結合關鍵詞驗證及語者驗證之雲端身份驗證系統

A Cloud Speaker Authentication System Based on Keyword Verification and Speaker Verification

邱義欽　　范雋彥　　林伯慎
Yi-Chin Chiu　　Chuan-Yen Fan　　Bor-Shen Lin
國立臺灣科技大學 資訊管理學系
Department of Information Management
National Taiwan University of Science and Technology
david_yc_chiu@yahoo.com　kynwu.tw@gmail.com　bslin@cs.ntust.edu.tw

摘要

電腦和網際網路的誕生，讓人們的生活越來越便利。而隨著行動裝置的快速發展，人類的生活方式更是產生了非常大的變革，不僅需要的資訊，信手拈來便可以獲得；許多企業所提供的新興商品與服務交易，更是在彈指之間便可以順利完成。因此，如何在網際網路上提供使用者方便、快速、彈性、可靠的身份驗證，並免除使用者記憶及輸入一大堆用戶名稱及密碼的負擔，便成為一個重要的課題。本研究結合了關鍵詞驗證和語者驗證技術，讓使用者不需要記憶及輸入冗長與煩雜的資訊，只要對著智慧型行動裝置說話，身份辨識系統便可以在網際網路的環境中對使用者來進行身份驗證。我們以隱藏式馬可夫模型和高斯混合模型分別實作了關鍵詞驗證模組與語者驗證模組，並以分散式架構實作出雲端即時身份辨識系統。我們以 TCC-300 語料進行語者模型參數和訓練流程的調校實驗，以改進語者驗證效能的訓練流程；並對背景語者篩選方法及性別相關模型進行實驗，探討不同條件下的系統設計方法。實驗的結果顯示，在語者模型之混合數設定為 15、迭代次數設定為 10，背景語者的數目設定為 50 人的情況下， F 值可以達到 0.9875，展現出不錯的效能。

關鍵詞：雲端身份驗證，分散式辨識，關鍵詞驗證，語者驗證，高斯混合模型

一、緒論

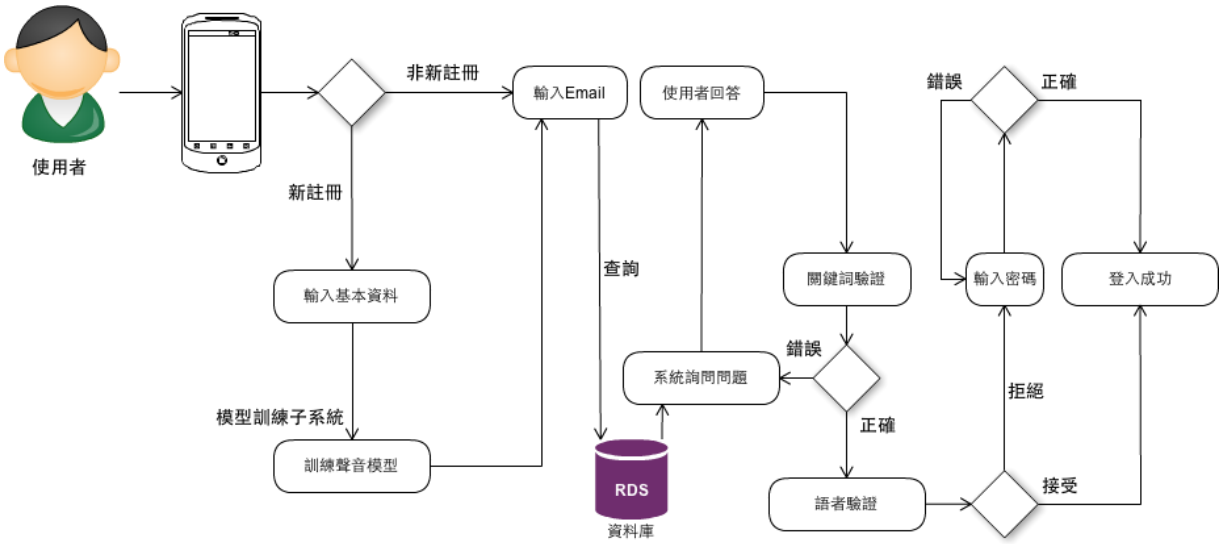
使用者身份的驗證是任何系統安全上的基本議題，特別因為網路的匿名特性，就變得益形重要。傳統上身份驗證方法主要可分為三大類：所是(Who one is)、所知(What one knows)、所有(What one has)。「所是」驗證的目標為確認為其本人，所使用的特徵通常為生物特徵，如指紋、虹膜等。而常用的鑰匙、門禁卡、信用卡、或是會員卡則是根據「所有」來判斷；這類方法不易確認持有者為擁有者本人，而易產生遺失、偽造而遭冒用的情形。因此，有時會加上簽名、照片或錄音等方法來減低安全風險。在網路環境常

用的「用戶名稱加密碼」驗證方式，則是根據「所知」來判斷。這類方式也有一些限制與缺點，包括使用者必須會打字、裝置必須能提供軟硬體鍵盤、密碼可能被盜取等。因為這些缺點，衍生出許多問題，包括使用者必須經常修改密碼、常須記憶多組密碼、並可能忘記密碼等不便，這甚至會造成使用障礙。有鑑於此，發展一種方便、快速、具有彈性、免除記憶許多密碼、又不用擔心被盜用的身份驗證方法，對於網路服務就非常重要。

由於人類發聲器官的生理結構本身就具有獨一性，而說話的習慣和口音，也不易被模仿或複製，這都使得語音驗證可能成為一種普遍可靠的身份驗證方式。在生物度量(Biometrics)的技術分類上，語音認證如同簽名(Signature)，是少數同時兼具生理(Physiological)和行為(Behavioral)兩類特性的生物度量方法，在信用卡、網路銀行、電子商務的驗證、登入與存取控制等應用，具有很大的潛力。而語音驗證同時又具有方便、快速、非侵入性等特性，在應用的推廣上，相較於指紋較可避免隱私權的爭議，這是其另一優勢。雖然語音驗證技術並非完美，但這不是無法克服的障礙。因為事實上，幾乎沒有驗證技術不會產生任何漏洞，只要適當地設計系統，就能在效率與風險間達到折衷。例如，已被廣泛應用的銀行帳號、信用卡或旅行支票的簽名，股票下單的錄音，或銀行個人資料存取的問答錄音等方式，都會有潛在的安全缺陷。然而，這些缺陷並沒有阻礙這些驗證方式被廣泛應用。只要能運用不同技術彼此互補，並在系統設計上適當降低並控制風險，不完美的技術仍可以產生符合需求的應用。本研究就是希望藉由語音、語者驗證和其它身分驗證方法適當地結合，實現出更友善的網路認證模式。在駭客猖獗的網際網路，也可做為防止身份遭盜用的重要防線。

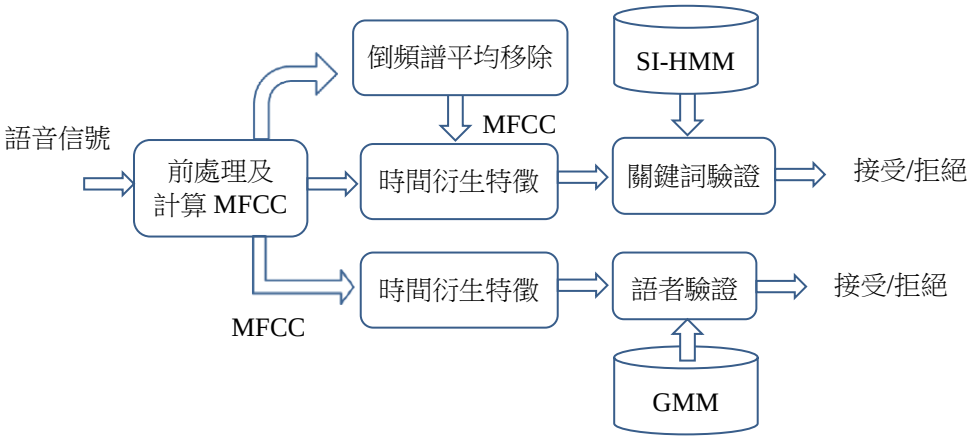
二、語音驗證方法

隨著網路服務的普及，如何應用語音驗證技術以提供使用者更方便、快速、具有彈性、並準確的網路身份驗證服務，極具有重要性。我們提出了可結合關鍵詞驗證技術(Keyword Verification)和語者驗證技術(Speaker Verification)的分散式多重驗證架構。此架構能夠對於使用者的語音，同時驗證是否為其所宣稱的身份[5,6]，以及該語音內容是否符合使用者所預設之關鍵詞彙[7]，如住家地址、公司行號、就學單位或電話等。本身份驗證系統的使用情境如圖一所示。使用者以行動設備登入雲端服務時，驗證系統首先判斷使用者是否為新註冊者，若使用者為第一次登入，則須輸入基本資料，並且錄下一段長度約六十秒的語音供系統訓練語者模型使用。模型訓練完成後，使用者可以用輸入帳號(E-mail)的方式登入。系統會根據使用者預設之問題或個人資料來詢問，像是使用者的基本資料，住址、居住縣市等。若使用者回覆的語音符合答案，並且語音驗證結果確為使用者本人，則系統給予存取資料之權限。此方式可同時驗證「所知」(關鍵詞)及「所是」(說話者)，增加身份驗證的效率和可靠度。若能再結合「所有」的驗證方法(如 RFID)，就可以提供不同安全等級的多重驗證的方式，讓網路驗證服務可以更快速且具有彈性。



圖一、系統使用流程圖

語音驗證的基本系統如圖二所示。首先，語音信號透過音框擷取取出長度為 256 個取樣點的音框，接著進行預強調濾波並乘上漢明視窗。接下來經過快速傅立葉轉換、梅爾濾波器、對數運算、及離散餘弦轉換，可以得到梅爾頻率倒頻譜係數(Mel-Frequency Cepstrum Coefficients, MFCC)。梅爾倒頻譜係數可再計算其時間衍生特徵，稱為動態梅爾倒頻譜係數。這些特徵共同組成了辨識用的特徵向量。由於關鍵詞驗證模型為語者不特定(Speaker Independent, SI)的隱藏式馬可夫模型(HMM)，在計算時間衍生特徵之前，須對 MFCC 特徵進行倒頻譜平均移除(Cepstrum Mean Subtraction, CMS)計算，以消除語者發聲通道和錄音通道的差異。而語者驗證模型是各語者的高斯混合模型(GMM)，必須保留語者發聲通道差異，故不須對 MFCC 特徵進行倒頻譜平均移除計算。本系統將 15 維的 MFCC 以及一次與二次差分特徵組成共 45 維的特徵向量；而特徵向量的序列即為為兩個驗證系統的特徵。在圖二中，特徵向量的序列會以串流的方式傳入關鍵詞驗證和語者驗證模組，進行即時同步的搜尋比對，產生驗證結果。



圖二、特徵參數擷取流程圖

語音驗證技術主要的作法概述如下：

(a) 關鍵詞驗證

關鍵詞偵測與驗證是傳統語音辨識技術的一環，傳統除了聲學辨識單位的研究外，還包括了聲學模型訓練、語者調適、信心度量[9]、鑑別式訓練法(如最小化音素錯誤法)、以及決策模組(如支撐向量機分類器[10-11])等。本系統中使用隱藏式馬可夫模型做為關鍵詞驗證的模型，並使用梅爾倒頻譜係數(MFCC)[8]及其衍生時間特徵作為辨識特徵。由於驗證系統須對一句語音同時進行語者驗證與關鍵詞驗證，且關鍵詞為使用者自訂問題的答案，因此關鍵詞驗證需要能夠處理不特定關鍵詞。本系統使用的模型是以 TCC-300 語料訓練而成的語者不特定模型，包括 113 個右相關聲母模型、37 個前後文無關韻母模型、以及一個靜音模型。本論文的關鍵詞驗證除使用關鍵詞模型外，也使用聲韻母模型作為填充模型，並加入懲罰值進行關鍵詞偵測[11]。

(b) 語者驗證

語者驗證方式可以概略區分為文本相關(Text-Dependent)及文本無關(Text-Independent)。所謂文本相關，是指系統提示使用者所要說的語音內容必須與系統錄製時所說的語音內容相同或相關，例如某一個特定的數字或字串；而文本無關則需要可以提示使用者說出任意內容的文句，並加以驗證。文本無關的系統具有較好的彈性及安全性，可以提供給使用者更好的保護。本系統中希望可以提示使用者回答個人預設問題的答案或個人資料，同時進行關鍵詞及語者驗證；這些內容具有變動性，因此本系統的語者驗證模組必須能夠提供文本無關的方式來進行驗證。

在語者驗證技術研究上，國內有些著重在基礎技術研究，例如：高斯混合模型[1][2]、鑑別式訓練法[3]；有些著重於應用系統研發，例如：結合語音與人臉辨識區域的門禁系統[4]等。一般的架構是先使用高斯混合模型或是隱藏式馬可夫模型進行聲學辨識，得到聲稱語者及反語者的機率，再經由一決策模組做最後的決策。常用的決策模組有相似度比率測試(Likelihood Ratio Test, LRT)、類神經網路、支撐向量機等。本系統是以高斯混合模型結合相似度比率測試來進行語者驗證。

GMM 模型的公式表示如下：

$$p(\vec{x}\,|\,\lambda)=\sum_{i=1}^{M} w_i b_i(\vec{x}) \tag{1}$$

其中 M 是混合數，$\vec{x}$ 是維度為 D 的特徵向量，$b_i(\vec{x})$ 是高斯分佈，而 w_i 是各個高斯的權重，必須滿足的限制 $\sum_{i=1}^{M} w_i = 1$。高斯分佈的定義如下所示：

$$b_i(\vec{x})=\frac{1}{(2\pi)^{D/2}\,|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)^T\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)\right\} \tag{2}$$

其中 $\vec{\mu}_i$ 是平均向量，Σ_i 是共變異矩陣。高斯混合模型的參數可以用 λ 表示如下：

$$\lambda = \{w_i, \vec{\mu}_i, \Sigma_i\} \quad i = 1,\ldots,M \tag{3}$$

至於語者模型訓練的過程，簡單地陳述如下：

a. 初始值設定：將每一位語者其訓練語料之特徵向量，所計算出來的平均向量及共變異矩陣作為高斯混合模型中第一個高斯分佈的參數，其混合權重值則設定為 1.0。

b. 若模型中高斯分佈的總數小於系統所設定的混合數，則進行高斯分佈的分割程序。本系統選取混合權重值最高的高斯分佈來進行分割。

c. 藉由期望值最大化(Expectation Maximization, EM) [13]演算法去重新估測模型參數包括權重、平均向量及共變異矩陣至系統所設定的迭代次數為止。

d. 重複 b、c 步驟，直到模型中高斯分佈的混合數到達系統所設定的混合數為止。

期望值最大化演算法的重估公式如下：

$$p(i \mid \vec{x}_t, \lambda) = \frac{w_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} w_k b_k(\vec{x}_t)} \tag{4}$$

$$w_i = \frac{1}{T}\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda) \tag{5}$$

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)\vec{x}_t}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)} \tag{6}$$

$$\bar{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)x_t^2}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)} - \bar{\mu}_i^2 \tag{7}$$

根據訓練好的高斯混合模型λ，我們對於任何不確定的語音觀測序列X可以計算$P(X|\lambda)$。基於高斯混合模型就可以發展出不同的語者驗證方法，例如 Resenberg 等人提出群正規化計分方法[14-16]，事先由非宣告語者的語料所訓練出來的仿冒者模型，亦稱為反語者模型(Anti-Speaker Model)。或是從全部註冊語者的語料訓練出一個共有的模型，稱為全域語者模型(Global Speaker Model)。反語者模型或全域語者模型可通稱為背景語者模型，而根據正規化計分法，可以使用宣告語者模型機率之對數值和背景語者模型機率之對數值相減，作為一決策變數。如果其值大於一門檻值 θ 則接受為宣告語者。語者驗證的計算公式如下：

$$S(X \mid k) = \log\left[p(X \mid \lambda_k)\right] - \log\left[p(X \mid \bar{\lambda}_k)\right] \underset{\text{拒絕}}{\overset{\text{接受}}{\gtrless}} \theta \tag{8}$$

$$\log p(X \mid \bar{\lambda}_k) = \log\left\{\frac{1}{B}\sum_{b=1}^{B} p(X \mid \lambda_b)\right\} \tag{9}$$

$$\log p(X \mid \lambda_k) = \frac{1}{T}\sum_{t=1}^{T} \log p(\vec{x}_t \mid \lambda_k) \tag{10}$$

我們稱公式(8)為群正規化計分函數，其中 $\bar{\lambda}_k$ 代表宣告語者 k 的反語者模型，B 則為背

景語者人數。背景語者所提供的相似度正規化可以拉大宣告語者與仿冒語者之相似度值，使得門檻值能夠較容易被設定。由公式(8)計算所得的 $S(X|k)$值再與門檻值 θ 做比較，判斷其是否為宣告語者。

(一)反語者模型選擇方法

在此我們是依據語者模型距離測量的方法，找出語者在語者模型資料庫中的同質語者集合(Cohort Speaker Set)。距離測量的方法則是採取 Bhattacharyya 距離來量測聲學模型之間的距離。假設我們給定兩個高斯分佈，$G_1=G(\mu_1;\Sigma_1)$及 $G_2=G(\mu_2;\Sigma_2)$，則兩個高斯分佈之間的 Bhattacharyya 距離，計算方式就如下式所示：

$$D_{BA}(G_1,G_2)=\frac{1}{8}(\mu_1-\mu_2)\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}(\mu_1-\mu_2)^t+\frac{1}{2}\ln\frac{\left|\frac{1}{2}(\Sigma_1+\Sigma_2)\right|}{|\Sigma_1|^{\frac{1}{2}}|\Sigma_2|^{\frac{1}{2}}} \tag{11}$$

因為每個 GMM 模型包含了多個高斯混合，所以要計算兩個語者 GMM 模型的距離，可以對兩群 GMM 模型間的兩兩高斯距離($D_{BA}(G_1,G_2)$)計算其加權和。GMM 模型可以用來衡量兩語者語音的相似度，在後面的實驗中我們將使用它來挑選和宣告語者聲音最接近的語者，作為背景語者。

(二)驗證效能計算

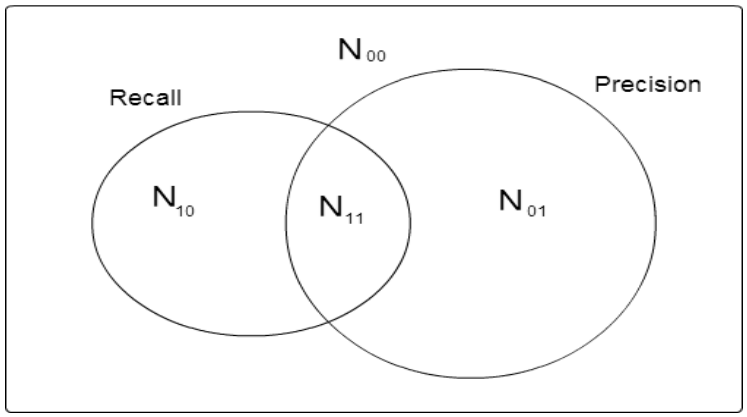
根據語者驗證的驗證流程，驗證的結果可以分成四大類，如圖三所示。其中 N_{10} 代表宣告語者的語音被系統拒絕的個數(錯誤拒絕)，N_{11} 為宣告語者的語音被系統接受的個數，N_{01} 則是非宣告語者的語音被誤認為宣告語者而接受的個數(錯誤接受)，N_{00} 指的是非宣告語者的語音被系統正確拒絕的個數。根據 N_{10}、N_{11}、N_{01}、N_{00} 四個統計值，我們可以分別計算精確率(Precision)、召回率(Recall)以及 F 度量(F-Measure)如下：

$$precision=\frac{N_{11}}{N_{11}+N_{01}} \tag{12}$$

$$recall=\frac{N_{11}}{N_{11}+N_{10}} \tag{13}$$

$$F=\frac{2\times precision\times recall}{precision+recall} \tag{14}$$

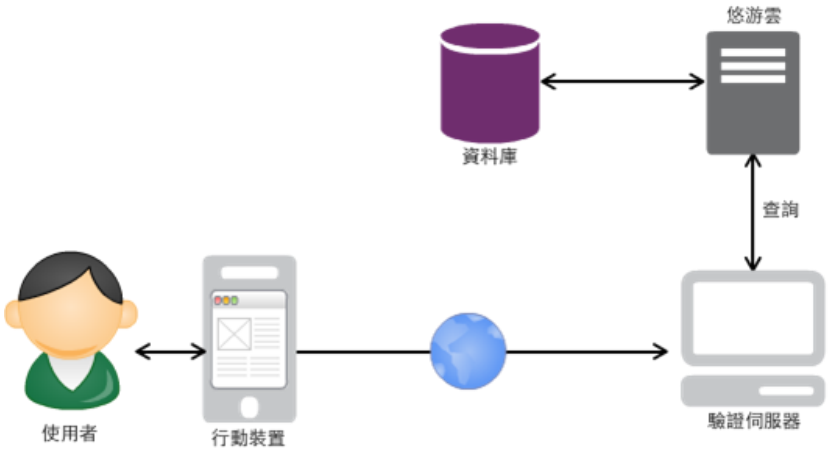
在語者驗證系統中，門檻值 θ 的值會影響系統的效能。門檻值設定的過高，容易使真實語者被系統拒絕，使得錯誤拒絕率(False Rejection Rate)提高；門檻值設定的過低，會使仿冒者容易被系統誤判為宣告語者，使得錯誤接受率(False Acceptance Rate)上升。由於此兩難狀況，在系統調校的時候通常會使用折衷的效能指標 F 值進行最佳化，會找到使 F 值最大的門檻值 θ 作為最終系統設定的門檻值。



圖三、Precision 與 Recall 的計算

三、分散式架構

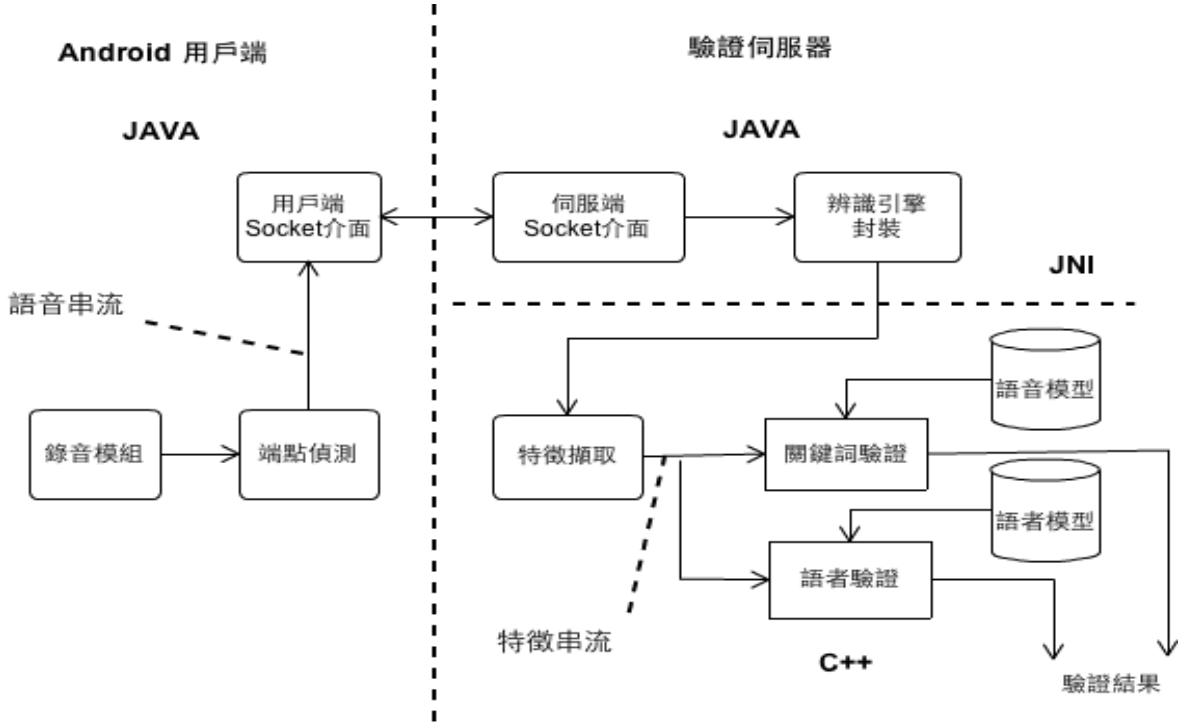
在網際網路環境下，使用者可能從任意地點、任意裝置存取資訊，因此驗證系統須對各種行動裝置提供驗證的功能。然而各種行動裝置的計算能力不同，語音辨識及語者驗證的演算法需要龐大的計算量，不容易在所有的裝置上做到即時性，因此一個分散式的系統架構具有解決此問題的潛在能力。所以，我們以用戶端及伺服端的架構來設計一個分散式的語音驗證系統，其硬體架構如圖四所示。用戶端為使用者的手持行動裝置運行 Android 作業系統；伺服端包含了兩個實體，一個是驗證伺服器，運行的是 Windows 7 作業系統；另一個是悠遊雲伺服器，提供使用者資料庫(MySQL)的存取，運行的是 Linux Fedora 作業系統。



圖四、硬體架構圖

核心的驗證模組是以分散式的架構分別實現於 Android 用戶端以及驗證伺服器，如圖五所示。用戶端使用麥克風錄製語音並進行端點偵測，並將錄製到的語音串流透過 Socket

模組即時同步傳送至驗證伺服端。端點偵測主要使用能量及過零點率作為端點偵測的特徵。伺服端在接收到語音串流後會進行即時同步的特徵計算以及辨識搜尋。在辨識伺服器的實作上，由於辨識與驗證計算複雜度較高，所以特徵擷取與辨識的核心是以 C++實作，以達到高效率的即時辨識。而 Socket 接收語音串流的部分則是以 Java 語言實作，透過 Java 原生介面(Java Native Interface, JNI)規範呼叫 C++編譯之驗證引擎原生碼。



圖五、語音驗證模組的分散式架構圖

四、語者模型實驗

為了訓練較佳的語者驗證模型，以提升平台驗證效能，我們以 TCC-300 語料庫進行實驗來調校訓練的程序以及參數設定。首先，我們針對 GMM 模型所使用的混合數、以及訓練的迭代次數進行基礎實驗。在以下實驗中若未特別提及，均是使用所有非宣告語者(共 102 位)的模型組成反語者模型，而使用模型機率的算術平均作為反模型機率。我們從 TCC-300 語音資料庫中選取 103 位語者所錄製之短句語料作為實驗所使用的訓練及測試語料。其中男性語者有 51 位，女性語者有 52 位，平均每位語者有 65 句語料。每位語者的語料中取 90%語料(總長度平均約 115 sec)來訓練模型，其餘 10%(總長度平均約 15 sec)則作為測試語料。TCC-300 語音資料庫的取樣頻率為 16kHz，資料型態為 16-bit Wav 格式。

(一)混合數的實驗

本實驗主要的目的是找出語者驗證系統所需要使用之高斯混合模型其混合數的適當值。當模型之迭代次數設定為 10 時，其實驗的結果如表一所示。我們可以發現：當 GMM

模型之混合數愈多時，其 F 值也會隨之上升。由此我們可以得知：混合數愈多愈能描述出語者的發聲特性，但是相對地計算量以及所需時間也會增加；混合數較少時，系統效能就會下降。例如混合數為 5 時，其 F 值僅有 0.9091；若混合數到達 15 時，則 F 值可以到達 0.9868；而當混合數超過 15 時，系統效能則呈現飽和狀態。雖然當混合數到達 60 時，其 F 值可以到達 0.9993，但是因為我們使用了 102 位背景語者來做測試，因此若使用 60 個 Mixture 會使得對高斯混合模型計算 Likelihood 時的計算量達到 6120 個高斯分佈，計算的時間將會大幅增加，高斯混合模型的數目也已經超過了語音辨識模型中所使用的個數。在分散式的計算架構下，這樣的複雜度雖然仍然可以做到即時辨識，但是為了減少計算的負載量，我們選擇 Mixture 15 來進行後續的實驗。因此在考量兼顧語者驗證系統反應之即時性及驗證效能的情形下，我們將本專案實際系統的模型混合數設定為 15 來訓練每一位語者之 GMM 模型。

表一、對不同混合數的效能變化

Mixture	Iteration	Recall	Precision	F Measure
5	10	0.9104	0.9078	0.9091
10	10	0.9633	0.9719	0.9676
15	10	0.9883	0.9854	0.9868
20	10	0.9927	0.9927	0.9927
25	10	0.9927	0.9927	0.9927
30	10	0.9956	0.9941	0.9949
35	10	0.9956	0.9971	0.9963
40	10	0.9985	0.9985	0.9985
45	10	0.9985	0.9985	0.9985
50	10	0.9985	0.9985	0.9985
55	10	0.9985	0.9985	0.9985
60	10	1.0000	0.9985	0.9993

(二)迭代次數的實驗

本實驗主要目的是找出語者驗證系統所需要使用之高斯混合模型在每次分裂後之迭代次數的適當參數值。當模型之混合數設定為 15 時，其實驗結果如表二所示。當迭代次數愈多時，其 F 值並無顯著變化，大約在 0.98 至 0.99 之間。又因每次迭代在訓練時都會耗費較多時間，所以本專案在訓練模型時，將迭代次數參數固定為 10。

表二、對不同迭代次數的效能變化

Mixture	Iteration	Recall	Precision	F Measure
15	5	0.9868	0.9796	0.9832
15	10	0.9883	0.9854	0.9868
15	15	0.9868	0.9868	0.9868
15	20	0.9868	0.9810	0.9839
15	25	0.9883	0.9854	0.9868
15	30	0.9883	0.9825	0.9854
15	35	0.9868	0.9912	0.9890
15	40	0.9883	0.9926	0.9904
15	45	0.9897	0.9868	0.9883
15	50	0.9897	0.9839	0.9868

(三)背景語者人數與挑選方式之實驗

為了瞭解背景語者數目以及篩選方式對於驗證效能的影響，我們使用了兩種挑選背景語者的方式，一種是根據 Bhattacharyya 距離找出和宣告語者最相近的 N 名語者做為反語者模型，另一種則是以隨機方式挑選。我們將模型之混合數設定為 15、迭代次數設定為 10，背景語者的數目 N 從 10 遞增至 50 時，F 度量的實驗結果如表三所示。由表三可以看出，當背景語者人數愈多時，其 F 值也會隨之上升。由此我們可以得知：背景語者人數愈多愈能夠增加反語者模型的鑑別度，提升驗證之效能，但是相對地計算量及所需時間也會增加。若是背景語者的人數選取低至 10 人，則以 Bhattacharyya 距離(表三中標記為 B-Distance)選取背景語者的方式會降至 0.9038，但很明顯地仍然比以隨機的方式來挑選反語者的效能為佳。這顯示了 Bhattacharyya 距離能夠正確地找出和宣告語者相近的語者，而訓練出較具有鑑別力的決策邊界函數，其原理類似支撐向量機分類器。考量系統實際運作時，註冊人數可能會隨著系統使用時間遞增；若語者總數很大時，勢必無法將所有語者模型都用來計算反語者模型機率。此時，篩選背景語者就是讓速度與效能達到折衷的可行策略。本實驗也同時驗證了以 Bhattacharyya 距離加權方式計算兩個高斯混合模型距離的適切性。

表三、以不同方式來挑選背景語者模型

背景語者人數	B-Distance	Random
10	0.9038	0.7310
20	0.9398	0.8854
30	0.9556	0.8866
40	0.9582	0.9137
50	0.9716	0.9258

註：B-Distance 為在挑選背景語者時不分男女性別，而以距離宣告語者 GMM 模型最近之 N 位語者模型作為其背景語者模型，N 則為其背景語者人數。Random 為在挑選背景語者時不分男女性別，而以隨機的方式挑選 N 位語者模型作為其背景語者模型。

(四)以性別來區分背景語者之實驗

過去在語音辨識上使用和性別相關的語音模型，對於語音辨識的效能均能夠產生提升的效果。我們想探究在語者識別上使用性別相關的模型對於驗證效能是否也能夠有所幫助，因此將男女的語料分開，僅從與測試語者性別相同的語者中挑選背景語者來進行實驗。我們將模型之混合數設定為 15、迭代次數設定為 10，背景語者的數目 N 從 10 遞增至 50，篩選背景語者的方式則是採用 Bhattacharyya 距離來篩選。在篩選背景語者時，我們會以測試語者的性別(假定系統已預先判斷性別)來選取性別相同的背景語者，或從全部語者中篩選背景語者來進行實驗，實驗結果如表四所示。由表四結果可以看出，在同樣背景語者人數下，以同性別語者來選取反語者模型會比從所有語者篩選為佳；這顯示了同性別的背景語者更能正確地區別宣告語者與近似語者，亦即性別相關的模型可以有較佳的鑑別力。然而我們必須注意到，此驗證效能的提升是在假定系統已正確辨別語者性別的條件下而達成，系統的設計中必須增加可靠的性別決策模組。

表四、是否使用性別相關的背景語者模型

背景語者人數	性別相關	全部語者
10	0.9132	0.9038
20	0.9504	0.9398
30	0.9632	0.9556
40	0.9753	0.9582
50	0.9875	0.9716

(五)測試語音資料時間長度之實驗

於上述的各項實驗中，每一位語者都使用了其全部的註冊語音資料來訓練其語者模型，全部的測試語音資料來進行驗證測試。為了瞭解測試語音資料的時間長度對於驗證效能會產生什麼樣的影響，因此我們將模型之混合數設定為 15、迭代次數設定為 10，每一位語者使用其全部的註冊語音資料來訓練其語者模型，而只使用部份的測試語音資料來進行驗證測試時，其實驗的結果如表五所示。我們可以發現：當測試語音資料之時間長度愈長時，其 F 值也會隨之上升。由此我們可以得知：測試語音資料之時間長度愈長愈能描述出語者的發聲特性，但是相對地計算量以及所需時間也會增加；測試語音資料之時間長度較短時，系統效能就會下降。例如當測試語音資料之時間長度為 10 個音框(約 0.1 sec)時，其 F 值僅有 0.8407；若測試語音資料之時間長度到達 100 個音框(約 1 sec)時，則 F 值可以到達 0.9787；而當測試語音資料之時間長度超過 150 個音框(約 1.5 sec)時，系統效能則呈現飽和狀態。因此在考量兼顧語者驗證系統反應之即時性及驗證效能的情形下，我們將本專案實際系統的測試語音資料其時間長度設定為 100 個音框(約 1 sec)來進行每一位語者之驗證測試。

表五、對不同測試語音資料時間長度的效能變化

Testing Data Frame	Recall	Precision	F Measure
10	0.8253	0.8567	0.8407
20	0.9134	0.9311	0.9222
30	0.9339	0.9651	0.9493
40	0.9471	0.9743	0.9605
50	0.9633	0.9676	0.9654
60	0.9486	0.9878	0.9678
70	0.9677	0.9720	0.9698
80	0.9780	0.9638	0.9708
90	0.9794	0.9709	0.9751
100	0.9765	0.9808	0.9787
150	0.9853	0.9824	0.9839
200	0.9897	0.9839	0.9868
250	0.9868	0.9839	0.9853
300	0.9883	0.9854	0.9868

註：

(1) Testing Data Frame 中的數值所代表的是於每一筆測試語音資料中從頭開始擷取的音框數。若該筆測試語音資料的音框數不足(測試語音資料的音框總數 ＜ Testing Data Frame 中所設定的音框數)，則以該筆測試語音資料中全部的資料來進行驗證測試。

(2) 根據本專案系統中語音資料取樣時所採取的音框擷取方式，100 個音框約等於 1 sec 的時間。

(六)註冊語音資料(訓練資料)時間長度之實驗

從上述的實驗中我們不難發現：當測試語音資料之時間長度設定為 100 個音框(約 1 sec)時，系統即可呈現出不錯的驗證效能。因而使我們想要進一步地探討：當我們開始減少註冊語音資料的時間長度時，會對系統效能造成什麼樣的影響？為了瞭解註冊語音資料的時間長度對於驗證效能會產生什麼樣的影響，因此我們將模型之混合數設定為 15、迭代次數設定為 10，每一位語者只使用部份的註冊語音資料來訓練其語者模型，而且也只有使用一部份的測試語音資料(100 個音框，約 1 sec)來進行驗證測試時，其實驗的結果如表六所示。我們可以發現：當註冊語音資料之時間長度愈長時，其 F 值也會隨之上升。由此我們可以得知：訓練語音資料之時間長度愈長愈能描述出語者的發聲特性，但是相對地計算量以及所需時間也會增加；訓練語音資料之時間長度較短時，系統效能就會下降。例如當訓練語音資料之時間長度為 1000 個音框(約 10 sec)時，其 F 值僅有 0.8415；若訓練語音資料之時間長度到達 6000 個音框(約 60 sec)時，則 F 值可以到達 0.9427；而當訓練語音資料之時間長度超過 12000 個音框(約 120 sec)時，系統效能則呈

現飽和狀態。因此在考量兼顧語者驗證系統反應之即時性，以及可以透過系統多重驗證機制中其他不同的驗證方式來互補其效能的情形下，我們將本專案實際系統的註冊語音資料其時間長度設定為 6000 個音框(約 60 sec)來訓練每一位語者之 GMM 模型。

表六、對不同訓練語音資料時間長度的效能變化

Training Data Frame	Recall	Precision	F Measure
1000	0.7915	0.8983	0.8415
3000	0.9222	0.9235	0.9229
6000	0.9178	0.9690	0.9427
9000	0.9530	0.9701	0.9615
12000	0.9765	0.9779	0.9772
16000	0.9765	0.9794	0.9779
18000	0.9765	0.9808	0.9787

註：

(1) Training Data Frame 中的數值所代表的是每一位語者於其註冊語音資料中從頭開始擷取的音框數。若該位語者註冊的語音資料音框數不足(註冊語音資料的音框總數 ＜ Training Data Frame 中所設定的音框數)，則以該位語者所註冊全部的語音資料來訓練其 GMM 模型。

(2) 每一筆測試語音資料從頭開始擷取 100 個音框。若該筆測試語音資料的音框數不足(測試語音資料的音框總數 ＜ 100 個音框)，則以該筆測試語音資料中全部的資料來進行驗證測試。

(3) 根據本專案系統中語音資料取樣時所採取的音框擷取方式，100 個音框約等於 1 sec 的時間。

五、結論

在語音辨識和驗證技術逐漸成熟以及個人行動裝置快速普及下，如何將結合語音相關技術應用在身份認證系統，以改善使用者認證服務的速度及流程，越來越受到重視。關鍵詞擷取技術與語者驗證技術可分別使用 What one knows 以及 Who one is 的驗證方法來進行驗證。兩者的結合可以提高身份驗證的可靠度。本計畫成功地結合了上述兩種驗證方法，在一個分散式的網路環境中達到了即時多重驗證，我們並製作了一個技術展示系統，系統可以透過使用者輸入的地址資料來進行驗證，如果使用者不知道正確的地址、或其語音非宣稱者本人的語音，驗證系統均可能加以拒絕，因此可增加驗證系統的可靠性。此外此驗證方式具有彈性，未來可進一步結合 RFID 驗證方式，提供雲端服務更有彈性的認證方式。例如雲端服務中，查詢個資、修改密碼或進行交易等會有不同安全等級需求，驗證系統可以多次詢問或多重驗證的方式來提升安全性。我們相信這是一個未來應用的重要趨勢。

為了達到較佳的驗證效果，我們也作了一系列的實驗來調校系統的參數，包括混合數、

迭代次數、背景語者人數、背景語者挑選方式、是否使用性別相關模型、測試語音資料及註冊語音資料之時間長度等。實驗結果顯示當混合數為 15 以上、迭代次數為 10 以上時就可以達到穩定的效能；背景語者人數上升時效能可以持續提升，當測試語音資料或註冊語音資料之時間長度增加時，也可以產生相同的效果，但是計算量會增加，因此應考慮伺服器的計算負載是否過重因而影響辨識速度；背景語者的挑選方式則顯示使用 Bhattacharyya 距離挑選和宣稱語者最相近的語者作為背景語者遠較隨機方式挑選為佳；性別相關的實驗則顯示只從與宣稱語者同性別的語者中挑選背景語者，會比從所有的語者中挑選為佳，也就是性別相關的語者模型具有較佳的鑑別力。根據實驗的結果，我們使用較佳的訓練流程和參數設定來訓練模型，並應用在我們的展示系統中。未來希望將雲端語音驗證包裝成為網路服務，以提供用戶在網路環境中方便、快速、彈性、可靠地身份認證服務。

致謝

本研究承蒙國科會專題研究計畫「為雲端服務而設計之智慧終端應用安全套件」的部份經費補助，方得以完成本研究，謹此致謝。

參考文獻

[1] 吳金池，“語者辨識系統之研究”，國立中央大學電機工程研究所碩士論文，2002 年。

[2] 謝忠穎，“正交式高斯混合模型之語者驗證系統”，中興大學電機工程學研究所碩士論文，2009 年。

[3] 趙怡翔，“鑑別式訓練法於語者驗證之研究”，交通大學資訊科學與工程研究所博士論文，2009 年。

[4] 蔡仲齡，“含語者驗證之小型場所人臉辨識門禁系統之研發”，國立成功大學工程科學研究所碩士論文，2008 年。

[5] Reynolds Douglas A.,“An Overview of Automatic Speaker Recognition Technologies”, ICASSP, p. 4072-4075, 2002.

[6] Campbell Joseph P.,“Speaker Recognition: A Tutorial”, Vol. 85, No. 9, Proceedings of the IEEE, 1997.

[7] Huang, X., A. Acero, and H. W. Hon, Spoken Language Processing, Prentics Hall, New Jersey, 2001

[8] R. Vergin and D. O’Shaughnessy and A. Farhat,“Generalized Mel Frequency Coefficients for Large-Vocabulary Speaker-Independent Continuous-Speech Recognition”, IEEE Trans. on Speech and Audio Processing, Vol. 7, No. 5, pp. 525-532, September 1999.

[9] Jiang H.,“Confidence Measures for Speech Recognition: A Survey”, Speech Communication, 2005.

[10] Campbell W. M., etc,“Support Vector Machines Using GMM Supervectors for Speaker Verification”, IEEE Signal Processing Letters, Vol. 13, No.5, 2006.

[11] 黃冠達，“應用支撐向量機於中文關鍵詞驗證之研究”，國立臺灣科技大學資訊管理研究所碩士論文，2007 年。

[12] Leggetter C. J. and Woodland P. C.,“Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models”, Computer Speech and Language, p. 171-185, 1995.

[13] T. K. Moon,“The Expectation-Maximization Algorithm”, IEEE Signal Processing Magazine, Vol. 13, No. 6, pp. 47-60, November 1996.

[14] A. E. Rosenberg, J. Delong, C. H. Lee, B. H. Juang and F. K. Soong,“The Use of Cohort Normalized Scores for Speaker Recognition”, Pro. ICSL 92. Banff, pp. 599-602. Oct. 1992.

[15] Chi-Shi Liu, Hsiao-Chuan Wang and Chin-Hui Lee,“Speaker Verification Using Normalized Log-Likelihood Score”, IEEE Trans. on Speech and Audio Processing, pp. 57-60, Jan. 1996.

[16] 游智翔，“整合高斯混合與具性能指標支撐向量機模型之語者確認研究”，國立中央大學電機工程研究所碩士論文，2008 年。

[17] Mika, S., G. Rätsch, J. Weston, B. Schölkopf, and K. R. Müller,“Fisher Discriminant Analysis with Kernels”, in Proc. Neural Networks for Signal Processing IX, Madison, WI, USA, pp. 41-48, 1999.