公告本

申請日期	P0.P.7
案　　號	P01777462
類　　別	G10L 15/00

（以上各欄由本局填註）

A4
C4

548630

發明 ~~新型~~ 專利說明書

一、發明 ~~新型~~ 名稱	中　文	利用映射以用於自動語音辨識之系統及方法
	英　文	"SYSTEM AND METHOD FOR AUTOMATIC VOICE RECOGNITION USING MAPPING"
二、發明 ~~創作~~ 人	姓　名	1. 奎英揚 YINGYONG QI 2. 哈里納斯　格魯達迪 HARINATH GARUDADRI 3. 寧畢 NING BI
	國　籍	1. 美國 2. 加拿大 3. 中國
	住、居所	1. 美國加州聖地牙哥市梅克斯特巷6609號 2. 美國加州聖地牙哥市歐維多街9435號 3. 美國加州聖地牙哥市布魯茲維廣場14209號
三、申請人	姓　名 （名　稱）	美商奎康公司 QUALCOMM INCORPORATED
	國　籍	美國
	住、居所 （事務所）	美國加州聖地牙哥市摩豪斯大道5775號
	代表人 姓　名	菲力普 R. 華德渥斯 PHILIP R. WADSWORTH

（由本局填寫）

承辦人代碼：	A6 B6
大　　類：	
ＩＰＣ分類：	

本案已向：

國（地區）	申請專利，申請日期：	案號：	，□有 □無主張優先權
美國	2000 年 09 月 08 日	09/657,760	☑有 □無主張優先權

有關微生物已寄存於：　　，寄存日期：　　，寄存號碼：

（請先閱讀背面之注意事項再填寫本頁各欄）

裝 訂 線

五、發明說明（ 1 ）

發明背景

I. 領域

本發明一般而言係關於通訊範疇，更特定而言係關於一種創新及改良的語音辨識系統與方法。

II. 背景

語音辨識(VR)代表了最重要的技術之一來賦予具有模擬智慧的機器來辨識使用者或使用者語音指令，並有利於人機介面。VR亦代表瞭解人類語言的一關鍵技術。利用技術來由一語音信號回復成一語言訊息的系統稱之爲語音辨識器。此處所使用的該名詞"語音辨識器"通常代表任何說話致動的使用者介面裝置。

該VR的使用(也通常稱之爲說話辨識)爲了安全性理由也成爲日益重要。舉例而言，VR可用來取代一無線電話鍵盤上的按鈕之人工操作。此在當一使用者在駕車時要打電話時，特別地重要。當使用沒有VR的電話時，該駕駛者必須由駕駛盤上移開一隻手，並在按壓按鈕來撥打電話時注視該電話鍵盤。這些動作增加了一車禍的可能性。一說話啓動電話(即設計有語音辨識的電話)將允許駕駛者來撥打電話而可持續地注視路面。此外，一車用手持聽筒系統可允許駕駛者在打電話時雙手都保持在架駛盤上。

語音辨識裝置可分類爲說話者相關(SD)或說話者無關(SI)裝置。說話者相關裝置比較常見，其被訓練辨識來自特殊使用者的指令。相反地，說話者無關裝置能夠接受來自任何使用者的語音指令。爲了增加一給定VR系統的效能，

五、發明說明（ 2 ）

不論是說話者相關或說話者無關，皆需要訓練來使該系統具有正確的參數。換言之，該系統需要在其能夠最佳化地操作之前進行學習。

一說話者相關 VR 裝置基本上是以兩個階段來操作，一訓練階段及一辨識階段。在訓練階段，該 VR 系統提示該使用者講出在系統詞彙中的單字一次或兩次(基本上爲兩次)，所以該系統能夠學習到這些特殊單字或片語的該使用者說話之特性。一車用免持聽筒的範例詞彙可包含鍵盤上的數字；關鍵字"打電話"，"傳送"，"撥號"，"取消"，"清除"，"加入"，"刪除"，"歷史"，"程式"，"是"及"否"；一預定數目的名字通常稱之爲同事，朋友或家庭成員。一旦完成訓練，在辨識階段中該使用者可藉由說明該訓練過的關鍵字來啓始打電話，該 VR 即藉由比較該說出的發音與先前訓練的發音(儲存成樣板)及採取最佳符合者來辨識。舉例而言，如果名字"John"爲該訓練的名字之一，使用者可說出詞句"Call John"來打電話給 John。該 VR 系統將可辨識單字"打電話"及"John"，並將使用者先前已輸入成 John 的電話號碼之來撥號。訓練的系統及方法。

一說話者無關 VR 裝置也使用一訓練樣板，其包含一預定大小的預錄詞彙(例如某些控制單字，0 到 9 的數字，及是與否)。大量的說話者(例如 100)必須錄製其說出每個詞彙中的單字。

不同的說話者無關 VR 裝置可產生不同的結果。舉例而言，一說話者無關(SI)隱藏 Markov 模型(HMM)引擎可產生

五、發明說明（ 3 ）

與一說話者無關動態時間扭曲(DTW)引擎不一樣的結果。結合兩種引擎的結果可使得一系統比僅使用一種引擎的結果更具有較佳的辨識準確性及較低的拒絕率。

一說話者相關VR及一說話者無關VR可產生不同的結果。一說話者相關引擎使用關於一特定使用者的樣板來執行辨識。一說話者無關引擎使用來自一使用者集合的範本產生的樣板來執行辨識。因爲說話者特定樣板較接近於一給定使用者的說話樣式，SD引擎提供比SI引擎要佳的準確性。但是，SI引擎的好處是使用者在使用該系統之前不需要經過"訓練過程"。

其需要一種結合不同形式的引擎之系統與方法。結合多種引擎可提供增強的準確性，並使用大量的資訊在該輸入語音信號。一種結合VR引擎的方法揭示於美國專利申請編號09/618,177，名爲"語音辨識的結合引擎系統與方法"("Combined Engine System and Method for Voice Recognition")，於2000年7月18日提出，其授權給本發明的授讓人在此引用做爲參考。

一決策邏輯VR系統可使用探索邏輯來設計決策規則。該決策邏輯基本上以每個引擎的測試發音及最上端候選者(單字)樣板之間的測定距離開始。舉例而言，假設使用兩個引擎(引擎D及H)。d_1及d_2代表測試發音與引擎D的最上端兩個候選者單字之間的距離，而h_1及h_2代表測試發音與引擎H的最上端兩個候選單字之間的距離。d_g及h_g分別代表該測試發音與引擎D及H的"垃圾"樣板之間的

五、發明說明（ 4 ）

距離。該垃圾樣板用來代表所有不在該詞彙中的單字。該決策邏輯包含在這些測定的距離與一組預先定義的臨界值的一系列比較。但是，該比較規則及臨界值需要以試誤法爲基礎來分析並調整，部份因爲其不能夠被系統化地最佳化。此過程非常耗時且困難。此外，該探索規則可以是與應用相關的。舉例而言，一新的規則組合需要被分析，如果對每個引擎使用上端 3 個單字而非上端 2 個單字時。其有可能該組辨識無雜訊語音的規則將與那些辨識有雜訊語音的規則會不相同。

因此，其需要一種解決來自複數個不同 VR 引擎的不同結果之系統與方法。

發明概要

上述的具體實施例係提出語音辨識的系統與方法。在一具體實施例中，其提供一種結合複數個語音辨識引擎來改善語音辨識之方法。該方法較佳地是包含耦合複數個語音辨識引擎到一映射模組。每個 VR 引擎產生一假定，即單字候選者，然後該映射模組應用一映射函數來由該複數個 VR引擎產生的假定中選擇一假定。

在一具體實施例中，說話者無關語音辨識引擎被結合。在另一具體實施例中，說話者相關語音辨識引擎被組合。在又另一具體實施例中，一說話者無關語音辨識引擎結合於一說話者相關語音辨識引擎。

在一具體實施例中，一說話者無關語音辨識引擎爲一動態時間扭曲語音辨識引擎。在一具體實施例中，一說話者

五、發明說明（ 5 ）

無關語音辨識引擎爲一隱藏 Markov模型。在一具體實施例中，一說話者相關語音辨識引擎爲一動態時間扭曲語音辨識引擎。在一具體實施例中，一說話者相關語音辨識引擎爲一隱藏 Markov模型。

圖式簡單說明

本發明的特徵，目的及好處將藉由以下的詳細說明，並參考所附圖面，即可更爲瞭解，其中相同的參考符號在所有圖面皆相對應，其中：

圖 1 所示爲具有三種語音辨識引擎的一語音辨識系統的具體實施例；

圖 2 所示爲包含一 DTW 引擎及一 HMM 引擎的一語音辨識系統；及

圖 3 所示爲具有兩個語音辨識引擎的語音辨識系統之具體實施例。

發明詳細說明

在一具體實施例中，如圖 1 所示的語音辨識系統 100 具有三種語音辨識引擎，其能夠隔離的單字辨識工作：一動態時間扭曲說話者無關(DTW-SI)引擎 104，一動態時間扭曲說話者相關(DTW-SD)引擎 106，及一隱藏 Markov 模型(HMM)引擎 108。這些引擎係用於指令單字辨識及數字辨識來提供豐富的說話式使用者介面給一掌上型裝置執行的日常工作，例如像是一行動電話，個人數位助理(PDA)等。在另一具體實施例中，該語音辨識系統 100 包含一 DTW-SI 104 及一 DTW-SD 引擎 106。在又另一具體實施例

五、發明說明（ 6 ）

中，該語音辨識系統 100 包含一 DTW-SI 引擎 104 及一 HMM 引擎 108。在又另一具體實施例中，該語音辨識系統 100 包含一 DTW-SD 引擎 106 及一 HMM 引擎 108。在一具體實施例中，該 HMM 引擎 108 爲說話者無關。在另一具體實施例中，該 HMM 引擎 108 爲說話者相關。本技藝的專業人士將可瞭解到，其可使用任何在本技藝中已知的 VR 引擎。在又另一具體實施例中，結合複數個其它 VR 引擎形式。對於本技藝的專業人士亦可瞭解該引擎可用任何組合來建構。

根據一具體實施例，如圖 1 所示，一語音辨識系統 100 包含一類比到數位轉換器(A/D) 102，一 DTW-SI 引擎 104，一 DTW-SD 引擎 106 及一 HMM 引擎 108。在一具體實施例中，該 A/D 102 爲一硬體 A/D。在另一具體實施例中，該 A/D 102 係以軟體實施。在一具體實施例中，該 A/D 102 及該引擎 104，106，108 係實施爲一個裝置。本技藝的專業人士可瞭解到，該 A/D 102 及引擎 104，106，108 可實施及分佈在任何數目的裝置之間。

該 A/D 102 係耦合於該 DTW-SI 引擎 104，該 DTW-SD 引擎 106 及該 HMM 引擎 108。該 DTW-SI 引擎 104，該 DTW-SD 引擎 106 及該 HMM 引擎 108 係耦合於一映射模組 110。該映射模組採用引擎 104，106，108 的輸出爲其輸入，並產生對應於一語音信號 s(t) 的單字。

該語音辨識系統 100 可存在於像識一無線電話或一車用免持聽筒。一使用者(未示出)說出一單字或片語，產生一

五、發明說明（ 7 ）

語音信號。該語音信號以一習用的換能器(未示出)，被轉換到一電子語音信號 s(t)。該語音信號 s(t)被提供給 A/D 102 ，其根據一已知的取樣方法來轉換該語音信號到一數位化語音樣本，例如像是脈衝編碼調變(PCM)， A-法則或μ-法則。在一具體實施例中，基本上每秒鐘有 N 個 16 位元語音樣本。因此，對於 8,000 Hz 的取樣頻率，N=8,000 ，對於 16,000 Hz 的取樣頻率， N=16,000 。

該語音樣本係提供給該 DTW-SI 引擎 104 ，該 DTW-SD 引擎 106 及該 HMM 引擎 108 。每個引擎處理該語音樣本，並產生假定，即該語音信號 s(t)的候選單字。然後該映射模組映射該候選單字到一決策空間，其被評估來選擇最可反應該語音信號 s(t)的候選單字。

在一具體實施例中，該語音辨識系統包含兩個 VR 引擎，如圖 2 所示。該語音辨識系統 100 包含一 DTW 引擎 112 及一 HMM 引擎 114 。在一具體實施例中，該 DTW 引擎爲一說話者無關 VR 引擎。在另一具體實施例中，該 HMM 爲一說話者相關 VR 引擎。在一具體實施例中，該 HMM 引擎爲一說話者無關 VR 引擎。在另一具體實施例中，該 HMM 引擎爲一說話者相關 VR 引擎。

在這些具體實施例中，該系統同時具有 DTW 及 HMM 的好處。在一具體實施例中， DTW 及 HMM 樣板係在一訓練階段期間明確地產生，其中該語音辨識系統被訓練來辨識輸入語音信號。在另一具體實施例中， DTW 及 HMM 樣板係在該語音辨識系統的典型使用期間隱含地產生。範例性

五、發明說明（ 8 ）

訓練系統及方法係描述在美國專利申請編號 09/248,513 ，名爲"語音辨識拒絕方案"("VOICE RECOGNITION REJECTION SCHEME")，其於 1999 年 2 月 8 日立案，其授權給本發明的受讓人，在此完全引用做爲參考，及美國專利申請編號 09/225,891 ，名爲"語音信號的分段化及辨識之系統及方法"("SYSTEM AND METHOD FOR SEGMENTATION AND RECOGNITION OF SPEECH SIGNALS")，其於 1999 年 1 月 4 日立案，其授權給本發明的受讓人，在此完全引用做爲參考。

該語音辨識系統的一組所有詞彙單字的樣板係儲存在任何習用形式的非揮發性儲存媒體，例如像是快閃記憶體。此允許該樣板在關閉該語音辨識系統 100 的電源時，仍保留在該儲存媒體中。在一具體實施例中，該組樣板係以說話者無關樣板建構系統來建構。在一具體實施例中，指令字元被包含在一 VR引擎詞彙中。

該 DTW 技術在本技藝中爲熟知的，其揭示於 Lawrence Rabiner 及 Biing-Hwang Juang 所著，"語音辨識基礎"("Fundamentals of Speech Recognition")，頁 200-238 ，(1993)，其完全在此引用做爲參考。根據該 DTW 技術，一棚架藉由對於儲存在一樣板資料庫的每個發音繪製要測試的該發音的時間序列相對於一時間序列而形成。然後正要測試的發音即被每個點地(例如每 10 ms)與在樣板資料庫中的每個發音做比較，一次一個發音。對於每個在該樣板資料庫中的發音，正在測試的發音即被調整，或對時間"扭曲"

五、發明說明（ 9 ）

，其可在特殊點上被壓縮或擴張，直到達到與樣板資料庫中的發音達到最可能的匹配。在每個時間點上，該兩個發音被比較，且在該點(零成本)宣告一匹配或宣告一不匹配。在一特殊點的不匹配之事件中，正在測試的發音被壓縮，擴張，或視需要而不匹配。該處理會持續到兩個發音彼此已經完全地比較過。有可能大量(基本上成仟地)不同調整的發音。具有最低成本函數的調整之發音(即需要最小數目的壓縮及/或擴張及/或不匹配)即被選出。在類似於 Viterbi 解碼演算法之方式中，該選擇較佳地是藉由該樣板資料庫中該發音中的每個點向回看來執行，以決定具有最低整體成本的路徑。此允許決定出最低成本(即最爲接近匹配的)調整的發音，而不用依靠產生不同調整的發音之每個可能之"強力"方法。然後在該樣板資料庫中的所有發音之最低成本調整的發音即被比較，而具有最低成本者被選出，做爲最接近匹配到該測試的發音之儲存的發音。

雖然在一 DTW 引擎 104 中的 DTW 匹配方案及一 HMM 引擎 108 中的 Viterbi 解碼爲類似，該 DTW 及 HMM 引擎利用不同的前端方案，即特徵擷取器，以提供特徵向量到該匹配階段。爲此原因，該 DTW 及 HMM 引擎的錯誤樣式相當地不同。具有組合引擎的語音辨識系統可得到錯誤樣式中差異之好處。藉由適當地結合來自兩個引擎的結果，其可達到一較高的整體辨識準確率。更重要地是，其可達到所需要辨識準確性的較低拒絕率。

在一具體實施例中，操作相同詞彙組合的說話者無關語

五、發明說明（ 10 ）

音辨識引擎被結合。在另一具體實施例中，說話者相關語音辨識引擎被結合。在又另一具體實施例中，兩個引擎係運作在不同的詞彙組合上。

每個引擎產生一輸出，其爲在其詞彙中所說出的單字。每個輸出包含該輸入信號的一單字候選者。未對應於該輸入信號的單字即被拒絕。範例性拒絕方案係揭示於美國專利申請編號 09/248,513 中，其在此完全引用做爲參考。

準確的語音辨識對於一嵌入式系統相當困難，部份由於其受限的計算資源。爲了增加系統的準確性，語音辨識係使用多重辨識引擎來完成。但是，不同的 VR 引擎可產生不同的結果。舉例而言，一引擎可選擇"Jane"及"Joe"做爲最上端候選單字，然而其它的 VR 引擎可選擇"Julie"及"Joe"做爲最上端的兩個候選者。這些不同的結果需要被解決。一答案必須提出，即一候選單字需要被選出。該 VR 系統必須根據要生效的多個引擎之這些後選單字來達到一決定。

在一具體實施例中，結合了 X(X= 2, 3,...)個引擎，每個引擎產生 Y (Y= 1, 2,...)候選單字。因此，在 X*Y 個候選者中僅有一個是正確的答案。在另一具體實施例中，每個引擎可產生不同數目的候選者。

在具有兩個引擎 D 及 H 之具體實施例中，d_1及d_2代表測試發音與引擎 D 的最上端兩個候選單字之間的距離，而h_1及h_2代表測試發音與引擎 H 的最上端兩個候選單字之間的距離。變數d_g及h_g分別代表該測試發音與引擎 D 及 H

五、發明說明（ 12 ）

的距離。W_j為一組候選單字，其中索引 j 為該組編號，而 N 為組的數目。每組具有一些候選單字，該數字為一正整數。索引 i 為 VR 引擎編號。

每個 VR 引擎也產生該測試發音 T_u及該詞彙之外的單字樣板 W_g之間的距離 D_g。一詞彙中單字為在一 VR 引擎的詞彙中的單字。一詞彙外單字為不在一 VR 引擎的詞彙中單字。

如果該映射函數的結果大於一臨界值，所評估的候選單字為正確，該輸入即被接受。否則，該輸入被拒絕。

表 1 所示為具有一 DTW 引擎及一 HMM 引擎的具體實施例之距離的矩陣，其中來自每個引擎的最上端兩個單字被選擇成為該候選組合。D_1及 D_2為來自該 DTW VR 引擎的最上端兩個候選單字，而 H_1及 H_2為來自該 HMM VR 引擎的最上端兩個候選單字。

在具有兩個 VR 引擎的具體實施例中，其中一 VR 引擎產生 X 距離，而其它引擎產生 Y 距離，即產生一整體 X*Y 候選單字。

僅有來自一候選組合的一個單字將被辨識，並決定如果該辨識將被拒絕/接受時即可做出決定。在一具體實施例中，對於由該候選組合中選擇一單字，並做出決策來接受或拒絕，皆使用一線性映射函數。

每組候選單字 W_i, i=1, 2, 3, 4 ，具有其對應的測定向量，如表 1 所示。

五、發明說明（ 13 ）

表 1

W_1：	$D_1^{W_1}$	$D_2^{W_1}$	D_g	$H_1^{W_1}$	$H_2^{W_1}$	H_g
W_2：	$D_1^{W_2}$	$D_2^{W_2}$	D_g	$H_1^{W_2}$	$H_2^{W_2}$	H_g
W_3：	$D_1^{W_3}$	$D_2^{W_3}$	D_g	$H_1^{W_3}$	$H_2^{W_3}$	H_g
W_4：	$D_1^{W_4}$	$D_2^{W_4}$	D_g	$H_1^{W_4}$	$H_2^{W_4}$	H_g

D 代表一 DTW 引擎。H 代表一 HMM 引擎。$D_1^{W_i}$ 爲 T_u 與 W_i 之間的距離。$D_2^{W_i}$ 爲排除 W_i 的第二最佳候選者的距離。D_g 代表 T_u 及該垃圾樣板之間的距離。$H_1^{W_i}$，$H_2^{W_i}$，H_g 分別代表對於 DTW 引擎之相同的量。

該線性映射函數具有以下的形式：

$M_i(D,H)=C_0+c_1D_1^{W_i}+c_2D_2^{W_i}+c_3D_g+c_4H_1^{W_i}+c_5H_2^{W_i}+c_nH_g$，其中 $c_i(i=0,1,\ldots,n)$ 在一具體實施例中爲一實數常數，且在另一具體實施例中爲一語音參數。在索引 i 上的上限爲 n。該上限 n 等於該語音辨識系統中 VR 引擎的數目加上每個 VR 引擎的候選單字之數目。在每個 VR 引擎具有兩個 VR 引擎及兩個候選單字的具體實施例中，n=6。n 的計算示於下：

兩個 VR 引擎	2
第一 VR 引擎的兩個候選單字	+2
第二 VR 引擎的兩個候選單字	+2
	----------
	n=6

單字辨識及單字接受的決策規則如下：

1. 最大化 $M_i(D,H)$ 的單字被選擇爲要被辨識的單字；及

2. 該辨識在當 $M_i(D,H)>0$ 時被接受，在當 $M_i(D,H)\leq 0$ 時

五、發明說明（ 14 ）

被拒絕。

該映射函數可視需要被建構及被訓練來最小化錯誤接受/拒絕錯誤。在一具體實施例中，該常數 c_i (i=0,1,...,n)係由訓練獲得。在該訓練過程中，每個測試樣本的識別爲已知。一單字(在 W_1， W_2， W_3及 W_4之間)的測定向量被標記爲正確(+1)，而其它則標記爲不正確(-1)。訓練決定了該係數向量 $c=c_i(i=0,1,...n)$的數值，藉以最小化錯誤分類的數目。

向量 b 爲一代表每個訓練向量的正確/不正確性質的向量，而 W 爲該測定矩陣，其中每列爲一測定向量 $D_1^{w_i}$， $D_2^{w_i}$，D_g，$H_1^{w_i}$，$H_2^{w_i}$，H_g，(i=1,...4)。在一具體實施例中，該係數 c 係由計算 W 的虛擬倒數來獲得：

$c=(W^TW)^{-1}W^Tb$

此程序最小化該均分根錯誤(MSE)。在另一具體實施例中，先進的錯誤最小化程序，例如最小化該整體錯誤計數，也可用來求解係數向量 c。對於本技藝的專業人士也可瞭解到，在本技藝中已知的其它錯誤最小化程序可用來求解係數向量 c。

該映射函數方法可同樣應用到多個(>2)引擎及多個(>2)單字候選者。當有 L 個 VR 引擎及每個產生 N 個單字候選者時，該通用的映射函數具有以下形式：

$$M_i(c,V)=C_0+\sum_{l=1}^{L}\sum_{k=1}^{N}c_k^l V(l)_k^{w_i}$$

C_0爲該臨界值常數。c_k^l爲 VR 引擎 l 的第 k 個映射常數。$V(l)_k^{w_i}$爲來自 VR 引擎 l 的單字候選者 W_i的第 k 個距離。

五、發明說明（ 15 ）

在一具體實施例中，該映射函數爲非線性。一或多個變數/參數係用於取代係數的映射函數中。在一具體實施例中，用於該映射函數中的該一或多個變數/參數爲來自一VR引擎的語音參數。對於本技藝的專業人士亦可瞭解到一或多個變數/參數可爲取自該語音信號 s(t) 的測定或處理之語音參數。

藉此，已揭示一種結合語音辨識引擎之創新及改良之方法及裝置。那些本技藝的專業人士將可瞭解到，配合此處揭示的具體實施例所提出的不同的說明邏輯區塊，模組及映射可實施爲電子硬體，電腦軟體，或兩者的組合。該不同說明的元件，區塊，電路及步驟通常係以其功能來說明。是否該功能以硬體或軟體實施係依據施加在整體系統上的特殊應用及設計限制。專業人士可瞭解到在這些狀況之下硬體及軟體的互換性，及如何地最佳來實施每個特殊應用的所述之功能。如同範例，配合此處所揭示具體實施例所述的不同說明邏輯區塊，模組及映射可用執行一組韌體指令之處理器，一特定應用積體電路(ASIC)，一現場可程式閘極陣列(FPGA)或其它可程式化邏輯裝置，分離式閘極或電晶體邏輯，分離式硬體元件，例如像是暫存器，任何習用可程式軟體模組及一處理器，或其任何的組合設計來執行此處所述之功能者，來實施或執行。該 A/D 102，該VR引擎，及該映射模組 110 可較佳地執行在一微處理器中，但另外，該 A/D 102，該 VR 引擎，及該映射模組 110 可用任何習用的處理器，控制器，微控制器或狀態機器來執

A5
B5

四、中文發明摘要（發明之名稱：利用映射以用於自動語音辨識之系統及方法）

一種結合語音辨識引擎104，106，108，112，114，並使用一映射函數以解決個別語音辨識引擎104，106，108，112，114的結果之間差異之方法與系統。說話者無關語音辨識引擎104及說話者相關的語音辨識引擎106被結合。隱藏Markov模型(HMM)引擎108，114及動態時間扭曲(DTW)引擎104，106，112亦被結合。

英文發明摘要（發明之名稱："SYSTEM AND METHOD FOR AUTOMATIC VOICE RECOGNITION USING MAPPING"）

A method and system that combines voice recognition engines 104, 106, 108, 112, 114 and resolves differences between the results of individual voice recognition engines 104, 106, 108, 112, 114 using a mapping function. Speaker independent voice recognition engines 104 and speaker-dependent voice recognition engines 106 are combined. Hidden Markov Model (HMM) engines 108, 114 and Dynamic Time Warping (DTW) engines 104, 106, 112 are combined.

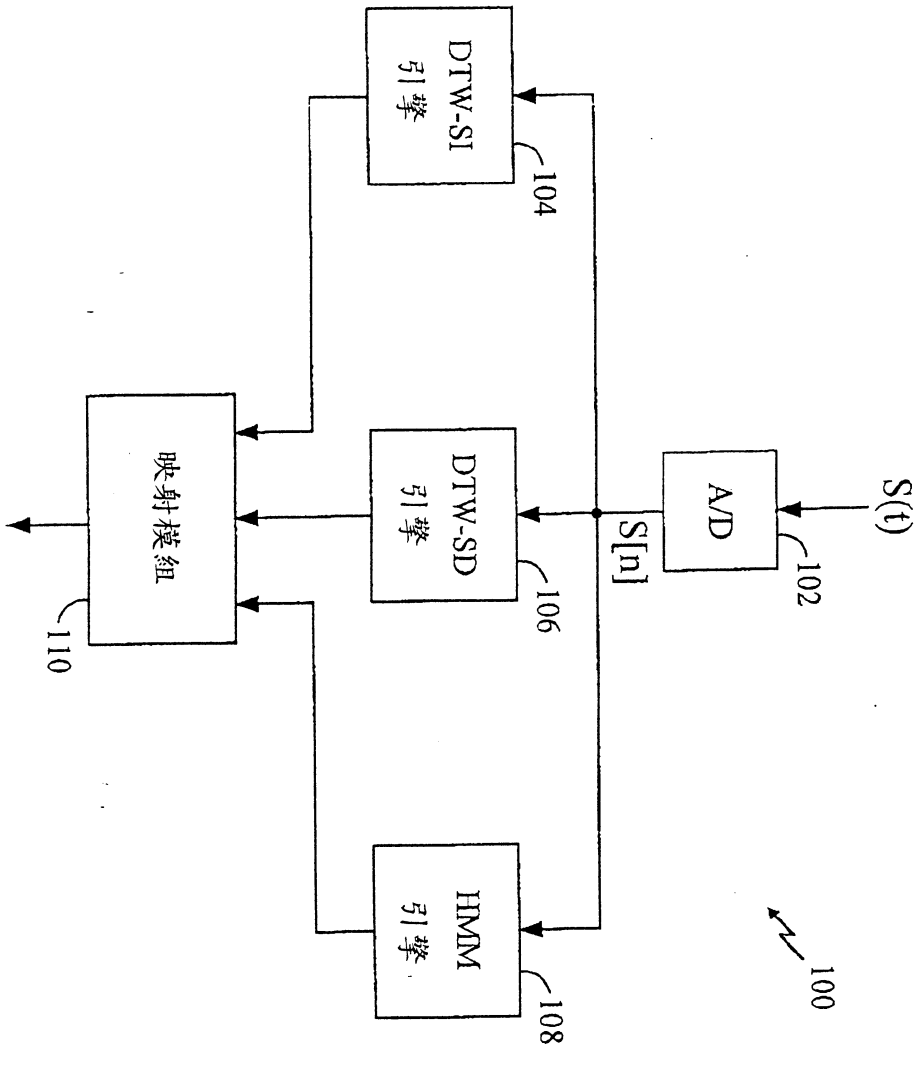

圖 1

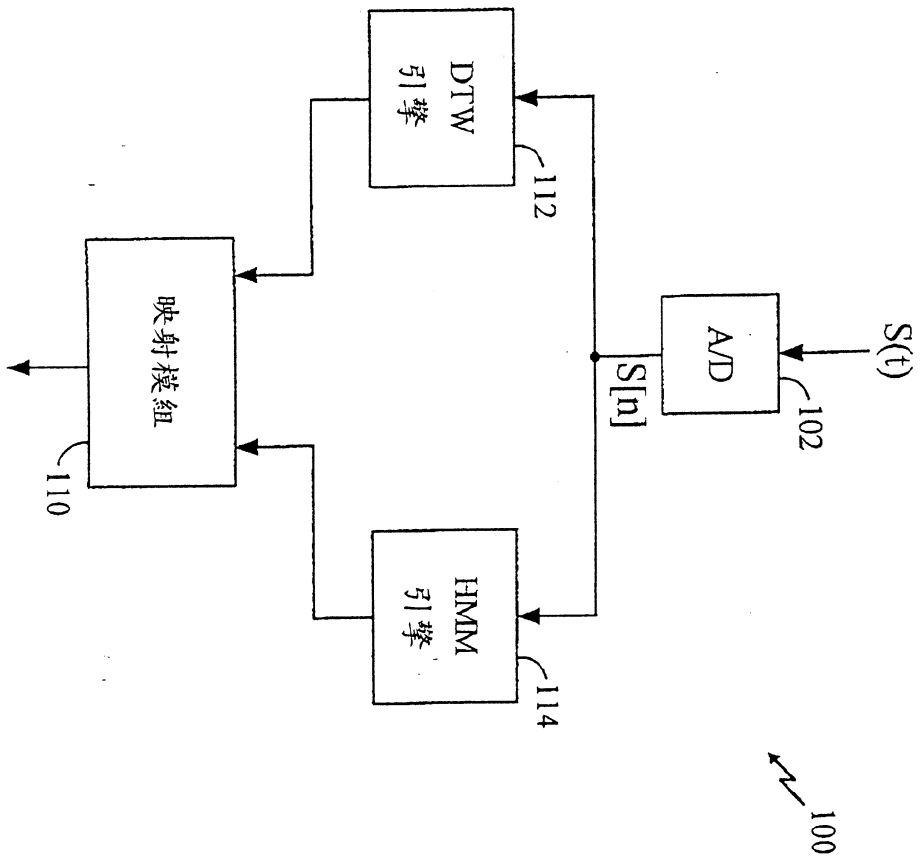

圖 2

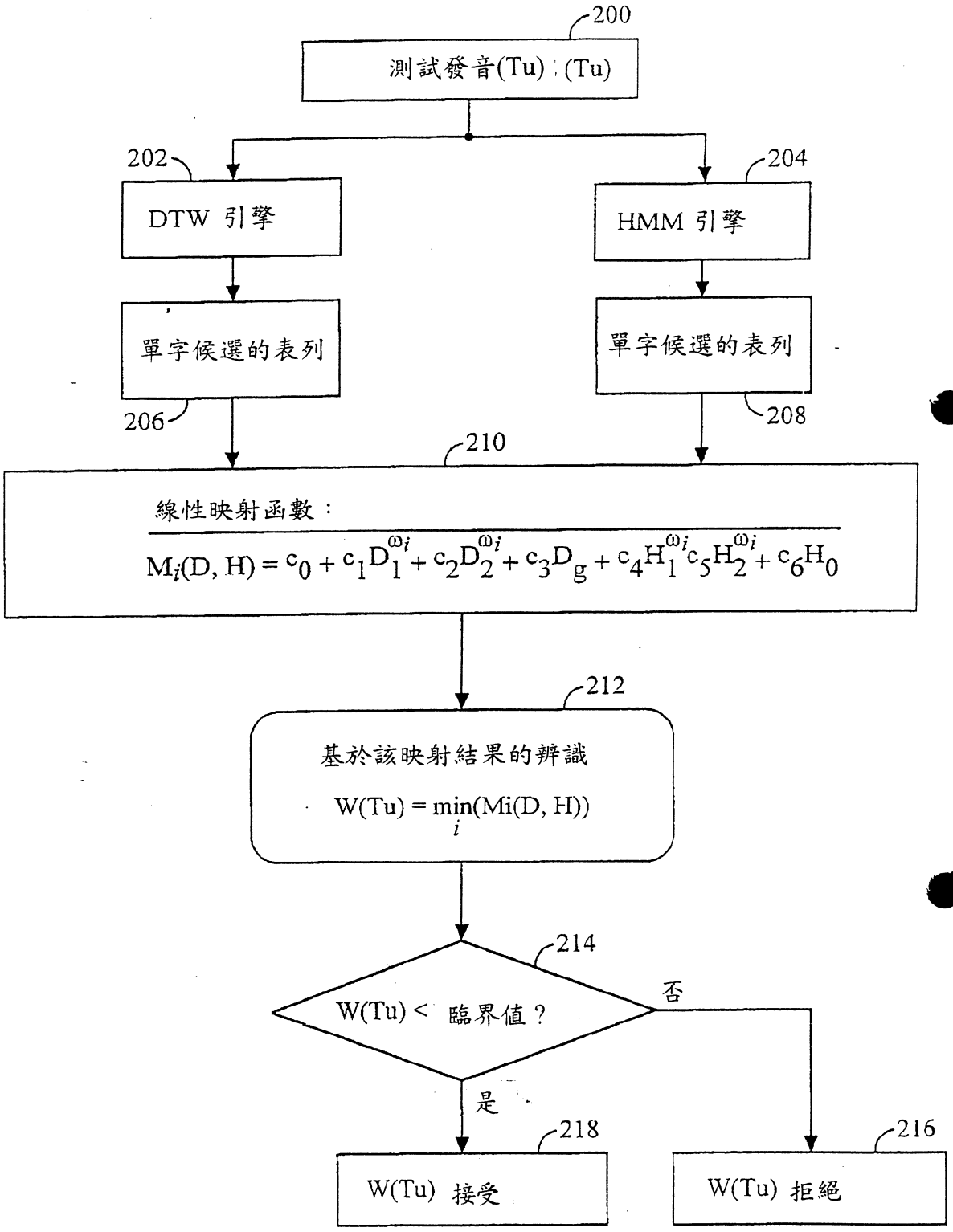
200
測試發音(Tu)：(Tu)
202
DTW 引擎
204
HMM 引擎
單字候選的表列
單字候選的表列
206
208
210
線性映射函數：
$M_i(D, H) = c_0 + c_1 D_1^{\omega_i} + c_2 D_2^{\omega_i} + c_3 D_g + c_4 H_1^{\omega_i} c_5 H_2^{\omega_i} + c_6 H_0$
212
基於該映射結果的辨識
$W(Tu) = \min_i(Mi(D, H))$
214
W(Tu) < 臨界值？
否
是
218
216
W(Tu) 接受
W(Tu) 拒絕

圖 3

五、發明說明（ 11 ）

的"垃圾"樣板之間的距離。該垃圾樣板用來代表所有不在該詞彙中的單字。

在一具體實施例中，由 VR 引擎產生的候選者中選擇一候選者的決定係基於來自該測定空間（$d_1,d_2,...d_g$ 及 $h_1,h_2,...h_g$）映射到該決策空間（接受/拒絕該測試發音做為在該表列中單字之一）。在一具體實施例中，該映射為一線性映射。在另一具體實施例中，該映射為一非線性映射。

根據一具體實施例，由具有一 DTW 為主的 VR 引擎及一 HMM 為主的 VR 引擎之裝置執行的方法步驟之流程圖，其示於圖 3。在步驟 200 中，可得到一測試發音 T_u。一旦得到該測試發音 T_u，一 DTW 語音辨識分析係在步驟 202 中執行於該測試發音 T_u 上，而一 HMM 語音辨識分析係在步驟 204 中執行於該測試發音 T_u 上。在步驟 206 中，可得到一組 DTW 候選單字 D_i。在步驟 208 中，得到一組 HMM 候選單字 H_i。在步驟 210 中，一線性映射函數應用到每個 DTW 候選單字 D_i 及每個 HMM 候選單字 H_i。在步驟 212 中，一候選字元的辨識係基於該線性映射結果。在步驟 212 中，具有最小映射函數值的候選單字係選擇為一辨識單字 $W(T_u)$。在步驟 214 中，該辨識的字元 $W(T_u)$ 的映射函數值係與一臨界值比較。如果該辨識字元 $W(T_u)$ 的映射函數值小於該臨界值，該辨識的單字即在步驟 216 中被拒絕。如果該辨識字元 $W(T_u)$ 的映射函數值大於該臨界值，該辨識的單字即在步驟 218 中被接受。

$D_i^{w_j}$ 為一測試發音 T_u 200 及詞彙中單字 W_j, j=1,2,...N 之間

五、發明說明（ 16 ）

行。該樣板可存在於 RAM 記憶體，快閃記憶體，ROM 記憶體，EPROM 記憶體，EEPROM 記憶體，暫存器，硬碟，一可移除碟片，CD-ROM，或任何其它在本技藝中已知的儲存媒體。該記憶體(未示出)可整合到任何前述的處理器(未示出)。一處理器(未示出)及記憶體(未示出)可存在於一 ASIC（未示出)。該 ASIC 可存在於一電話中。

先前對本發明具體實施例之說明被提供來使得本技藝之專業人士可以製作或使用本發明。對這些具體實施例的不同修正對本技藝之專業人士將可立即瞭解，而此處所定義的基本原理可應用到其它具體實施例中，而不使用本發明的設施。因此，本發明並不是受限於此處所示的具體實施例，而是根據此處所揭示符合於該原理及創新特徵之最廣範圍。

元件符號說明

100　語音辨識系統

102　類比到數位(A/D)轉換器

104　動態時間扭曲說話者無關(DTW-SI)引擎

106　動態時間扭曲說話者相關(DTW-SD)引擎

108　隱藏 Markov 模型(HMM)引擎

110　映射模組

112　DTW 引擎

114　HMM 引擎

六、申請專利範圍

1. 一種語音辨識系統，其包含：

   複數個語音辨識(VR)引擎，每個語音辨識引擎用來產生一單字候選者；及

   一映射模組，其用來採用來自該複數個 VR引擎的單字候選者做為輸入，並基於一映射函數來選擇一單字候選者。

2. 如申請專利範圍第 1 項之語音辨識系統，其中該複數個語音辨識引擎包含一說話者無關語音辨識引擎。

3. 如申請專利範圍第 1 項之語音辨識系統，其中該複數個語音辨識引擎包含一說話者相關語音辨識引擎。

4. 如申請專利範圍第 2 項之語音辨識系統，其中該複數個語音辨識引擎包含一說話者相關語音辨識引擎。

5. 如申請專利範圍第 4 項之語音辨識系統，其中至少一說話者無關語音辨識引擎為一動態時間扭曲語音辨識引擎。

6. 如申請專利範圍第 4 項之語音辨識系統，其中至少一說話者無關語音辨識引擎為一隱藏 Markov 模型語音辨識引擎。

7. 如申請專利範圍第 4 項之語音辨識系統，其中至少一說話者相關語音辨識引擎為一動態時間扭曲語音辨識引擎。

8. 如申請專利範圍第 4 項之語音辨識系統，其中至少一說話者相關語音辨識引擎為一隱藏 Markov 模型辨識引擎。

六、申請專利範圍

9. 如申請專利範圍第 1 項之語音辨識系統，其中該映射函數線性地映射來自一測定空間的該單字候選者到一決策空間。

10. 如申請專利範圍第 1 項之語音辨識系統，其中該映射函數非線性地映射來自一測定空間的該單字候選者到一決策空間。

11. 如申請專利範圍第 1 項之語音辨識系統，其中該單字候選者係由一單字候選者樣板及該發音之間的距離來代表。

12. 如申請專利範圍第 11 項之語音辨識系統，其中該映射模組將來自每個 VR 引擎的每個距離乘以一係數，並將該乘積加上另一個係數 C_0，藉此產生一總和。

13. 如申請專利範圍第 12 項之語音辨識系統，其中一單字候選者係基於該總和來選擇。

14. 如申請專利範圍第 1 項之語音辨識系統，其中該映射函數為：

$$M_i(F,S) = C_0 + c_1 F_1^{w_i} + c_2 F_2^{w_i} + c_3 F_g + c_4 S_1^{w_i} + c_5 S_2^{w_i} + c_n S_g$$

其中 F 為第一語音辨識引擎，S 為第二語音辨識引擎，$F_1^{w_i}$ 為發音 T_u 與候選單字 W_i 之間的距離，$F_2^{w_i}$ 為排除 W_i 的第二最佳候選者的距離，D_g 代表 T_u 及一垃圾樣板之間的距離，$S_1^{w_i}$ 為發音 T_u 與 W_i 之間的距離，$S_2^{w_i}$ 為排除 W_i 的第二最佳候選者的距離，S_g 代表 T_u 及該垃圾樣板之間的距離，及 $c_i(i=0,1,...n)$ 為一係數，而上限 n 等於 VR 引擎的數目加上每個 VR 引擎的候選單字的總和之總和。

六、申請專利範圍

15. 如申請專利範圍第 14 項之語音辨識系統，其中該係數為一實數常數。

16. 如申請專利範圍第 14 項之語音辨識系統，其中該係數為一語音參數。

17. 如申請專利範圍第 1 項之語音辨識系統，其中該映射函數為：

$$M_i(c,V)=C_0+\sum_{l=1}^{L}\sum_{k=1}^{N}c_k^l V(l)_k^{w_i}$$

其中 C_0 為一臨界值常數，c_k^l 為 VR 引擎 l 的第 k 個映射常數，而 $V(l)_k^{w_i}$ 為來自 VR 引擎 l 的單字候選者 W_i 的第 k 個距離。

18. 一種語音辨識之方法，其包含：

獲得一測試發音的至少一候選單字；及

基於一映射函數，由該至少一候選單字選擇一辨識的單字。

19. 如申請專利範圍第 18 項之方法，其中該映射函數線性地映射來自一測定空間的至少一候選單字到一決策空間。

20. 如申請專利範圍第 18 項之方法，其中該映射函數非線性地映射來自一測定空間的至少一候選單字到一決策空間。

21. 如申請專利範圍第 18 項之方法，其中該單字候選者係由一單字候選者樣板與該測試發音之間的距離來代表。

22. 如申請專利範圍第 21 項之方法，其中該映射函數將每

六、申請專利範圍

個距離乘以一係數，並將該乘積與另一個係數 C_0 相加，藉此產生一總和。

23. 如申請專利範圍第 22 項之方法，其中一辨識的單字係基於該總和來選擇。

24. 如申請專利範圍第 18 項之方法，其中該映射函數為：

$$M_i(F,S)=C_0+c_1F_1^{w_i}+c_2F_2^{w_i}+c_3F_g+c_4S_1^{w_i}+c_5S_2^{w_i}+c_nS_g$$

其中 F 為第一語音辨識引擎，S 為第二語音辨識引擎，$F_1^{w_i}$ 為發音 T_u 與候選單字 W_i 之間的距離，$F_2^{w_i}$ 為排除 W_i 的第二最佳候選者的距離，D_g 代表 T_u 及一垃圾樣板之間的距離，$S_1^{w_i}$ 為發音 T_u 與 W_i 之間的距離，$S_2^{w_i}$ 為排除 W_i 的第二最佳候選者的距離，S_g 代表 T_u 及該垃圾樣板之間的距離，及 $c_i(i=0,1,...n)$ 為一係數，而上限 n 等於 VR 引擎的數目加上每個 VR 引擎的候選單字的總和之總和。

25. 如申請專利範圍第 24 項之方法，其中該係數為一實數常數。

26. 如申請專利範圍第 24 項之方法，其中該係數為一語音參數。

27. 如申請專利範圍第 18 項之方法，其中該映射函數為：

$$M_i(c,V)=C_0+\sum_{l=1}^{L}\sum_{k=1}^{N}c_k^lV(l)_k^{w_i}$$

其中 C_0 為一臨界值常數，c_k^l 為 VR 引擎 l 的第 k 個映射常數，而 $V(l)_k^{w_i}$ 為來自 VR 引擎 l 的單字候選者 W_i 的第 k 個距離。

28. 一種語音辨識之方法，其包含：

六、申請專利範圍

獲得一測試發音；

分析該測試發音；

基於該測試發音的分析來提供該測試發音的至少一候選單字；

應用一映射函數到該至少一候選單字；

基於該至少一候選單字的該映射函數值，而由該至少一候選單字選擇一候選單字；及

比較該選擇的候選單字之映射函數值與一臨界值。

29. 如申請專利範圍第 28 項之方法，進一步包含基於該比較來接受該選擇的候選單字。

30. 如申請專利範圍第 28 項之方法，進一步包含基於該比較來拒絕該選擇的候選單字。

31. 一種語音辨識之方法，其包含：

基於一數位化發音來產生複數個單字候選者，該產生係使用複數個不同的語音辨識技術；及

應用一映射函數到該複數個單字候選者來選擇一辨識的單字。

32. 如申請專利範圍第 31 項之方法，進一步包含基於比較一臨界值與該辨識單字的該映射函數值的結果而拒絕該辨識的單字。

33. 如申請專利範圍第 32 項之方法，進一步包含基於比較一臨界值與該辨識單字的該映射函數值的結果而接受該辨識的單字。