**三、研究計畫內容（以中文或英文撰寫）：**

1. **研究計畫所要探索的課題**
2. **探索課題概述**

設想如下情境：

甄老84歲, 由40餘歲未出嫁小女兒(阿信)照顧. 房間分布如圖1. 甄老有早上6:00起身

在客廳(圖1右方Living Room)繞圈, 扶杖走路運動的習慣. 而阿信要到上午7:30才會起

床.

某日早晨甄老依習慣在客廳繞圈, 忽然不注意摔倒, 痛楚難當,又沒力氣自己爬起來, 只

得連叫"阿信." 阿信驚醒, 急忙由臥室(圖1下方Bedroom)到客廳，檢視甄老情況，打電

話找來兄弟姊妹, 一番討論後, 決定叫救護車運送至較近的XX醫院, 照過X光後, 發現

是髖骨骨折, 必須住院開刀.

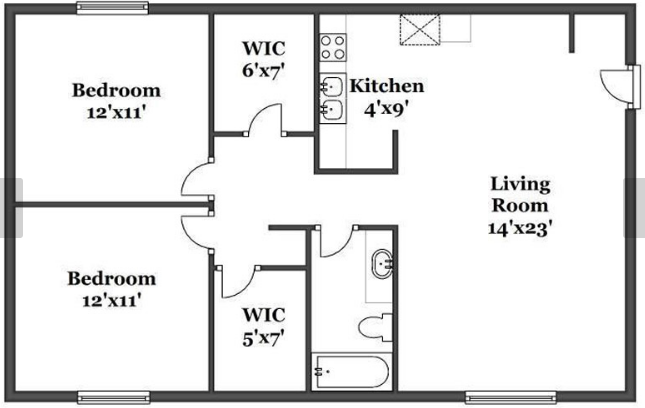


圖1. 典型的室內環境 (取自http://kalamazoo.apartments/hillside-village/floor-plans/)

本計畫目的即是探索發展機器人音源定位及聲音辨識技術，能夠陪伴老人；像前述情境的阿信一樣，在老人摔倒等情況發生時，辨識聽到的聲音，決定聲音來源方向位置，迅速移動到發出聲音處，進行必要處置。

**計畫創新特點(相關名詞術語稍後說明):**

1. 應用nengo神經系統模擬工具，建立聲音訊號由外耳、中耳、內耳、腦幹(含主要神經核: MSO、LSO、IC、DLNN等)上傳，經由MGB傳至聽覺A1、A2皮質區的整合仿生計算模型。
2. 以腦幹各神經核計算模型的連接與運作，模仿人類應付回音環境(reverberant environment)及相對運動聲源的機制。
3. 應用深度學習網路(deep learning network)演算法，以來自MGB的生理訊號，訓練聽覺皮質區建立Spaun架構機器人所使用的語意指標(semantic pointer)，並用以辨識聲音。例如：察覺重物摔倒聲，聽到機器人自己的名字等。此一語意指標可與Spaun機器人其他感官產生的語意指標儲存至其他腦皮質區，進行資訊融合(fussion)、推理、驅動運動部件等活動。
4. 運用計算機圖學內的3D射線追蹤演算法，產生接近實際房間環境如圖2的室內脈衝響應函數(Room Impulse Response，RIR)。室內脈衝響應函數與聲源發出之聲音訊號計算摺積(convolution)後，可產生接收位置收到的訊號，作為產生回音環境(reverberant environment)及相對運動聲音訊號之用。
5. 應用機器人作業系統(Robot Operating System，ROS)所附的導航及移動程式庫，以及Gazebo機器人模擬系統，呈現3D場景，顯示機器人在如圖2的房間配置環境中，轉頭面向聲音來源，靜止或向聲源位置移動等的行為過程。
6. 以上各功能之整合。

**簡化探索課題的假設：**

1. 單一聲源：處理多個聲源是通常認為困難的Audio Scenario Analysis問題，擬作為本研究後續計畫的一部分。
2. 擬真人頭部之人形機器人：以便應用網路上可取得的假人頭量測到的頭部轉移函數(Head Related Transfer Function，HRTF)。一般機器人頭部與人類頭部的相似程度較低，其HRTF要另行計算或量測。任意頭部外形對應的HRTF計算，主持人已有構想，同樣可作為本研究後續計畫的一部分。
3. 聲音辨識種類: 僅考慮不同雜訊背景下的國語數字0到10的發音及若干特殊聲音如重物落地聲(模仿老人摔倒發出的聲音)、老人呻吟聲、機器人名字等。不牽涉日常語彙之語音辨識，擬列為未來機器人與人對話系統研發目標之一。
4. 機器人的移動和導航系統利用ROS現有程式庫：暫不考慮另外開發仿生模擬模組。
5. 適度簡化各神經單元的仿生模型：參考相關文獻，減少不相關功能的實作，盡量求取擬真程度與計算效率和結構複雜度的平衡。
6. **探索課題背景**

隨著人口高齡化，子女忙於工作，照顧獨居老人或兩老的工作可能要由照護工擔任。但是老人照護工作繁瑣又缺乏成就感，很難找到足夠的人手。同時，由於少子化，能像阿信一樣有兄弟姐妹分攤心理與生理壓力的情況，越來越少見。因此老人陪伴機器人的研發，刻不容緩。世界先進國家如丹麥、英國、日本、義大利等國都陸續有安養中心導入機器人，取代部分人力，也受到安養中心老人的歡迎。這些機器人的一項主要功能，是感知人類的情緒，與老人對話[[1]](#footnote-1),[[2]](#footnote-2),[[3]](#footnote-3),[[4]](#footnote-4),[[5]](#footnote-5)。台灣在老人陪伴機器人的研發方面也有進展，例如華康科技開發的「康粟平台機器人」，已捐贈南開科大，合作開發監控老人健康等功能[[6]](#footnote-6)。

除了媒體報導的機器人辨識人類情緒及與人對話等功能之外，本計畫所提的聽音辨位與聲音辨識，也是學界與業界極力探討的研究主題；進入Google Scholar搜尋，光是2015及2016兩年，所發表的機器人辨識聲源位置(localization)相關論文，就達到80餘篇，而且近年來，每年發表的相關論文數均在成長。以下簡述本研究相關的各領域背景知識及相關重要文獻回顧。

* 1. **聲源定位: 使用麥克風陣列**

談到機器人的聲源定位，許多人先想到的便是麥克風陣列技術。根據2015

年發表的一篇survey paper (S. Argentieria, P. Danès, P. Souères, 2015)，在2000年左右，開始有「機器人聽覺」(robot audition)論文發表。這約15年期間，聲源定位是相關論文的大宗。Argentieria等人在那篇論文中，將機器人聲源定位的方法分成兩大類：仿效人耳的雙麥克風(binaural)技術與多麥克風陣列(microphone array)技術，並且指出機器人聽覺技術，與傳統聲音定位技術，如水下聽音系統相較，必須符合：1)幾何限制(geometry constraint)，麥克風系統必須能夠裝設在機器人平台上；2)即時限制(real time constraint)，訊號處理必須在短時間內完成，並且做出反應；3)頻率限制(frequency constraint)，處理的訊號必須含括語音頻帶(約為300 Hz到3,300Hz)；4)環境限制(environment constraint)，必須應付實際聽覺環境，包括回音與雜訊，以及聲源與機器人間的相互運動。雙麥克風(binaural)技術在下一小節說明，此處簡介多麥克風陣列技術。

依據(S. Argentieria, P. Danès, P. Souères, 2015)，麥克風陣列技術主要有三類：

1. MUSIC (MUltiple SIgnal Classification)。利用多個麥克風收到的訊號，算出其共變異矩陣(covariance matrix)，計算其特徵值與特徵向量，建立一個偽頻譜(pseudo-spectra)函數。求取使pseudo-spectra函數極大的參數，便可估計聲源個數及其位置。假設多個聲源之間有某種統計關係，便可以處理回音環境下的聲音訊號。但是MUSIC演算法處理的對象是一個特定頻率的弦波，所以要對足夠多頻率進行計算，需要相當長的計算時間。雖然已有若干技術改進，MUSIC演算法不容易滿足即時與頻率兩項限制。
2. 訊號相關函數(correlation function)。假設陣列中第*i*與第*j*個麥克風，收到含有雜訊的聲音訊號分別為與，則兩訊號的相關函數為

其中代表包含聲源訊號及雜訊的期望值計算。求取極大值對應的**值，可以近似相同音源發出之聲音訊號抵達兩個麥克風的時間延遲(Time Delay(s) of Arrival，TDOA) 。令兩麥克風的空間距離為，在遠場平面波的假設下，可以由簡單幾何圖形，導出聲源位置與兩麥克風連線夾角(azimuth)的關係式

並且由此決定聲源方位(代表音速)。文獻上可以查到許多有效估計訊號相關函數的方法，例如將訊號的Fourier轉換相乘，其次乘上依照需求，對各頻率的某權重函數，再以逆Fourier轉換，獲得訊號相關函數。顯然，遠場平面波的假設對低頻訊號以及一般較近距離的對話不易成立。如果有多個聲源位於回音環境中，訊號相關函數會出現不止一個尖峰，使聲源與聲波反射影響不易分辨，定向效能也會下降，成為這種方位估計法的最大缺點。

1. 波束形成(beamforming)。假設聲波由聲源位置發出，進入第*n*個麥克風，

形成接收到的訊號。此一訊號經過若干延遲與權重，相當於通過一個脈衝響應為濾波器(頻率響應為，是波數，*f*為頻率)，加入總和，得到輸出訊號

根據系統對不同聲源位置以及聲源發射的訊號，將系統輸出訊號平方，可以算出一個能量地圖(energy map)

表達出*t*時間所收到的位置聲源發射的功率。所以在能量地圖上求出的尖峰便可能是各聲源位置的評估。遠場及近場聲源的濾波器可以表成不同的形式；例如，平面波遠場對應的濾波器，是坐標系原點指向位置的單位向量，是第*n*個麥克風的位置向量。同時可以證明，這種方位估計方法的解析度受陣列的場型(beam pattern或array pattern)限制：

考慮遠場聲源時，，成為方位角的函數。其最大值對應方位角附近形成場型的主波束(main beam)，波束寬度決定陣列的解析度：如果兩個聲源位在同一個主波束內，就無法區分。一般而言，波束寬度約與陣列大小成反比，所以要有足夠解析度時，會需要較大的陣列。陣列的大小，則是以波長度量。因此波束形成法在頻率較低(波長較長)時，需要較大的陣列，極可能違反幾何或頻率限制。

* 1. **聲源定位: 使用雙麥克風**
  2. **人類聽覺系統**
  3. **神經系統模擬軟體Nengo**
  4. **異常(anomy)聲音偵測**
  5. **語意指標(semantic pointer)與Spaun機器人架構**
  6. **回聲相關之建築聲學**
  7. **3D 射線追蹤**
  8. **機器人作業系統及Gazebo機器人模擬軟體**

1. 說明機器人辩識聲音來源方位的重要性
2. 說明機器人辨識聲音來源方位方法上的限制
3. 指出探索課題: 回音環境下移動人形機器人的仿生聲源定位及聲音識別

Biologically-Inspired Sound Source Localization and Sound Identification for a Mobile Humanoid Robot in a Reverberant Environment

何謂複雜環境?

何謂仿生計算?

聽覺週邊圖

腦幹聽覺訊號路徑圖

系統方塊圖

Nengo概念圖

1. 既有的方位辨識方法

4.1 麥克風陣列方法

4.2 雙耳辨識方法

4.4 仿生方法

聽覺訊號傳播處理路徑

HRTF

外耳

中耳

內耳

MSO

LSO

IC

1. **研究方法及進行步驟**
2. 仿生聽覺系統建立
   1. HRTF from Brian
   2. 聽覺周邊 : IPEM 模型
   3. Jeffrey model for MSO in Nengo
   4. LSO model in Nengo
   5. IC model in Nengo
   6. 聽覺皮質區
3. 語意指標產生
4. 室內脈衝響應函數計算
5. 與ROS運動與導航系統的銜接
6. Gazebo軟體的應用
7. **預期達成目標及可行性評估指標**

**預期達成目標**

1. 產生一篇國際會議論文，預計2018年6~7月，計畫接近結束時投稿。
2. 所完成的軟體系統，預備以下列五種測試場景類型，在單一聲源的假設下，進行數值實驗。

各實驗中的機器人擬真人頭HRTF取自於 Brian Hear Python程式庫(官網http://www.briansimulator.org/docs/hears.html#head-related-transfer-functions)，及CIPIC資料庫(Algazi, Duda, Thompson, &Avendano, 2001)(官網

http://interface.cipic.ucdavis.edu/sound/hrtf.html)，並比較所產

生的差異。

第三、四種場景類型使用roomsim (Campbell, Palomaki, &Brown, 2005)

(MATLAB程式原始碼位於<https://sourceforge.net/projects/roomsim/>)

產生室內脈衝響應函數BRIR(Binaural Room Impulse Responses)，

由於roomsim的計算方法是以簡單的聲源多次鏡射影像，求反射聲波的

傳播路徑及聲音強度，無法應用到實際的複雜房間環境，所以我們另外

以計算機圖學的3D射線追蹤方法，求算複雜房間環境的BRIR，並應用

於第五種測試場景類型。

1. 所完成的Python程式，移植到Robot Operating System (ROS)(官網http://www.ros.org/)，以ROS所附機器人模擬系統Gazebo(官網http://gazebosim.org/)呈現3D場景，顯示機器人在無窮空間/長方形空房間內，靜止或運動時，轉頭面向聲音來源的行為過程。
2. 完成上述以3D射線追蹤方法計算更實際房間環境的BRIR程式，並公開為ROS下的一個process。

**測試場景類型1**

探討無反射(anechoic)環境下的靜止機器人聲源定位以及語意指標建立準

確率，與文獻已知數值實驗結果或神經生理特性比較，並討論對應物理生

理機制。

聲源：為簡化問題，設定俯仰角(elevation) 0o; 水平角(azimuth) -90o至

90o， 間隔30o，正前方為0o。

與機器人頭部中心距離1.28 m

雜訊: 不同訊雜比(Signal to Noise Ratio，SNR)的White Gaussian、冷氣

聲、與交通背景雜訊。

場景說明：

場景1.1:

聲音訊號：低頻弦波(100 Hz、100 x 151/15 Hz、100 x 152/15 Hz、…、1500 Hz)，預期ITD影響較ILD為大。

場景1.2:

聲音訊號：高頻弦波(1200 Hz、1200 x 151/15 Hz, 1200 x 152/15 Hz、…、

18,000 Hz)，預期ILD影響較ITD為大。

場景1.3:

聲音訊號：自然語音(五個英語字：”hello,” “look,” “fish,”

“coffee,” “tea”)，(Liu et al., 2008) 指出這五個單字包含相當大

範圍的日常語音型態)。雖然我們人形機器人的頭部形狀與

(Liu et al., 2008)的機器人頭部不同，語音特徵也有差異，但是仍可與

其結果參考比較。

場景1.4:

聲音訊號：國語數字0到10的發音、重物落地聲(模仿老人摔倒發出的聲

音)、老人呻吟聲、老人呼喚機器人名字、門窗開啟聲、玻璃破裂聲。

數值實驗觀察要點:

場景1.1~1.4: 求得單聲源方位水平角，畫出類似

(Liu, Erwin, &Wermter, 2008)之圖9(d)的圖表(複製於本計畫書圖 1)，

並探討其準確性及相關物理生理機制。



圖 1.聲音方位辨識結果呈現，圖中方塊大小代表Estimated Angle附近的神經脈衝數與整體脈衝速率的比例大小。本圖取自(Liu, Erwin, &Wermter, 2008)文中的Fig. 9(d)。

場景1.4: 對於國語數字0到10的語音，以及特殊聲音，產生各聲音訊號的語意指標，估算判定語意指標的正確率並對較重要的參數如聲源方向角度作圖。

**測試場景類型2**

探討回音環境(reverberant environment)下，靜止機器人聲源定位以及語

意指標建立的正確率，與文獻結果及測試場景類型1的各對應場景所得比較，討論相關物理與神經生理機制。

多重反射回音環境:

長方形空房間(shoebox 模型)，使用與

(Youssef, Argentieri, &Zarader, 2012)相同的設定。主要參數如下：

5 x 4 x 2.7 m，灰泥(acoustic plaster)牆、木頭地板、混凝土天花板；

濕度50%；溫度20oC，同時考慮空氣吸收率及聲音強度隨距離產生的衰減。

此時回音時間。

牆壁的吸收係數可以修改，得到RT60等於0.45s與0.7s的回音，一併加入數值實驗，比較不同RT60的結果。

機器人頭部:中心位置: (2, 2, 1.5)m。

聲源: 使用與(Youssef et al., 2012)相同的設定。為簡化問題，設定俯仰角0o;

水平角 -45o至45o，間隔5o，正前方為0o。與頭部中心的距離由1m 到2.8m，間隔0.45m。

雜訊：與測試場景類型1相同。

**場景說明:**

場景2.1: 聲音訊號為自然語音(五個英語字：”hello,” “look,”

“fish,” “coffee,” “tea”)，結果可以參考(Youssef et al., 2012)(Youssef, Argentieri, &Zarader, 2013)(Mroueh, Marcheret, Goel, &Ibm, 2015)。

場景2.2: 聲音訊號為國語數字0到10的發音、重物落地聲(模仿老人摔倒

發出的聲音)、老人呻吟聲、老人呼喚機器人名字、門窗開啟聲、玻璃破裂

聲。

**數值實驗觀察要點:**

場景2.1~2.2: 以不同RT60之值，畫出聲源方位水平角誤差與RT60的相關

曲線，並探討其準確性及相關物理生理機制。

場景2.2: 對於國語數字0到10的語音，以及特殊聲音，產生各種聲音訊

號的語意指標，估算判定語意指標的正確率並對RT60統計作圖。

**測試場景類型3**

探討無反射環境下，機器人在不同速率下直線或繞圈運動，求出聲源定位以

及語意指標建立的正確率，相對於時間的變化，與文獻結果及測試場景類型1的各對應場景所得比較，討論相關物理與神經生理機制。

聲源：為簡化問題，設定俯仰角(elevation) 0。

測試開始時，位於機器人正前方，與機器人頭部中心距離1.28 m。

雜訊: 不同訊雜比(Signal to Noise Ratio，SNR)的White Gaussian、冷氣

聲、與交通背景雜訊。

**場景說明:**

場景3.1: 聲音訊號：低頻弦波，同場景1.1。

場景3.2: 聲音訊號：高頻弦波，同場景1.2。

場景3.3: 聲音訊號：自然語音，同場景1.3。

場景3.4: 聲音訊號：同場景1.4。

場景3.5: 聲音訊號: 同場景1.4。

當聲音為重物落地聲或呻吟聲或機器人名字時，機器人直線移動到發出聲

音處。

**數值實驗觀察要點:**

場景3.1~3.3: 求得聲源方位水平角，畫出各弦波頻率下，誤差與時間的

關係圖，並探討其準確性及相關物理生理機制。

場景3.4: 求得聲源方位水平角，畫出聲源定位以及語意指標建立的正

確率，相對於時間的變化，並探討其準確性及相關物理生理機制。

場景3.5: 聲音辨識為重物落地聲時，或呻吟聲，或機器人名字，觀察機

器人是否直線移動到發出聲音處。

**測試場景類型4**

探討回音環境下，直線或繞圈運動機器人聲源定位以及語意指標建立的正確

率，與文獻結果及測試場景類型2的各對應場景所得比較，討論相關物理與神經生理機制。

多重反射回音環境: 同測試場景類型2。

機器人頭部: 起始中心位置: (2, 2, 1.5)m。

聲源: 起始時位於機器人正前方，與機器人頭部中心距離1.28 m。

雜訊: 不同訊雜比(Signal to Noise Ratio，SNR)的White Gaussian、冷氣

聲、與交通背景雜訊。

**場景說明:**

場景4.1: 聲音訊號：自然語音，同場景1.3。

場景4.2: 聲音訊號：同場景1.4。

場景4.3: 聲音訊號: 同場景1.4。

當聲音為重物落地聲或呻吟聲或機器人名字時，機器人直線移動到發出聲

音處。

**數值實驗觀察要點:**

場景4.1~4.2: 以不同RT60之值，畫出聲源方位水平角誤差與RT60的相關

曲線，並探討其隨時間變化的準確性及相關物理生理機制。

場景4.2: 對於國語數字0到10的語音，以及特殊聲音，產生各聲音訊號

的語意指標，估算判定語意指標的正確率隨時間變化的關物理生理機制，並

對RT60統計作圖。

場景4.3: 聲音訊號: 同場景1.4。

當聲音為重物落地聲或呻吟聲或機器人名字時，觀察機器人是否直線移動

到發出聲音處。

**測試場景類型5**

測試以3D射線追蹤方法計算更實際室內環境的BRIR程式正確性，並用於

類似圖1所示的房間分布環境，其中當然也可放置家具。

**場景說明：**

場景5.1: 利用場景類型3和4的shoebox模型與設定，計算BRIR。

場景5.2: 利用圖1房間配置，不放家具，計算BRIR。

場景5.3: 利用圖1房間配置，安放數個形狀簡單的家具如桌椅，計算BRIR。

場景5.4: 利用場景5.3的房間和家具，聲音訊號同場景1.4，對於國語數字0到10的語音，以及特殊聲音，產生各聲音訊號的語意指標，估算判定語意

指標的正確率隨時間變化的關物理生理機制，並對RT60統計作圖。

場景5.5: 利用場景5.3的房間和家具，聲音訊號同場景1.4，當聲音為重

物落地聲或呻吟聲或機器人名字時，機器人以ROS中的navigation system，移動到發出聲音處。

**數值實驗觀察要點:**

場景5.1: 檢驗所得的BRIR是否與採用場景類型3和4的shoebox模型計

算結果一致。

場景5.2~5.3:檢視所得BRIR的各個尖峰位置是否有反射機制合理對應。

場景5.4~5.5:分別與場景4.2及4.3比較，討論其差異。

**可行性**

本研究橫跨數個研究領域，不易全部通曉。但主持人近年多所涉獵，雖尚無具體成果，卻也建立若干基礎，並非從零開始。本研究探索之課題，主持人以目前所知及研究生程度判斷，有信心於一年期限內，完成大部分研究目標。

**評估指標**

以各場景通過測試的情形(場景類型與通過場景數)為評估指標，尤其場景5.5為本計畫所有概念之總驗收。請參見各測試場景說明。

**參考文獻**

Algazi, V. R., Duda, R. O., Thompson, D. M., &Avendano, C. (2001). THE CIPIC HRTF DATABASE Creative Advanced Technology Center 1500 Green Hills Road Scotts Valley , CA 95066. *Signal Processing*, (October), 99–102.

Campbell, D. R., Palomaki, K. J., &Brown, G. (2005). A MATLAB simulation of“ shoebox” room acoustics for use in research and teaching. *Computing and Information Systems J*, *9*, 48–51.

Liu, J., Erwin, H., &Wermter, S. (2008). Mobile robot broadband sound localisation using a biologically inspired spiking neural network. *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, 2191–2196. http://doi.org/10.1109/IROS.2008.4650760

Mroueh, Y., Marcheret, E., Goel, V., &Ibm, M. I. T. (2015). Deep Multimodal Learning for Audio-Visual Speech Recognition, 2130–2134. http://doi.org/10.1109/ICASSP.2015.7178347

Youssef, K., Argentieri, S., &Zarader, J. (2012). Towards a Systematic Study of Binaural Cues, 1004–1009.

Youssef, K., Argentieri, S., &Zarader, J. L. (2013). A learning-based approach to robust binaural sound localization. *IEEE International Conference on Intelligent Robots and Systems*, 2927–2932. http://doi.org/10.1109/IROS.2013.6696771

1. 賴麗秋, 機器人好好玩 陪伴、照護、運動、學習、遊戲一手通包, 2016年7月1日中國時報電子報開卷，<http://magazine.chinatimes.com/lifeplus/20160701004431-300507> [↑](#footnote-ref-1)
2. <https://read01.com/JLQj86.html> [↑](#footnote-ref-2)
3. <https://www.inside.com.tw/2016/09/19/alice_cares> [↑](#footnote-ref-3)
4. <https://www.youtube.com/watch?v=YgqKlskR1JY> [↑](#footnote-ref-4)
5. 2014年5月26日TVBS報導http://news.tvbs.com.tw/tech/532908 [↑](#footnote-ref-5)
6. 2016年9月7日中國時報電子報http://www.chinatimes.com/realtimenews/20160907005248-260405 [↑](#footnote-ref-6)