個家庭中空間的大小相異、擺設不同，這些因素都造成了不同的空間響應進而影響人們所發出的語音訊號。再加上隨著時間會有不同的聲音干擾，如電視、音樂、其他人交談聲等，都使得語音辨識正確率大幅地降低。為了解決這個問

題，我們將從影響長距離語音辨識正確率的因素開始探討，並嘗試提出解決之方法以及進行初步的實驗。

本文章節組織如下，第2節針對長距離語音辨識問題進行解析，分析各個影響長距離語音辨識正確率之因素。第3節則回顧了過去解決長距離語音辨識主要的幾個方法。第4-8節則開始針對長距離語音辨識進行初步之實驗，包括收音裝置之選擇、結合長距離語音語料之語音模型調適、距離對詞語驗證的相對鑑別能力之影響以及深度神經網路在長距離語音辨識正確率上之提升。

2 . 問題解析

室內長距離語音辨識的問題如何定義？或許可按難易程度分成三個層次來看：

- 免手持(hands-free)語音辨識
  不須嘴巴靠近傳統手持或頭戴式麥克風，可對著一定距離外的麥克風講話。典型的例子是車用語音辨識。
- 自由距離(distance-free)語音辨識
  不須嘴巴靠近麥克風的近距離，也不須特定距離。所謂的遠距(far field)並沒有一定的距離規格，但距離越遠音量越小、辨識效果自然越差。可能的應用如機器人。
- 自由角度(orientation-free)語音辨識
  使用者可完全忽略麥克風的位置，可從任何方位、任何角度來下語音指令。這是典型科幻片中的終極理想。

由以上層次分類可知，主要因素都是有關說話人相對於麥克風的空間與心理關係。因此接下來先分析幾個麥克風的相關重要概念。

2.1 麥克風規格

麥克風是將聲波訊號轉換成電波訊號的一種換能器(transducer)。聲波與電波都是一種物理場(field)，也就是都是一種以時空為變數的物理量，因此兩者有類比性，請見表1。

表 1　聲波與電波物理量比較表

聲場	聲壓(p)	粒子速(u)	聲強(I)
	sound pressure	particle velocity	sound intensity
單位	$Pa=N/m^2$	m/s	W/m^2
電場	電壓(v)	電流(i)	功率(P)

雖然定義上功率是電壓與電流相乘，但隨著時間變化的電壓與電流之間可能存在著相位差，且不像直流電有一常數值，所以實用上為求方便一般都會採用均方根rms (root mean square)值代表電壓與電流，以平均功率值代表功率，這使得直流電的簡單數學關係仍然可以成立。類比於此，聲場的物理量在實用上也採取類似做法。另外，在實際度量上，會採用跟一個參考值的比值取對數，也就是類似電波訊號的分貝(dB)的表示法，如表2所示（註：SPL與SVL都是rms值）：

表 2　聲場物理量的分貝(dB)單位量度

聲場量度	聲壓度	聲速度	聲強度
	SPL (sound pressure Level)	SVL (sound velocity Level)	SIL (sound intensity Level)
量度公式	$20*Log_{10}(p/p_0)$	$20*Log_{10}(u/u_0)$	$10*Log_{10}(u/u_0)$
參考值	$p_0=2*10^{-5}$ N/m^2	$u_0=5*10^{-8}$ m/s	$I_0=1*10^{-12}$ W/m^2

雖然有關麥克風的規格有一個國際標準IEC60268-4[1]，但實際上麥克風製造商大多並未遵循，因此市售各種麥克風的規格相當混亂。為此英國有一個網站（原本是書與光碟）蒐集了幾乎所有專業麥克風產品，並盡量將所有規格整理一致[2]。以下只介紹與本文最相關的幾個基本麥克風規格。

2.1.1 麥克風靈敏度(sensitivity)

麥克風是將聲能轉換成電能的換能器，所以最基本的規格就是其轉換增益(gain)，一般稱為麥克風的**靈敏度**(sensitivity)。麥克風感測的

物理量是聲壓，也就是聲波造成空氣介質震動所形成的氣壓變化，其基本單位稱為帕(Pa: Pascal)，也就是牛頓每平方米(N/m^2)。因此，靈敏度就是每單位聲壓所轉換的電壓；更精確一點說，就是以1 kHz聲波輸入每1 Pa (rms)聲壓所轉換的輸出電壓有多少mV (rms)[3]。這裡的輸出電壓一般都指無輸出負載(unloaded)的開路(open circuit)電壓。

若以工程慣用的單位，則1帕聲壓相當於94分貝聲壓度，也就是1 Pa = 94 dB SPL，因為$20*Log_{10}(1/(2*10^{-5}))\fallingdotseq 94$。類似的情況，麥克風輸出電壓也常用表3所示的兩種分貝單位：

表 3　麥克風輸出電壓常用分貝(dB)單位

電壓量度單位	dBu	dBV
量度公式	$20*Log_{10}(v/v_0)$	
參考電壓v_0值	775 mV	1.0 V
假設負載R值	600 Ω	1000 Ω

這兩個參考電壓值，都是來自於定義當參考輸出功率P_0=1 mW時，在假設負載R上的輸出電壓，亦即$v_0=\sqrt{R\times P_0}=\sqrt{R\times 10^{-3}}$。但作為一個標準電壓單位，只要計算輸出電壓與參考電壓值的比，就不必管實際上的負載值與功率值了。

2.1.2 麥克風動態範圍

麥克風的動態範圍是由可以正常轉換能量的最大值與最小值來決定。所謂正常轉換的最大值，係指在不超過一轉換失真度前提下之最大可輸入聲壓，稱為最大聲壓度(Maximum SPL)[4]；而最小值就是在零聲壓、也就是沒有任何聲壓輸入下，因著麥克風裝置本身所產生的自身雜訊(self-noise)電壓輸出[5]。

一個高品質的麥克風，其最大聲壓度通常設定在不產生高於0.5%THD (Total Harmonic Distortion)失真度的條件下。失真容忍度越高，其最大聲壓度就越大。例如若有一麥克風的最大聲壓度是在1%THD的條件下，則可猜測其在0.5%THD下的最大聲壓度，就可能是要比原規格再減少6dB[4]。

通常自身雜訊的規格是以能產生與自身雜訊相同電壓輸出的等效聲壓度來表示，這樣的聲壓度代表一種雜訊地板值(noise floor)的概念，也就是只要是低於此聲壓度的聲音訊號，在經過此麥克風轉換後的電壓，將會埋沒在麥克風自身雜訊所形成的雜訊地板值下。

所以麥克風動態範圍就是指其從自身雜訊到最大聲壓度之間的聲壓範圍、或其轉換的電壓範圍。表4以較淺顯的數值來呈現一範例：若靈敏度為7.75 mV/Pa、最大聲壓度為114 dB_{SPL}、自身雜訊為34 dB_{SPL}，則動態範圍為80 dB。

表 4　麥克風動態範圍計算範例

	聲壓		電壓	
	Pa	dB_{SPL}	mV	dBu
max SPL	10	**114**	77.5	-20
↑乘加差距	x10	+20	x10	+20
sensitivity	1	94	**7.75**	-40
↓除減差距	$x10^{-3}$	-60	$x10^{-3}$	-60
self-noise	10^{-3}	**34**	7.75μV	-100

2.2 距離影響一：訊號衰減

2.2.1 音量衰減導致錄音SNR降低

對於點聲源來說，其聲波在空間自由傳播的聲強會隨距離的平方而衰減，也就是符合物理的平方反比定律(inverse square law)[3]；因此聲壓的衰減與距離成反比關係。例如正常交談距離一米的聲壓約為62dB_{SPL}，而安靜居家的背景雜訊聲壓約為36 dB_{SPL}[3]，如此可得訊雜比SNR為26dB。若所有條件不變，則其他距離的聲壓衰減與相對應SNR如表5所示：

表 5　不同距離之聲壓衰減與相對應SNR

距離米	0.1	1	2	4	10
mPa	252	25.2	12.6	6.3	2.52
dB_{SPL}	82	62	56	50	42
SNR(dB)	46	26	20	14	6

因此麥克風的規格選擇自然成為重點，特別是靈敏度與收音距離直接相關，需要越遠距收音就需要更大靈敏度的麥克風。表6是不同用途下通常麥克風的靈敏度規格範圍[3]：

表 6　不同用途下麥克風靈敏度規格範圍

麥克風用途	正常靈敏度範圍	
	mV/Pa	$dBu/94dB_{SPL}$
近距、手持	2 ~ 8	-52 ~ -40
正常錄音室使用	7 ~ 20	-41 ~ -32
遠距收音	10 ~ 50	-38 ~ -24

不過最終影響語音辨識率的乃是錄下語音訊號的SNR，這裡的N只先考慮收音設備的雜訊，尚不含環境聲學雜訊。牽涉到的是整體數位收音設備，包括：麥克風、放大器、類比數位轉換器 (ADC: Analog-to-Digital Converter)等的動態範圍。基本上麥克風的動態範圍越大、ADC的有效位元數(effective number of bits)越大，可容許的動態範圍就越大，就可以容許在較遠錄音距離時還可保持好的SNR。當然這通常也代表錄音設備成本較高。

2.2.2 不同音素之音量衰減

不同音素因距離衰減而對語音辨識產生的影響其實更為複雜。語音訊號大致可分為兩類：有聲(voiced)與無聲(unvoiced)。有聲語音來自於聲帶振動所產生的週期性聲波；而無聲語音則是聲帶無震動，因此只是無週期性的氣流擾動。有聲語音包含母音、半母音、與有聲子音；無聲語音則都是子音。在語音學和音系學中有所謂響音層級(sonority hierarchy)來區分不同音素的響度，例如英文的響音層級從高到低排列順序如下[6]：

開母音[a] > 中母音[e o] > 閉母音[i u] >
接近音[r] > 邊音[l] > 鼻音[m n ŋ] >
有聲擦音[z v ð] > 無聲擦音[s f θ] >
有聲塞音[b d g] > 無聲塞音[p t k]

理論上來說，響音層級越低，其受距離衰減的影響越嚴重。但實際上，用於語音辨識的特徵並非只有音量，反而主要是頻譜特性、也有時間方面的特徵。因此在各種因素結合以後問題就變得更複雜。尤其當所謂遠距離是有許多不同距離的變化時，就會使得辨識時的語音跟聲學模型訓練語料不匹配的問題更加嚴重。

2.3 距離影響二：室內迴響

2.3.1 DRR變異性

從我們多年的實驗中可知，長距離語音辨識在室內因迴響所產生的問題其實才是最大的挑戰。不同的房間、麥克風的位置、講話者的距離與方向等都會造成很大的變異性。許多室內迴響能量甚至超過直接波，也就是直接波對迴響能量比 DRR (Directive-to-Reverberant Ratio) < 0 dB 的情形也很平常。因為如前面所述，直接波隨著傳播的距離等比衰減，但是房間迴響音場(reverberant field)是室內所有反射波的總和(ensemble)，因此基本上與聲源的距離無關。當室內空間中某一點與聲源的距離，恰好使直接波衰減成與迴響音場的聲壓大小相同，則這樣的距離就稱為臨界距離(critical distance)[3]。

大部分學術論文都是提出麥克風陣列來處理長距離語音辨識問題[7]，但其實麥克風陣列解決的主要是來自與聲源直接波不同方向的環境噪音聲波，並無法完全消除無方向性的迴響聲壓(不過多麥克風至少可以提升訊號能量，因此SNR還是會比單麥克風好)。我們過去也發展麥克風陣列的技術，但實際應用發現，在無響室錄製的語料庫，實驗結果對消除不同方向的噪音干擾非常好；但在實際房間進行測試，就與房間的迴響特性有很大的關係，好壞變化很大。

學術界也有一些研究是關於去除迴響(de-reverberation)的技術，但應用成熟度都還嫌不足。例如兩前年IEEE AASP (Audio and Acoustic Signal Processing) 技術委員會舉辦的IEEE AASP CHALLENGES 2013就是有關殘響處理技術的REVERB (REverberant Voice Enhancement and Recognition Benchmark) 挑戰[8]。結果在語音強化(enhancement)評測的部分，現有技術的確有降低殘響能量的效果，但在主觀品質評量中卻發現無法提升語音品質。另外在語音辨識(recognition)評測的部分，去殘響演算法只是提升辨識率的眾多技術之一。最好的結果是結合了8個麥克風的去雜訊前端、各種最先進語音辨識後端技術(DNN-HMM、模型調適、強健特徵擷取、多條件訓練)、以及多系統整合。如此複雜的系統，其最佳詞辨識錯誤率(WER)

比起頭戴式麥克風的結果還高出約三分之二[9]。由此可見殘響問題之難解。

2.3.2 RT_{60}變異性

不同房間大小與反射特性會造成迴響延遲時間的差異，這也更增加迴響問題的難度。一般都將迴響時間訂為當直接聲源停止後迴響聲場衰減60dB (1000倍)所需的時間，稱為RT_{60}。有一經驗公式為$RT_{60}=(0.16V)/(S\bar{\alpha})$，當中$V$為房間體積(單位$m^3$)，$S$為房間內聲波接觸總表面積(單位$m^2$)，$\bar{\alpha}$為房間內聲波接觸面的吸收率平均值[3]。因此，當房間空間V越大、迴響時間越長；反之當室內人、物越多(S越大)，或吸音效果越好($\bar{\alpha}$越大)，則迴響時間越短。例如，假設一長寬高分別為6m, 3m, 2m的房間，若其$\bar{\alpha}=0.2$，也就是平均20%的入射聲波被吸收、80%反射，則此房間的迴響時間RT_{60}=0.8秒。

我們可將整個房間的迴響視為一種頻道特性(channel effect)，因此可透過收集在房間中的錄音語料對辨識器的聲學模型進行訓練，藉此提高辨識率。我們曾經嘗試在不同的房間重錄訓練語料庫，試圖將空間響應特性訓練進模型裡面。當測試與訓練符合時，的確能得到一些提升的效果，但還是難以回到跟近距離語音辨識一樣的水準。若考慮到實際使用時，難以直接取得與訓練語料庫一樣的空間響應條件時，結果就更不堪想像。因為每個房間的迴響特性不同，甚至若有不同的人在場、穿不同材質的衣服都可能造成變異的影響。

因此，若能藉由估測房間的聲學特性參數，如前述DRR、RT_{60}、$\bar{\alpha}$等，才有可能用來提升殘響處理與語音辨識的效果。傳統上，這些參數需透過測量房間聲學脈衝響應 (AIR: Acoustic Impulse Responses)來進行估測，但實際應用上卻不可行，所以越來越多的研究開始探討如何直接從收錄得到的語音或錄音訊號來估測這些聲學特徵參數。因此去年開始進行的IEEE AASP CHALLENGES 2014就是與此相關的評測：ACE (Acoustic Characterization of Environments) Challenge [10]。其主要目標為針對單一麥克風與多麥克風錄音進行DDR與RT_{60}的估測，最終的評測結果將於今年10/21舉辦之IEEE WASPAA研討會中公佈[11]。

2.4 距離影響三：環境噪音

2.4.1 距離使環境噪音更嚴重

環境噪音的問題原本並非與長距離語音辨識有直接相關。但由於手持或頭戴式麥克風的語音輸入可透過低靈敏度麥克風而避免收錄環境噪音以獲得較好的訊雜比；相較而言長距離語音辨識為了獲得足夠音量必須採用較高靈敏度的麥克風(參考表6)，因此環境噪音的問題就難以避免且更為嚴重。

在實際應用情境中，環境噪音可大分為具方向性與不具方向性兩類。具方向性的噪音來自於特定的聲源如電視機、電話鈴響、其他說話人等。不具方向性的噪音則是來自於房間裏外各種持續性噪音聲源迴響混雜而成，其實也就是環境噪音源結合室內迴響所產生的問題。若是噪音具有方向性，就能以麥克風陣列來解決。但若是不具方向性的噪音就難以靠指向性收音來克服。

最後，若是說話者不是正對麥克風說話時，也就是前述自由角度語音辨識的層次，則連語音訊號來源都不是直接波時，要透過麥克風方向性來分離輸入語音與環境噪音就更難成功。這時大概只能透過盲目聲源分離(BSS: Blind Source Separation)這類技術來將輸入語音跟其他噪音源進行分離。

2.4.2 語音噪音跟語音指令的分辨

所謂訊號跟雜訊(噪音)的定義：系統所要的輸入就是訊號，其他的就都是雜訊(噪音)。在實際環境中，除了要對系統輸入語音指令以外，其他任何人與人的交談，或是電視中播放人物講話的語音，對語音辨識器來說都是一種環境噪音。這類語音噪音在訊號本質上跟語音指令已經沒有分別，因此不可能透過訊號前處理的方法予以分離，而只能在語音辨識的層級加以分辨。

傳統近場語音辨識會透過機器按鍵來告知系統語音指令訊號的起始或終結。但長距離語音辨識的應用情境隱含距離的阻隔，使得按鍵啟動的方法不再適用。因此語音辨識器必須處於永遠啟動(always on)的狀態，且採取語音觸發(voice trigger)的手段。語音觸發有兩類：喚

醒指令(wakeup command)與前導指令(pilot command)，目的都是透過一特定的語音指令來區別語音噪音和語音指令。

喚醒指令的目的是要將系統由非聽令模式轉成聽令模式。也就是平常系統在非聽令模式時，任何輸入的語音都被視為環境噪音；一旦系統偵測出喚醒指令時，之後所收錄的語音就被視為指令語音。至於前導指令則是一種在任何語音指令前都必須包含的觸發指令，這有點像小朋友玩的「老師說」遊戲，遊戲者所講的指令必須前面有「老師說」這個前導指令才能構成有效指令，否則就視為無效的噪音。

3．相關研究

在針對長距離語音辨識中易受噪音干擾的問題上，如2.3.1所提到的，許多研究皆使用了麥克風陣列技術來嘗試解決。傳統的作法是將麥克風陣列做為語音辨識器的前處理，透過波束成型(Beamforming)演算法[12-16]，將語音強化後再交由語音辨識器進行辨識。然而語音辨識是作用在一從語音波形所擷取出來的特徵序列上，上述的波束成型演算法在訊號上所做的強化工作，如最大化訊雜比(Signal to Noise Ratio; SNR)或最小化強化後訊號與原始訊號誤差，並不一定會保證帶來特徵鑑別性的增加以提高辨識正確率。針對傳統做法的缺點，Seltzer等人便提出了最大似然估計波束成型(Likelihood-Maximizing Beamforming; LIMABEAM)演算法[17]，該演算法調整麥克風陣列之參數以最大化正確答案的似然估計(likelihood)來提高辨識正確率。Liao等人則利用填充模型(Filler Model)結合驗證技術來調整麥克風陣列之參數，最大化語音辨識的信心值[18]。上述兩種方法將語音辨識器與麥克風陣列不再視為兩個獨立運作的模組，而是將兩者緊密地結合在一起，讓麥克風陣列的目標不再是語音的強化而是辨識正確率的增加。

在克服房間迴響問題上，一個直覺的方法便是利用房間脈衝響應(Room Impulse Response)的反向濾波器來消除迴響，然而該反向濾波器通常都不是最小相位進而導致濾波結果的不穩定[19]。另外則有研究使用匹配濾波器(Match Filter)來消除迴響[20-21]，雖然匹配濾波器可以帶來訊雜比的提高，然而在辨識正確率的提升上幫助並不大，甚至在脈衝響應為已知的情況下也是如此[22]。另一方面，調適(Adaptation)技術也是被用來解決迴響的方法之一，Kumatani等人利用語者調適技術大幅地降低字錯誤率(Word Error Rate; WER)[7]。

近年來，隨著深度神經網路(Deep Neural Network; DNN)的發展，也開始有人利用神經網路來進行長距離語音辨識的研究。Swietojanski等人利用以深度神經網路為基礎的聲學模型來進行長距離語音辨識，發現可減少8%的字錯誤率[23]。Swietojanski等人更進一步地以迴旋神經網路(Convolutional Neural Network; CNN)來取代原有的神經網路，發現可再帶來6.5%的字錯誤下降率[24]。

為了可以對長距離語音辨識有更深一步的了解，我們將從單一麥克風開始，逐步地檢視關於收音裝置、收音環境、音素類別以及聲學模型種類對於長距離語音辨識的影響，以下章節將針對上述問題進行實驗與探討。

4．收音裝置

最前端之收音裝置為最先接觸到聲音的部份，對於最後的語音辨識結果具有重要的影響，如能接收到高訊雜比之聲音訊號，將有助於之後的語音特徵擷取。在實驗初期，我們針對下列市售麥克風進行測試：

(1) CGG麥克風模組 + 電腦音效卡(CGG)
(2) 鐵三角麥克風AT9903 + 鐵三角麥克風擴大機ATMA2(AT9903)
(3) BEHRINGER電容式麥克風ECM8000 + Tinsea專業話筒放大器mpa mini(ECM8000)
(4) ASUS Xtion(Xtion)

我們以500句測試句(電話語音、每句平均9.46個音節)[25]從高傳真的揚聲器在距離麥克風5公分的位置播放出來，再由各麥克風所收錄後的音檔之音節辨識正確率(無語言模型)做為評估麥克風好壞之標準，測試結果如表7所示：

表 7　市售麥克風音節辨識正確率

麥克風種類	音節辨識正確率
原始電話語音（Baseline）	66.00%
CGG	52.09%
AT9903	57.35%
ECM8000	48.58%
Xtion	59.23%

我們可以從結果發現，ASUS的Xtion所得到的辨識率為最高，而從訊號本身的訊雜比來看，Xtion也是最高的，並且它在長距離（1.5公尺或3公尺）仍可收到語者所發出之語音，不像其他麥克風的錄製結果，人耳已無法分辨出其中語音的部份（麥克風靈敏度不足）。因此最後我們選用ASUS的Xtion做為我們整個實驗的收音麥克風。

5．室內空間響應

在室內長距離語音辨識上，室內空間響應是影響最後辨識率的主要因素之一，而因為不同空間具有不同的空間響應，進而提高了解決該困難點的難度。

在長距離的條件下，我們可以從迴響對於聲音訊號的影響看出室內空間響應是如何影響聲音訊號的感測。在一般室內，當聲音從語者的嘴部發出後，經過四周牆壁、地板與天花板的反射再由收音裝置接收到，所得到的聲音訊號已是原始訊號疊加上由不同方向所反射回來的聲音訊號。

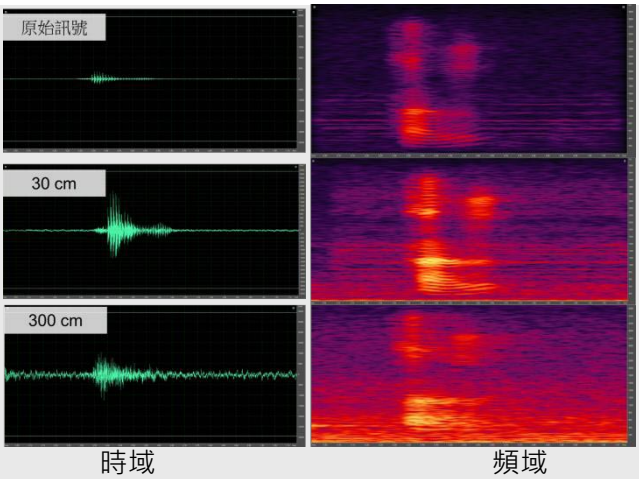


圖 1　迴響對聲音訊號在不同距離的影響

圖1顯示了長距離對於訊號時域及頻域的影響。我們以高傳真揚聲器分別在距離30公分與300公分播放原始信號再以Xtion錄製，30公分訊號頻譜上的諧波結構（Harmonic Structure）與原始訊號相比，已經略顯不明顯；而當距離拉長到300公分時，頻譜上的諧波結構已經模糊不清了，訊號在時域上的週期性也無法明顯的表現出來，而這正是影響語音辨識正確率的主要原因。

圖 2　長距離語料錄製設定

圖 3　（a）空間A與（b）空間B大小與佈置圖

而不同的空間大小與佈置，也會影響所錄製到的聲音訊號。我們以500句測試句從高傳真的揚聲器在距離麥克風300公分的位置播放出來(圖2)，在空間A與空間B(圖3)分別以Xtion錄製，最後的音節辨識正確率如表8所示。

我們可以發現空間A的容積（107m^3）明顯的比空間B（74m^3）大很多，再加上空間A有一面牆皆為玻璃為易反射之材質，因此空間A的RT_{60}約為0.5秒，比空間B的0.45秒來得大，這也反應到了最後的音節辨識正確率，空間A比空間B少了8%。

表 8　相異空間音節辨識正確率

	空間A	空間B
音節辨識正確率	20.58%	28.80%

6．結合空間響應之語音模型調適

由4、5節可以得知，收音裝置與空間響應會對聲音訊號造成影響，使得錄製到的聲音訊號與我們所用來訓練語音模型的訓練語料在通道的特性上有顯著的差異，進而降低辨識正確率。為了嘗試解決這個問題，我們初步希望可以透過錄製受到空間響應影響的語料，並結合調適技術，克服訓練與測試語料通道不匹配的問題，進一步提升辨識正確率。

6.1　長距離語音語料庫

我們於圖3中的空間B進行長距離語料的錄製，該空間長、寬、高分別為522公分、588公分、242公分，裡面擺放有方型會議桌以及12張具扶手之椅子。天花板為石膏板（吸音係數為0.02），牆壁為混凝土（吸音係數為0.02），兩者的吸音係數相較於隔音用絨布（吸音係數為0.63）都遠小的多，顯示對於聲音的反射嚴重。

錄製方式分為兩種，第一種為將既有的語料透過高傳真的揚聲器播放出來，並以麥克風進行錄製（回錄方式）。其中揚聲器為Genelec 8030A，麥克風為ASUS Xtion，揚聲器距離麥克風為300公分，麥克風為靠牆放置（圖2）。第二種為邀請語者至現場依照文稿錄音，語者所站位置同樣離麥克風為300公分（真人錄音）。

回錄方式所錄製的語料內容為電影查詢語料14818句，共約23.49個小時以及測試語料500句，約0.45個小時。真人錄音則請了50位語者依照小學4-6年級社會科知識提問相關內容文稿，每人隨機念20句，總共收錄約1.87個小時。

6.2　調適

根據我們之前調適的經驗，我們將長距離語音資料庫做為訓練語料的一部分，加上原始的近距離語音資料庫進行語音模型的訓練。在得到初始的語音模型後，再利用長距離語音資料庫進行調適。我們首先進行 Maximum Likelihood Linear Regression （MLLR）調適，其中MLLR適合用在調適語料數量不多的情況，其主要概念為將語音模型分成數群，再根據輸入的長距離語音資料庫針對不同群去計算相對的仿射變換（Affine Transform）矩陣，最後再將初始的語音模型經由變換矩陣轉換成較符合長距離語音特性之語音模型。接著我們再透過Maximum a posteriori（MAP）調適技術來進一步做語音模型的調整。MAP屬於直接調適模型參數的方法，其為針對模型中各別的參數做調整，因此對於沒有對應調適語料的參數就無法做參數的更新，因此MAP在具有大量調適語料的情況下才會有比較好的結果。

6.3　實驗結果

表9為最後以回錄語料以及真人語料測試調適過後語音模型的結果（無語言模型）。我們可以發現在不同的詞彙內容上，也造成對辨識率不同的影響。

表 9　調適後模型音節辨識正確率（%）

測試語料 / 模型類別	測試句500句 原始音檔	測試句500句 300公分	真人錄音996句 300公分
原始模型	66.84	23.58	25.16
加入長距離語料之語音模型	66.57	32.99	33.67
加入長距離語料後調適之語音模型	41.72	35.27	36.88

另外，我們可以看到原始語音模型在收音距離拉長後，語音辨識正確率大幅下降（66.84%→23.58%），而在加入長距離語料一同訓練後，辨識正確率可提昇10%左右（23.58%→32.99%）；再利用調適技術，可再將正確率提昇約3%（32.99%→35.27%）。

接著透過對錯誤率的分析，我們發現在收音距離拉長後，子音部分（特別是鼻音、塞擦音）正確率大幅下降；而辨識正確下降率前10名90%都是發音部位在前之子音(如圖4所示)，如ㄘ、ㄗ、ㄉ、ㄋ、ㄈ、ㄊ、ㄑ與ㄙ等，推測是因為子音的能量較低，語音訊號無法完整地傳遞到收音的麥克風，造成所得到的資訊不完整使得辨識率大幅下降。因此，在未來設計長距離語音辨識系統時，應特別對該類子音進行處理，以提高最後系統之效能。

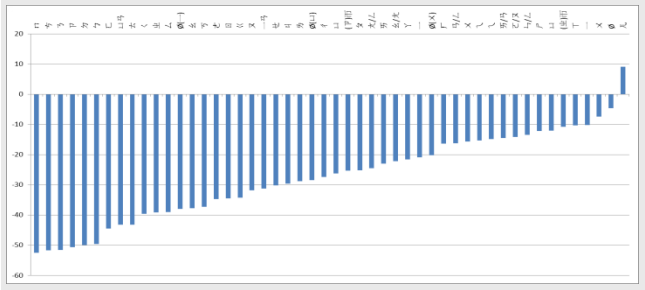


圖 4　各音素辨識正確下降率

7．音素類別與距離對詞語驗證的相對鑑別能力比較

7.1 不同距離之單音素詞語驗證模型建立

表 10　遠/近距離資料庫及單音素詞語驗證模型格式

資料庫描述		
	近距離（15公分）	遠距離（300公分）
文句內容	口語電影查詢	口語電影查詢
時數	23.75 hr	23.75 hr
錄音環境	智慧型手機	空間B回錄
HMM模型描述		
狀態數	3	
高斯混合數	32	
特徵參數	39維MFCC,並經過MVA[26]正規化處理	

為了比較不同距離情況下之詞語驗證效果，我們分別利用兩套資料庫來訓練兩組50個單音素詞語驗證模型，其資料庫描述與模型格式如表10所示。

HMM模型先以Maximum Likelihood (ML)法則訓練至收斂，接著將文稿音素隨機排列(Random Lexicon)後進行Sub-word based Minimum Verification Error(SB-MVE)之鑑別式訓練[27]，提高50個音素模型間的鑑別能力。

7.2 單音素詞語驗證模型能力比較

首先觀察兩組模型在進行鑑別式訓練前後之內部測試(inside test)效果，如表11上半部，在近距離情況下50個音素模型的平均驗證等錯誤率（Equal Error Rate；EER）從12.16%降至6.38%，具有47.5%之降幅。而遠距離之模型則有42.6%之降幅，EER從14.65%降至8.41%。如同我們預測，在長距離的情況下模型的鑑別能力勢必會變差，但鑑別式訓練法則還是有一定程度的效果。

接著使用測試句500句來進行外部測試（outside test），此套語料庫為麥克風錄音，與訓練的兩套資料庫差異性較大，期望可得到較客觀的比較結果。我們將測試語料經過同驗證模型使用之切音程序，將50種音素段落切出後進行EER的評估，其結果如表11下半部所示:

表 11　內、外部測試平均EER測試結果

內部測試結果		
模型	**ML**	**SB-MVE**
近距離	12.16%	6.38%
遠距離	14.65%	8.41%
外部測試結果		
模型	**ML**	**SB-MVE**
近距離	15.13%	9.92%
遠距離	19.07%	13.89%

由表11可得知，在外部測試中不管遠近距離的模型經鑑別式訓練(SB-MVE)後均呈現不到40%EER錯誤降低率，其中遠距離僅有27.2%改進，且經過鑑別式訓練之平均音素EER高達

13.89%，表示此模型可能容易受到環境差異之影響。

再來我們觀察各個音素在經過鑑別式訓練後，遠距相對近距離模型的EER增加程度。表12列出前10個增加程度最大之音素。其中ㄆ、ㄌ、ㄎ及ㄩ之EER增加超過了一倍，顯示ㄆ與ㄎ等破裂音，或像ㄌ與ㄩ等流音及介音之驗證模型對距離的敏感度較高。而ㄘ與ㄙ的遠距離驗證模型EER也超過了20%，再次顯示破裂音甚至摩擦音在遠距離的驗證效果可能不太理想。

表 12　Top-10 EER差異之音素列表

phone	近距EER	遠距EER	EER增加比例
ㄆ	7.05%	18.18%	158.1%
ㄌ	6.07%	15.65%	157.7%
ㄎ	6.17%	13.48%	118.4%
ㄩ	6.17%	13.38%	117.1%
ㄓ	7.96%	14.95%	87.7%
ㄈ	6.56%	11.97%	82.6%
ㄔ	13.28%	23.43%	76.4%
ㄙ	13.38%	22.93%	71.4%
ㄇ	7.37%	12.51%	69.7%
Ø(ㄩ)	6.54%	10.76%	64.6%

7.3 關鍵詞驗證能力比較

觀察完個別音素的驗證能力之後，我們接著實驗關鍵語詞在遠/近距離的驗證效果，藉此瞭解音素的驗證能力是否直接反應在上層詞語或句子當中。在此實驗中，我們分別在三種不同距離(近、中、遠)情境下錄製測試語料。其中近距離為使用者一般手持智慧型手機講話之情境，而中遠距離是用該近距離語音分別設定在150及300cm下所回錄而成。在測試集合外詞彙(Out Of Vocabulary set)方面，除了一般詞語內容，我們另外也錄製了會議人聲與電視節目播放聲音等測試資料。

我們採用[28]裡所提出之關鍵詞驗證演算法CM_3，將各個音素的驗證分數集合統計成最後的關鍵詞驗證分數。其運算式如下所示:

$$CM_3 = \frac{1}{N}\sum_{n} min(LLR_n^*, 0)$$

其中LLR為音素之log-likelihood ratio，N為關鍵詞內的音素數目。另外可以注意到，原始公式可將關鍵詞CM_3分數視為所有音素分數之均等加權(1/N)總和，亦即每個音素的比重是一樣的。

在本實驗中，我們想探討在不同距離條件下，音素的鑑別能力是否能反映在上述公式之比重上，鑑別能力越高給予較大之比重值，反之則給予較低值。如此依音素鑑別能力加權之CM_3計算應該可以得到更好之關鍵詞驗證效果。

表 13　10-fold CM3結果

	權重值				訓練EER	訓練臨界值	測試EER	測試臨界值
	Ø	ㄚ	ㄈ	ㄨ				
近距離								
平均值	0.44	0.10	0.18	0.28	0%	-1.48	0%	-1.77
標準差	0.02	0.02	0.05	0.04	0%	0.03	1%	0.27
中距離								
平均值	0.44	0.19	0.10	0.27	2%	-1.36	1%	-1.43
標準差	0.01	0.01	0.02	0.03	0%	0.04	2%	0.19
遠距離								
平均值	0.37	0.25	0.06	0.32	2%	-1.33	2%	-1.40
標準差	0.01	0.02	0.02	0.02	0%	0.07	3%	0.23

針對三種距離環境，我們請30個人每人5句錄製了關鍵詞〝阿福〞，另外每個人錄製了23句指令詞句作為集合外詞彙使用，如此形成三組環境相關的實驗語料。我們將每一組實驗語料又拆成10-fold validation sets，其中90%的語料用來搜尋各個音素最佳的比重值，10%語料用來測試該比重組合。表13為三種距離環境下10-fold CM_3之平均實驗結果。

由表13結果可以得知，ㄈ摩擦音如先前所觀察隨著距離環境變長，其權重值隨之下降。而母音(ㄚ、ㄨ)則是隨著距離拉長而越顯示其重要性。而在整體驗證效果上，均等權重值之EER在三種環境下之EER分別為為0.7%、3.4%、4.0%。上表結果顯示，在10-fold的實驗中當適當地調整音素間的權重值後，平均EER分別可進步到0%、1%、2%，比均等權重時來得好，但效果變化幅度較大並不夠穩定，顯示此一想法在概念上應該可行，值得收集更多資料後再深入探討。

8．基於深度神經網路之語音辨識技術

深度神經網路(Deep Neural Network; DNN)近年來在影像與語音辨識上獲得很好的表現。在語音辨識上，深度神經網路主要是用來取代原本高斯混合模型(Gaussian Mixture Model; GMM)的部份。我們利用訓練好的深度神經網路來計算每個觀察到的特徵參數在每個音素上的機率，接著再利用訓練好的隱藏馬可夫模型(Hidden Markov Model; HMM)來將觀察到的一連串特徵參數解碼成一連串的音素。

深度神經網路與原有之神經網路最大的差異便是在神經網路初始值的取得上，深度神經網路透過一層層神經網路的個別訓練，可以得到一個良好的初始值，這個初始值可以讓之後的神經網路訓練得到較以往更好的結果。

表14為利用深度神經網路進行語音模型訓練的結果(無語言模型)。我們可以觀察到在五層隱藏層的設定下，加入長距離語料後，深度神經網路所得到的語音模型相對於原始的語音模型，可以得到59.55%的音節辨識正確增加率(30.41%→48.52%)。我們並可以觀察到加入長距離語料訓練後之語音模型來進行近距離語音辨識測試所得到的結果，依然可以維持在原有的水準，顯示該語音模型具有較佳的強健性。

表 14　深度神經網路模型音節辨識正確率(%)

測試語料 / 模型類別	測試句500句 原始音檔	測試句500句 300公分	真人錄音996句 300公分
原始模型	78.80	30.41	27.89
加入長距離語料之語音模型	78.97	48.52	45.67

9．結論

本文從解析長距離語音辨識的困難點開始，分析了收音裝置及室內空間響應等主要影響長距離語音辨識正確率之因素，並嘗試結合長距離語音語料及調適技術來解決空間響應對訊號所造成的影響。從結果來看，加入長距離語料調適之後，雖可增加10%以上的辨識正確率，但仍無法達到可用的地步。從各個音素的辨識正確下降率來看，鼻音與塞擦音等子音易受距離之影響，導致該語音訊號無法完整地傳遞，造成所收到的訊號不完整而使得辨識率大幅下降。而在詞語驗證的實驗中也可看到類似的結果，在一些破裂音與摩擦音的驗證效果上，正確率都因距離增加而降低了。這也或許解釋了為何以深度神經網路為基礎的語音模型會具有較高的辨識正確率之原因，因為深度學習(Deep Learning)本身為一鑑別式訓練，在語音相關時域與頻域等資訊因距離而衰減時，鑑別式訓練可以增加各個語音模型間之距離，提高彼此的鑑別程度。在未來或許可以基於如深度神經網路這樣具高鑑別性之語音模型上，再針對不同音素之辨識結果進行權重之調整，以降低因距離而衰減的訊號物理特性之影響，提高最後的語音辨識正確率。然而深度神經網路會比傳統高斯混合模型需要較多的運算量，如何在運算資源與效能間取得平衡，是在採用深度神經網路前所必須要進一步研究的課題。

參考文獻

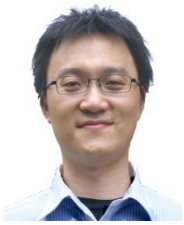
[1] International Electrotechnical Commission, IEC 60268-4:2014, Sound system

equipment - Part 4: Microphones, 2014.
[2] C. Woolf, and R. Streicher, Microphone-Data, http://microphone-data.com/.
[3] J. Eargle, The Microphone Book (Second Edition), Elsevier, 2004.
[4] C. Woolf, What the fields mean, http://microphone-data.com/help/#Max-SPL .
[5] M. Williams, The interpretation of the microphone data sheets, http://microphone-data.com/library/articles/ .
[6] W.D. O'Grady, J. Archibald, Contemporary linguistic analysis: An introduction (7th ed.), Toronto: Pearson Longman. pp. 70, 2011.
[7] K. Kumatani, J. McDonough, and B. Raj, "Microphone Array Processing for Distant Speech Recognition: From Close-Talking Microphones to Far-Field Sensors", IEEE Signal Processing Magazine, vol.29, no.6, pp.127–140, 2012.
[8] Audio and Acoustic Signal Processing Technical Committee, REVERB Challenge, IEEE AASP CHALLENGES 2013, http://reverb2014.dereverberation.com/.
[9] Keisuke Kinoshita, "Challenge summary", REVERB workshop, Florence, Italy,10 May 2014. http://reverb2014.dereverberation.com/workshop/proceedings.html
[10] Audio and Acoustic Signal Processing Technical Committee, ACE Challenge, IEEE AASP CHALLENGES 2014, http://www.ee.ic.ac.uk/naylor/ACEweb/index.html.
[11] IEEE, Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 21 October 2015.
[12] D. H. Johnson, and D. E. Dudgeon, Array Signal Processing, Englewood Cliffs, NJ: Prentice Hall, 1993.
[13] L. J. Griffiths, and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," IEEE Trans. Antennas Propagat., vol. AP-30, pp. 27–34, 1982.
[14] S. Nordholm, I. Claesson, and M. Dahl, "Adaptive microphone array employing calibration signals" , IEEE Trans. Speech Audio Processing, vol. 7, pp. 241–252, 1999.
[15] D. V. Compernolle, "Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings" , in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, vol. 2, pp. 833–836, 1990
[16] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters", IEEE Trans. Signal Processing, vol. 47, pp. 2677–2684, 1999.
[17] M.L. Seltzer, B. Raj, and R.M. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition", IEEE Trans. Speech and Audio Processing, vol.12, no.5, pp.489–498, 2004
[18] H.-C. Liao, Y.-F. Liao, and C.-H. Lee, "Maximum Confidence Measure Based Interaural Phase Difference Estimation for Noise Masking in Dual-Microphone Robust Speech Recognition", in INTERSPEECH-2011, pp. 473–476, 2011.
[19] S. Neely, and J. Allen, "Invertibility of a room impulse response", J. Acoust. Soc. Amer., vol. 66, no. 1, pp. 165–169, 1979.
[20] J. L. Flanagan, A. C. Surendran, and E. E. Jan, "Spatially selective sound capture for speech and audio processing", Speech Commun., vol. 13, no. 1–2, pp. 207–222, 1993.

[21] S. Affes, and Y. Grenier, “A signal subspace tracking algorithm for microphone array processing of speech,” IEEE Trans. Speech Audio Processing, vol. 5, pp. 425–437, 1997.
[22] B. Gillespie, and L. E. Atlas, “Acoustic diversity for improved speech recognition in reverberant environments”, in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, vol. 1, pp. 557–560, 2002.
[23] P. Swietojanski, A. Ghoshal, and S. Renals, “Hybrid acoustic models for distant and multichannel large vocabulary speech recognition”, in Proc. IEEE ASRU, pp.285–290, 2013.
[24] P. Swietojanski, A Ghoshal, and S Renals, “Convolutional neural networks for distant speech recognition”, IEEE Signal Process. Letters, pp. 172–176, 2014.
[25] R.-L. Chiou, and H.-C. Wang, “A preliminary test of MAT-160 speech database in connected syllables recognition”, in Proc. Int. Symp. Chinese Spoken Language Processing (ISCSLP), pp. 89–92, 1998.
[26] C.-P. Chen, and J.A. Bilmes, “MVA Processing of Speech Features”, IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.1, pp.257–270, 2007.
[27] R. Sukkar, ”Subword-Based Minimum Verification Error (SB-MVE) Training for Task Independent Utterance Verification” , in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, pp. 229–232, 1998.
[28] T. Kawahara, C.-H. Lee, and B.-H. Juang, “Combining keyphrase detection and subword-based verification for flexible speech understanding”, in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, vol. 1, pp 193–196, 1997.

作者簡介

廖憲正



工研院資通所資深研究員。專長為生理訊號處理、語音訊號處理與機器學習。

郭志忠

工研院資通所正研究員。國立清華大學電機工程博士。專長為語音訊號處理、計算語言學、音樂合成與音訊處理。
E-mail: cck@itri.org.tw

林政賢

工研院資通所聲音與文字處理技術部研究員。畢業於台北科技大學電子工程研究所。專長為語音訊號處理
E-mail: jslin@itri.org.tw