聲學場景改變之偵測技術

Acoustic Scene Change Detection Approach

林昶宏　張振魁　陳明彥
Chang-Hong Lin, Chen-Kuei Chang, Ming-Yen Chen

中文摘要

本研究提出在室內空間之中，利用脈衝音(Impulse Sound)自喇叭陣列週期性地發出脈衝音，透過聲波的反射與室內的物體、地面、牆面與屋頂等作用產生對應的殘響(Reverberation)，此殘響之特徵即可反應出室內空間的特性，我們稱為聲學場景(acoustic scene)。當室內空間中的狀態產生變化，如家具擺設位置改變、門窗開啟或關閉、人員數量的增減時，再次於此空間透過脈衝音反射取得的聲學場景即會顯現出差異，此時即可評估室內空間是否發生擾動。本技術透過擷取室內殘響的頻譜，計算頻譜與初始空間的頻譜的頻譜差量，再轉換為時域上的差值向量以表示聲學場景與初始聲學場景在各個時間點的差異。當差異的總和超過訓練時得到的門檻值時，則代表聲學場景改變。

Abstract

This work proposes an acoustic scene change detection approach in the indoor space by reverberation. Speakers emit impulse signals periodicity. Sound waves are then influenced with objects, the floor, walls, and roofs, and generate corresponding reverberation. This reverberation can represent the characteristic of the indoor space, which is called acoustic scene. When the state of the indoor space changes, such as furniture changes positions, windows are open or closed, and the number of people in the room increases or decreases, the acoustic scene would differ from the one before. In this way, we can evaluate whether the indoor scene changes or not. This work presents the difference between the current acoustic scene and the initial acoustic scene by calculating the difference between the current spectrogram and the initial spectrogram of the reverberation. When the sum of acoustic scene differences is greater than the threshold decided in the training stage, we say that acoustic scene change occurs.

關鍵詞(Key Words)

聲學場景轉換偵測(Acoustic Scene Change Detection；ASCD)

聲學場景建模(Acoustic Scene Modelling；ASM)

房間脈衝響應(Room Impulse Response；RIR)

喇叭陣列(Speaker Array；SA)

1 · 前言

居家安全在近幾年獲得越來越多的注意，其中最廣為人知的系統莫過於基於視覺分析的監控系統，例如透過網路攝影機監控居家環境的場景，當監視的範圍發生場景轉換時發送視訊給使用者。雖然基於視覺分析的監控系統[1,2]可以即時地在異常發生時通知使用者，而且攝影機僅在使用者不在時開啟，但是個人居家環境的隱私問題仍然是使用者卻步的原因。在其他可能的感測方法中，分析環境中的聲音是一種可行性高的方法。與攝影機相比，如果是家中無人的情況下，我們認為收集使用者家中的聲音就較不具隱私性的問題。

藉由分析聲音來理解環境的狀況可分為被動式以及主動式兩種類型。被動式的方法是透過麥克風監聽聲音以及分析環境聲音的特性來得到監聽環境的特性。如透過辨識聲音事件來理解環境。例如在Chu . *et al.*的方法中[3]，藉由擷取聲音參數，以及訓練特殊聲音事件模型來辨識是否有特定事件發生。這類型的方法主要是受限於需要訓練大量標記好的音檔來訓練聲音模型。在Tegborg *et al.*的研究中[4]，透過聲音的督普勒(Doppler, 都普勒)效應，分析聲音源移動的速度及來源角度。但是此方法在來源聲音的音壓太小時，分析出的頻率可能會不夠可靠。主動式的方法是透過揚聲器發出聲音，再透過接收器分析目標的反射音來得到目標的位置資訊。此種類型的方法常常搭配超音波感測器來建立陸上聲納系統(In Air Sonar System; IASS) [5]，但超音波感測器受限於聲音的特性，感測範圍通常較小。另一種主動式方法[6-8]是利用麥克風搭配揚聲器，分析麥克風收到的室內殘響訊號，來估測房間的脈衝響應，再透過分析房間的脈衝響應來取得房間的體積或是幾何(geometry)資訊。此類型的方法雖然可以獲得房間幾何資訊，例如牆壁大小、位置的資訊，但是卻沒有揭露對於較小的擺設是否具有相同的分析能力。

在本篇論文中，我們利用麥克風/揚聲器作為收發音裝置，來取得較大的感測範圍。我們分析殘響的反射音部分來建立空間的狀態資訊，稱作聲學場景。在訓練階段，我們建立初始環境的聲學場景誤差值範圍，當測試時的聲學場景與初始聲學場景的差值高過範圍時，則代表空間狀態改變。

2 · 反射音模型

假設喇叭與麥克風在同一位置，初始場景的殘響如下圖1左圖所示。當有人進入此場景時，人會將部分的脈衝音反射回麥克風，因此反射音(reflected sound)會疊加在初始殘響訊號上，如下圖1右圖所示。

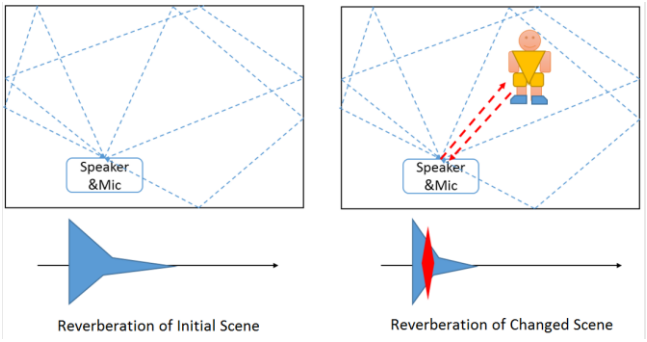

圖 1　反射音模型之示意圖。

令麥克風訊號為$\mathbf{y}$，初始場景的殘響為$\mathbf{i}$，$\mathbf{r}$為人造成的反射音，則我們將麥克風訊號定義為下式:

$$\mathbf{y}=\mathbf{i}+\mathbf{r} \tag{1}$$

其中$\mathbf{y}$, $\mathbf{i}$以及$\mathbf{r}\in\mathbb{R}^{L}$，$L$為殘響的長度。

從時頻域來看，令$Y(f,\tau)$、$I(f,\tau)$、$R(f,\tau)$分別為$\mathbf{y}$, $\mathbf{i}$以及$\mathbf{r}$的時頻圖(spectrogram)，其中f為頻帶的索引，τ為音框的索引，則麥克風訊號的頻譜$Y(f,\tau)$也能定義成$I(f,\tau)$與$R(f,\tau)$的疊加。

$$Y(f,\tau)= I(f,\tau)+ R(f,\tau) \tag{2}$$

假設第f個頻帶，第τ個音框的初始殘響期望值為$E[I(f,\tau)]$，則我們可以求得反射音$R(f,\tau)$

$$R(f,\tau)= Y(f,\tau)\text{-}E[Y(f,\tau)] \tag{3}$$

則第τ個音框的時域差值(temporal difference)$D(\tau)$為

$$D(\tau)=\sum_{\forall f} R(f,\tau) \tag{4}$$

3 ．聲學場景轉換偵測

我們提出的聲學場景轉換偵測方法可分成四個步驟，第一個步驟為殘響訊號擷取，裝置所收到的音訊串流先經過切割程序切成多筆殘響訊號。第二步驟是殘響訊號篩選，將音壓過大的殘響訊號視作離群樣本刪除。第三以及四個步驟則是聲學場景模型的建立以及比對。在建模階段收到的殘響訊號被用來建立聲學場景模型，在比對的階段收到的殘響訊號與模型比對後產生差異分數。比對階段可分成開發以及測試階段。若測試階段的差異分數大於開發階段的差異分數則代表聲學場景已經轉換。

3.1 殘響訊號擷取

本研究所提的殘響訊號擷取，主要是週期性地產生一脈衝音訊號並透過發聲裝置發送，經過空間中聲音的反射作用，再由收音裝置定期收集房間回聲的聲音串流。圖2為殘響訊號擷取的示意圖，為了截取回聲串流中的殘響訊號，本研究透過一個尖峰選取(peak selection)演算法找出可能的尖峰，每筆殘響訊號為尖峰及後L點所形成的訊號。每筆殘響可分為3個部分，分別為直達音(Direct Sound, DS)區段、第一反射音(Early Reflection, ER)區段(圖3的綠色區段)、以及多次反射音(Late Reflection, LR)區段。本研究只採集第一反射音區段的資訊，因為第一反射音區段的訊號具有受測物與裝置間的距離資訊。反觀直達音代表脈衝不經反射到達收音裝置的直達音，由於本研究的麥克風以及喇叭都在同一裝置上，因此直達音不帶有受測物的資訊。至於直達音區段雖包含有部分受測物的反射音，考量受測物的反射音的能量較為微弱，在直達音區段的能量也較為微弱，因此本研究未針對此區段進行分析。後續處理即針對殘響訊號第一反射音區段進行分析。

3.2 殘響訊號篩選

在真實的錄音環境中，除了聲學場景改變外，還有許多因素會影響室內空間的殘響。例如收發音裝置本身是否能穩定產生訊號，或是外在的環境噪音都是可能的變因。為了移除收發音裝置以及環境噪音的影響，本研究提出一個殘響訊號篩選的方法，主要確保每一筆殘響訊號的音壓一致，並且丟掉包含突發性噪音(如裝置不穩定的音壓輸出，或是類似脈衝的短時環境噪音)的殘響訊號。門檻值設定為訓練資料的音壓平均值，若訓練資料跟測試資料的殘響訊號能量小於此門檻值則移除不用。

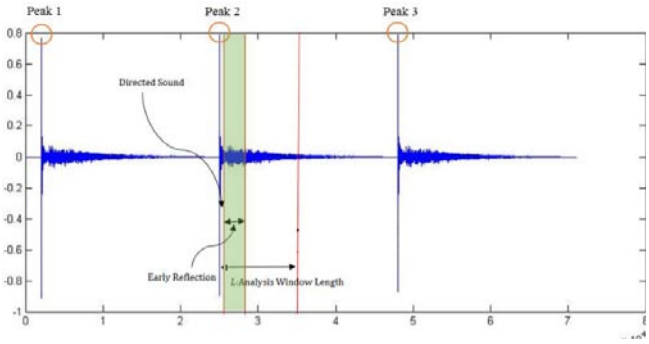

圖 2　殘響訊號擷取步驟示意圖。

3.3 聲學場景模型建立

本研究所提的聲學場景模型建立，主要是針對初始時間區段內所取得的殘響訊號，計算其頻譜期望值以及時域差值上界(Upper Bound, UB)計算兩個子步驟。在建模時，我們將蒐集到的殘響訊號分成兩部分，分別稱為訓練資料以及開發資料。第一個子步驟用訓練資料來計算頻譜期望值 $E[I(f,\tau)]$，第二個子步驟則透過(4)將開發資料轉換為時域差值，用來建立模型的上界 $b(\tau)$。$b(\tau)$的物理意義為非空間變異所造成的擾動，成因如麥克風穩定性以及背景噪音。假設利用開發資料得到多筆時域差值向量 $D(\tau,i)$，i 為資料的索引值，則上界 $b(\tau)$可由下式得到

$$b(\tau)=E[D(\tau,i)]+\ \alpha\ E[D(\tau,i)-\ E[D(\tau,i)]] \quad (5)$$

圖3為時域差值上界的示意圖，其中藍色實線為 $D(\tau,i)$，紅色實線則為 $b(\tau)$。調整 α 可提升系統模型的靈敏程度。

3.4 聲學場景模型之比對

本研究所提的聲學場景模型之比對，主要是比較目前的時域差值向量與初始空間時域差值向量的差異性是否過大來判斷室內狀態是否有擾動的產生。圖4 為模型比對的示意圖。利用(4)可計算出目前空間時域差值向量，如下圖 4 的藍色實線。為了得到較具統

計意義的時域差值向量，此方法計算時域差值向量短時間內的期望值。透過(5)的門檻值(下圖 4 的紅色虛線)計算，本研究比較時域差值向量的期望值超過門檻值的程度來判斷擾動是否產生。在(6)中，S 代表時域差值向量的期望值超過門檻值的程度，若 $D(\tau)>b(\tau)$，則累加 $D(\tau)$與 $b(\tau)$的差值。在圖 4 中，S 代表綠色區域的面積。

$$S=\sum_{\tau,\ \text{if } D(\tau)-b(\tau)>0}(D(\tau)-b(\tau)) \qquad (6)$$

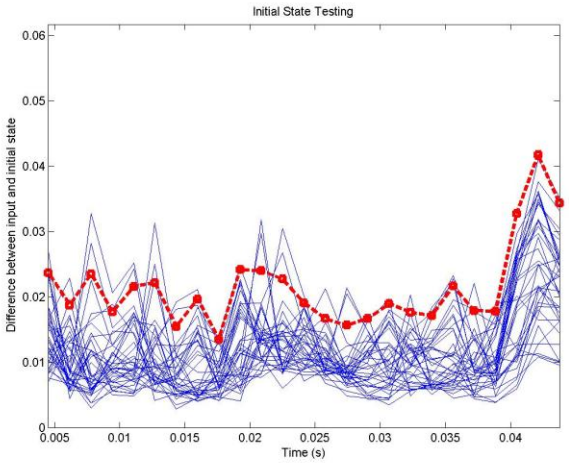

圖 3　空間狀態特徵模型之建模示意圖。

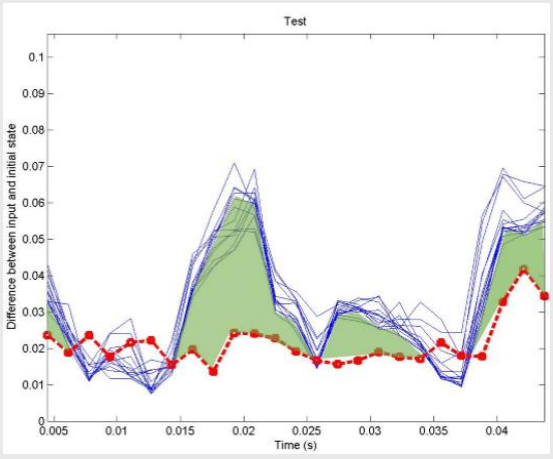

圖 4　空間狀態特徵模型之比對示意圖。

4．實驗數據

在我們的實驗中，我們採用陣列式喇叭(speaker array)以及雙聲道(Stereo)、全向型(Omnidirectional)麥克風來構成我們的脈衝音訊號感測裝置，裝置如下圖5所示。因為本篇的目標是分析反射音的距離資訊，所以我們只分析其中一支麥克風的聲音訊號。在我們實驗的過程中，我們發現喇叭也具有指向性，且一個喇叭的有效反射音收音角度約為30度，為了確保整個空間的反射音都能有效地被麥克風接收到，所以我們採用了4個喇叭來建置喇叭陣列。在我們的實驗中，所有的喇叭都發出相同的脈衝訊號，此外，為了確保前後時間點收到的殘響訊號不會互相混淆，脈衝訊號的時間間距需大於房間的殘響時間(Reverberation Time, RT)，我們設定為0.5秒。我們在室內環境音壓為40~43dB的狀態下，發射57dB的脈衝訊號，並收集其作用於房間後的殘響訊號。

圖 5　脈衝音訊號感測裝置。

實驗的初始受測環境如下圖 6 所示，房間大小為 6m*3.5m*2.5m，受測環境的主要擺設有 4 個物體，圖 6 的脈衝音訊號感測裝置放置於房間的角落。在裝置的左邊跟右邊分別有一張椅子跟桌子，桌上擺有電腦設備。在裝置的前方，距離 3 公尺的地板上，擺放一個保麗龍製的物品，大小為 0.4m*0.2m*0.1m。

圖6　實驗的初始受測環境。

在模型比對的階段，我們可透過計算(6)的S值來判斷測試時的殘響訊號是否接近初始的聲學場景，S值越小表示越接近初始的聲學場景，反之代表聲學場景轉換的可能性越高。為

了找出合適的門檻值來界定聲學場景轉換是否發生，我們對四種情境進行分析，分別為初始的受測環境、初始的受測環境+站立的人、初始的受測環境+房間內擺設改變、初始的受測環境+移動的人。以下是4種情境的設定以及比較。

4.1 初始的受測環境

在與初始的受測環境相同情況下，藉由觀察(6)的S值是否為一個很小的值(而且具有小的標準差)可來判斷初始模型的是否具有一致性。一致性越高，則越容易分辨聲學場景是否改變。在此情境中，訓練初始聲學場景模型的音檔長度為1分鐘，測試的音檔長度為10分鐘。測試時的S值如下表一，S值的平均值為0.0047，標準差為0.0045。

4.2 初始的受測環境+站立的人

在此情境中，受測環境比初始的受測環境多了一個站立的人，這個人身高1.77m，體重為73kg。在此情境的受測環境中，受測者站在房內的各個定點，屋內有九個定點，定點距離脈衝音訊號感測裝置有3種選擇:3公尺、4公尺、5公尺。定點的方位有30度、90度、120度3種選擇。每個定點的測試音檔長度為2分鐘。我們計算相同距離、不同方位下S值的平均值以及標準差，實驗結果如下表一所示。實驗數據顯示出，在多了受測者的情況下，S值的平均值都至少比初始的受測環境多了50倍左右，標準差則是多了至少4倍。代表反射音造成的S值是顯著的，初始狀態具有較高的一致性。此外，S值平均值隨著距離增加而下降，顯示出反射音的音壓隨著距離變長而衰減。

表 1 S值-增加站立的受測者的情況

	初始環境	3m	4m	5m
平均值	0.0047	0.4476	0.3170	0.2286
標準差	0.0045	0.0347	0.0399	0.0212

4.3 初始的受測環境+擺設改變

此情境與初始受測環境的差別是，我們改變地上物品(圖7a)的擺設方式。在初始的受測環境中，物品的擺放方式如圖7a。在此情境中，我們選擇了兩種擺設方式，如圖7b以及圖7c。在圖7b中，物品被平放在地板上，如同在距離裝置3m處，減少了與音波垂直的表面積。在圖7c中，物品被往後方平移約0.5m，相當於減少了3m處與音波垂直的表面積，並且增加了3.5公尺處與音波垂直的表面積。此兩種情況的S值顯示在表二，平移的情況比平放的情況有著更高的S平均值。與表一相比，平移的情況稍微小於受測者站立於3m的情況，等於站立於4m的情況。顯示出除了表面積外，物體的材質也影響反射音壓的重要因素。

圖7a.物品， 圖7b.物品(平放)，圖7c.物品(平移)

表 2 S值-改變屋內擺設的情況

	初始環境	平放	平移
平均值	0.0047	0.2845	0.3088
標準差	0.0045	0.0376	0.0238

4.4 初始的受測環境+移動的人

此情境是受測人在初始受測環境移動的狀況。受測者有3個可移動的區域，分別為距離脈衝訊號感測裝置2m~3m處、3m~4m處、以及4m~5m處。受測者在每個區域來回走動，移動範圍為30度到120度，每個區域的測試時間為2分鐘。此情境的S值如下表三所示，對於S值的期望值，移動的人的情況略小於站立的人的情況，與初始的受測狀態相比則至少大了20倍。對於S值的標準差，移動的人的情況則是比站立的人的情況大了兩倍，顯示出在受測者移動的狀態下，聲學場景改變的十分頻繁，所以有著較大的S值的標準差。

表 3　S值-增加移動的受測者的情況

	初始環境	2m~3m	3m~4m	4m~5m
平均值	0.0047	0.3496	0.3037	0.1190
標準差	0.0045	0.0801	0.0726	0.0399

在以上4種情境設定下，S值期望值都遠大於初始受測環境的S值期望值，且初始受測環境的S值標準差也遠小於其他的情境，我們可以透過此發現來界定出一個合理門檻值。假設在所有的情境下，S值都呈現為常態分佈，令初始受測環境的S值期望值為$\underline{S}_{mean}$，標準差為$\underline{S}_{std}$，則我們可以設定門檻值為$\underline{S}_{mean+}3\times\underline{S}_{std}$，當$S$值大於門檻值代表聲學場景被改變了。

以下我們進行一個受測者進出受測環境的模擬，來驗證演算法的有效性。我們假設實驗剛開始前3分鐘的聲學場景都沒改變，可用來訓練模型以及設定門檻值。第1分鐘用來訓練(5)的$b(\tau)$，2~3分鐘用來測試初始受測環境來得到門檻值。測試時間為4分鐘，第1分鐘跟第4分鐘為初始受測環境，2~3分鐘受測者進入受測環境，並在環境中進行隨機走動(Random Walk)。一共有兩個受測者進行模擬，受測者A身高為1.6m，體重55kg，受測者B為前面實驗的受測者。實驗數據如表四以及表五所示。

表四顯示出受測者A的漏失偵測率(Miss Detection Rate)為7.28%以及假情報率(False Alarm Rate)為6.32%，受測者B沒有漏失偵測而且假情報率只有3.62%。為了進一步降低漏失偵測以及假情報，不是每秒進行偵測，而是偵測時間增加為10秒，10秒內有效的偵測結果以投票方式決定聲學場景是否改變。結果如下表五所示，受測者A沒有假情報，漏失偵測率降為4%，受測者B則沒有任何的預測錯誤。

表 4　聲學場景轉換偵測結果-單次測試

受測者A	場景轉換(實際狀況)	場景不變(實際狀況)
場景轉換(預測)	92.42%	6.32%
場景不變(預測)	7.28%	93.68%
受測者B	場景轉換(實際狀況)	場景不變(實際狀況)
場景轉換(預測)	100%	3.62%
場景不變(預測)	0.00%	96.38%

表 5　聲學場景轉換偵測結果-累計測試

受測者A	場景轉換(實際狀況)	場景不變(實際狀況)
場景轉換(預測)	96.00%	0.00%
場景不變(預測)	4.00%	100%
受測者B	場景轉換(實際狀況)	場景不變(實際狀況)
場景轉換(預測)	100%	0.00%
場景不變(預測)	0.00%	100%

5．結論

本研究提出在室內空間之中，利用脈衝音作用於空間內的殘響訊號，建立聲學場景模型。之後週期性或透過喇叭陣列發出脈衝音，

取得當下之聲學場景模型並與初始聲學場景模型比對，得到擾動的程度可作為聲學場景之變化的依據。

實驗結果可歸納與整理如下：我們設定了4種情境，初始的受測環境、受測環境多了站立的人、初始受測環境的擺設改變、初始的受測環境多了一個在移動的人，我們發現在聲學場景未改變的情況下，擾動程度遠小於其他情境。就由這個發現我們訂出一個合適的門檻值，並且進行受測者進出受測環境的模擬，實驗數據顯示，我們提出的方法在偵測時間長度為10秒時，沒有假情報，且漏失偵測率為2%。

參考文獻

[1] D. Wu, S. Ci, H. Luo, Y. Ye, and H. Wang, ‘‘Video surveillance over wireless sensor and actuator networks using active cameras,’’ IEEE Trans. Autom. Control, vol. 56, no. 10, pp. 2467–2472, Oct. 2011.

[2] Y. Ye, S. Ci, A. Katsaggelos, Y. Liu, and Y. Qian, “Wireless video surveillance: A survey,” IEEE Access, vol. 1, pp. 646–660, Sep. 2013.

[3] S. Chu, S. Narayanan, and C.C.J. Kuo, “Environmental sound recognition with time-frequency audio features,” IEEE Trans. Audio, Speech, Language Processing, vol. 17, no. 6, pp. 1142–1158, 2009.

[4] V. Tegborg, M. I. Pettersson and I. Claesson, “Experimental results of passive imaging of moving continuous broadband sound sources within a sensor field,” IEEE Journal Oceanic Engineering, vol.36, no.1, pp.25-36, Jan. 2011.

[5] J. Steckel, A. Boen, and H. Peremans, “Broadband 3-D sonar system using a sparse array for indoor navigation,” IEEE Trans. Robotics, vol. 29, no. 1, pp. 161–171, Feb. 2013.

[6] I. Dokmanic, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, “Acoustic echoes reveal room shape,” in Proc. Nat. Acad. Sci. (PNAS), 2013, vol. 110, no. 30.

[7] F. Antonacci, J. Filos, M. Thomas, E. Habets, A. Sarti, P. Naylor, and S. Tubaro, “Inference of room geometry from acoustic impulse responses,” IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 10, pp. 2683–2695, Dec. 2012.

[8] N. R. Shabtai, Y. Zigel, and B. Rafaely, “Feature selection for room volume identification from room impulse response,” in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2009, pp. 249–252.

作者簡介

林昶宏

2014年於國立中央大學資訊工程研究所取得博士學位。目前任職於工研院資通所，擔任工程師一職。研究興趣為聲音訊號前處理、聲音訊號分析、以及機器學習。

張振魁

國立中央大學資管所碩士。目前任職於工研院資通所，擔任智慧聯網創新技術與服務組副組長以及兼任部門經理。研究主題為機器學習、大型商務軟體架構規劃，及使用者洞察、服務與商業模式規劃。

陳明彥

2009年於國立成功大學製造資訊與系統研究所取得博士學位，現為工研院資通所資深工程師。研究方向為自然語言處理、語意分析、以及生理訊號量測。