公告本

申請日期	91 3 26.
案號	91105907
類別	G10L 15/00

A4
C4

（以上各欄由本局填註）

577043

發明 ~~新型~~ 專利說明書

一、發明 ~~新型~~ 名稱	中文	使用內隱說話者匹配的聲音辨識系統
	英文	"VOICE RECOGNITION SYSTEM USING IMPLICIT SPEAKER ADAPTATION"
二、發明 ~~創作~~ 人	姓名	1.那倫拉那斯 馬拉亞斯 NARENDRANATH MALAYATH 2.千瓊 張 CHIENCHUNG CHANG 3.畢寧 NING BI
	國籍	1.印度 INDIA 2.美國 U.S.A. 3.中國 PEOPLE'S REPUBLIC OF CHINA
	住、居所	1.美國加州聖地牙哥市薩布瑞丘道10710號 10710 SABRE HILL DRIVE, #229 SAN DIEGO, CALIFORNIA 92128 UNITED STATES OF AMERICA 2.美國加州蘭寇聖塔菲市波撒達戴諾特道6076號 6076 VIA POSADA DEL NORTE RANCHO SANTA FE, CALIFORNIA 92067 UNITED STATES OF AMERICA 3.美國加州聖地牙哥市布里斯威廣場14209號 14209 BREEZEWAY PLACE SAN DIEGO, CALIFORNIA 92128 UNITED STATES OF AMERICA
三、申請人	姓名（名稱）	美商奎康公司 QUALCOMM INCORPORATED
	國籍	美國 U.S.A.
	住、居所（事務所）	美國加州聖地牙哥市摩豪斯大道5775號 5775 MOREHOUSE DRIVE SAN DIEGO, CALIFORNIA 92121-1714 UNITED STATES OF AMERICA
	代表人姓名	菲力普 R. 華德渥斯 PHILIP R. WADSWORTH

申請日期	
案　　號	
類　　別	

A4
C4

（以上各欄由本局填註）

發明 ~~新型~~ 專利說明書

一、發明 ~~新型~~ 名稱	中　文	
	英　文	
二、發明 ~~創作~~ 人	姓　名	4.安德魯 P. 狄亞寇　ANDREW P. DEJACO 5.蘇海 亞里　SUHAIL JALIL 6.哈里那斯 加路達里　HARINATH GARUDADRI
	國　籍	4.美國 U.S.A. 5.印度 INDIA 6.加拿大 CANADA
	住、居所	4.美國加州聖地牙哥市卡米尼多路9705號 9705 CAMINITO MOJADO SAN DIEGO, CALIFORNIA 92131 UNITED STATES OF AMERICA 5.美國加州聖地牙哥市馬亞琳達路10380號 10380 MAYA LINDA ROAD, APT. C-110 SAN DIEGO, CALIFORNIA 92126 UNITED STATES OF AMERICA 6.美國加州聖地牙哥市奧菲度街9435號 9435 OVIEDO STREET SAN DIEGO, CALIFORNIA 92129 UNITED STATES OF AMERICA
三、申請人	姓　名 （名　稱）	
	國　籍	
	住、居所 （事務所）	
	代表人 姓　名	

（由本局填寫）

承辦人代碼：	
大　　類：	
ＩＰＣ分類：	

A6
B6

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	2001年03月28日	09/821,606	☑有 □無主張優先權

有關微生物已寄存於：　　寄存日期：　　，寄存號碼：

五、發明說明（ 1 ）

背景

發明領域

本發明是有關於語音訊號處理。更特別是，本發明是有關於經由未監督訓練(unsupervised training)而達成改善效能之新的聲音辨識方法及裝置。

發明背景

聲音辨識為賦予機器模擬智慧之最重要的技術之一，以辨識使用者的聲音命令及使人機介面更加便利。使用恢復來自聲學語音訊號之語言訊息的技術之系統稱為聲音辨識(voice recognition，簡稱VR)系統。圖1顯示的是具有預先強調(preemphasis)濾波器102、聲學特徵選取(acoustic feature extraction，簡稱AFE)單元104、以及樣本匹配引擎110之基本的VR系統。AFE單元104會將一連串的數位聲音樣本轉換成稱為聲學特徵向量之一組測量值(例如，選取的頻率元件)。樣本匹配引擎110會使一連串的聲學特徵向量與VR聲學模型112中所包含的樣板匹配。VR樣本匹配引擎通常會使用動態時間歪曲(Dynamic Time Warping，簡稱DTW)或隱藏式馬克沃夫模型(Hidden Markov Model，簡稱HMM)技術。DTW及HMM為此技術中所熟知的，並且詳細說明於1993年，由Prentice Hall所出版之Rabiner, L. R.及Juang, B. H.的「語音辨識基礎」(FUNDAMENTALS OF SPEECH RECOGNITION)。當一連串的聲學特徵匹配於聲學模型112中的樣板時，識別的樣板會用來使輸出產生所希望的格式，如相對應於輸入語音之一語音談話的識別序列。

五、發明說明（ 2 ）

如以上所提到的，聲學模型112通常為HMM模型或DTW模型。DTW聲學模型可視為相關於需要辨識的每一段談話之樣板的資料庫。一般而言，DTW樣板包含一序列的特徵向量，其已於相關談話的許多樣本上平均化。DTW樣本匹配通常需要找出一種儲存的樣板，其到表示輸入語音的輸入特徵向量序列之距離最小。基於聲學模型之用於HMM中的樣板包含相關語音語調的詳細統計說明。一般而言，HMM樣板會儲存一序列的平均向量、變異向量及一組轉變機率。這些參數係用來說明語音單元的統計結果，並且會從語音單元的許多樣本中評估出這些參數。HMM樣本匹配通常需要產生模型中之每一種樣板的機率，其基於相關於輸入語音之一系列的輸入特徵向量。具有最高機率的樣板會選擇用來當作最適當的輸入語調。

「訓練」係有關於採集一個或更多說話者之特定語音片段或音節的語音樣本之過程，以產生聲學模型112中的樣板。聲學模型中的每一種樣板係相關於特定的談話或稱為聲階(utterance class)的語音片段。在聲學模型中，可能有多種樣板相關於相同的聲階。「測試」係有關於使聲學模型中的樣板與自輸入語音選取的一序列特徵向量匹配之程序。已知系統的效能係大大地取決於末端使用者的輸入語音與資料庫的內容之間的匹配程度，因此會使經由訓練所產生的參考樣板與用於VR測試的語音樣本匹配。

訓練之兩種一般的形式為監督訓練(supervised training)及未監督訓練。在監督訓練中，相關於每組訓練特徵向量的

五、發明說明（ 3 ）

聲階已知為先驗，提供輸入語音的說話者時常會提供對應於預定聲階之談話或語音片段的腳本。然後，由讀取腳本而產生的特徵向量會併入相關於正確聲階的聲學模型樣板之中。

在未監督訓練中，相關於一組訓練特徵向量的聲階不是已知為先驗。在一組訓練特徵向量可併入正確的聲學模型樣板之前，必須正確地識別出聲階。在未監督訓練中，一組訓練特徵向量之聲階的識別錯誤，會導致在錯誤的聲學模型樣板中做修飾。這樣的錯誤通常會降低，而不是提昇聲音辨識效能。為了避免這樣的錯誤，基於未監督訓練之聲學模型的任何修飾通常必須非常謹慎地施行。一旦相當確信已正確識別聲階，一組訓練特徵向量會併入聲學模型。這種必要的謹慎會使得經由未監督訓練而建造SD聲學模型的過程非常緩慢。直到以此方式建造出SD聲學模型，大部分的使用者可能不會接受這樣的VR效能。

最佳而言，在訓練及測試期間，末端使用者會提供語音聲學特徵向量，以致於聲學模型112會與末端使用者的語音非常匹配。適合單一說話者之個別的聲學模型也稱為說話者相依(speaker dependent，簡稱SD)聲學模型。產生SD聲學模型通常會要求末端使用者提供大量的監督訓練樣本。首先，使用者必須提供很多不同聲階的訓練樣本。再者，為了達成最佳的效能，末端使用者必須提供用於每種聲階之代表不同可能的聲音環境之多種樣板。因為大部分的使用者不能或不願意提供必要的輸入語音來產生SD聲學模型，

五、發明說明（ 4 ）

所以會訓練取代一般使用的聲學模型之許多存在的VR系統，使用許多「典型」說話者的語音。這樣的聲學模型稱為說話者獨立(speaker independent，簡稱SI)聲學模型，並且設計成對於廣泛範圍的使用者都具有最佳的效能。然而，SI聲學模型對於任意單一的使用者不會最佳化。使用SI聲學模型的VR系統與使用適合於使用者之SD聲學模型的VR系統一樣，對於特定使用者的執行都不佳。對於一些使用者而言，如具有濃厚外國口音的使用者，使用SI聲學模型的VR系統之效能很差，以致於其一點也不能有效地使用VR的服務。

最佳而言，對於每一個個別的使用者會產生一種SD聲學模型。如上所討論，使用監督訓練所建造的SD聲學模型是不切實際的。但是使用未監督訓練來產生SD聲學模型會花長的時間，在此期間，基於部分SD聲學模型的VR效能會非常差。在使用未監督訓練來產生SD聲學模型之前及期間，使VR系統的技術執行的相當好是需要的。

發明概要

在此所揭露的方法及裝置係針對使用說話者獨立(SI)及說話者相依(SD)聲學模型的結合之創新的及改善的聲音辨識(VR)系統。使用至少一種SI聲學模型與至少一種SD聲學模型結合，以提供一種層級的語音辨識效能，其至少與純粹的SI聲學模型之語音辨識效能相同。所揭露的混合SI/SD VR系統係連續使用未監督訓練，以更新一種或更多種SD聲學模型中的聲學樣板。然後，混合VR系統會使用更新過的

五、發明說明（ 5 ）

SD聲學模型，單獨或結合至少一種SI聲學模型，用以在VR測試期間，提供改善的VR效能。

在此使用字「示範性(exemplary)」，以表示「用來當作一種範例(example)、例子(instance)或實例(illustration)」。敘述為一「示範性具體實施例」之任何的具體實施例對於另一具體實施例，不需要視為較佳或有助益的。

圖式簡單說明

從下文中參考附圖解說的詳細說明，將可更明白本發明的特徵、目的及優點，整份圖式中相同的參考文字視為對應的相同事物，其中：

圖1顯示的是基本的聲音辨識系統；

圖2顯示的是根據一示範性具體實施例之聲音辨識系統；

圖3顯示的是執行未監督訓練的方法。

圖4顯示的是產生用於未監督訓練的結合匹配分數之一示範性方法。

圖5顯示的是執行使用說話者獨立(SI)及說話者相依(SD)匹配分數的聲音辨識(測試)之流程圖；

圖6顯示的是從說話者獨立(SI)及說話者相依(SD)匹配分數中產生結合匹配分數的一種方法；以及

發明詳細說明

圖2顯示的是如可於無線遠端台202中實施的混合聲音辨識(VR)系統之一示範性具體實施例。在一示範性具體實施例中，無線遠端台202係經由無線頻道(圖中未顯示)而與無線通訊網路(圖中未顯示)做通訊。例如，遠端台202可以是

五、發明說明（ 6 ）

與無線電話系統通訊的無線電話。熟習此項技術者將會了解到，在此所敘述的技術可同樣應用於固定(不可攜帶)或不需要無線頻道的VR系統。

在所顯示的具體實施例中，來自使用者的聲音訊號會轉換成麥克風(microphone，簡稱MIC)210中的電子訊號，並且會轉換成類比至數位轉換器(analog-to-digital converter，簡稱ADC)212中的數位語音樣本。然後，數位樣本流會使用預先強調(preemphasis，簡稱PE)濾波器214濾波，例如有限脈衝響應(finite impulse response，簡稱FIR)濾波器會使低頻訊號成份衰減。

然後，濾波過的樣本會於聲學特徵選取(acoustic feature extraction，簡稱AFE)單元216中做分析。AFE單元216會將數位聲音樣本轉換成聲學特徵向量。在一示範性具體實施例中，AFE單元216會於連續數位樣本的片段上，執行傅立葉轉換，以產生對應於不同頻率儲存格(bin)之訊號強度的向量。在一示範性具體實施例中，頻率儲存格具有根據聲響比例(bark scale)之變化的頻寬。在聲響比例中，每個頻率儲存格的頻寬會與此儲存格的中心頻率產生關聯，致使較高頻率儲存格具有比較低頻率儲存格更寬的頻帶。聲響比例係敘述於1993年，由Prentice Hall所出版之Rabiner,L.R.及Juang,B.H.的「語音辨識基礎」之中，並且為此技術中所熟知。

在一示範性具體實施例中，每個聲學特徵向量係選自於固定時間區間所採集之一連串的語音樣本。在一示範性具

五、發明說明（ 7 ）

體實施例中，這些時間區間會重疊。例如，聲學特徵可以從每10毫秒開始之語音資料的20毫秒區間中得到，以致於每二個連續區間會共有一個10毫秒片段。熟習此項技術者將會了解到的是，時間區間可取代為非重疊或具有非固定持續時間，而不違反在此所敘述之具體實施例的範圍。

藉由AFE單元216所產生的聲學特徵向量會送到VR引擎220，其會執行樣本匹配，以使基於一種或更多種聲學模型230、232、以及234的內容之聲學特徵向量特徵化。

在圖2之示範性具體實施例中，三種聲學模型係顯示為：說話者相依(SI)隱藏式馬克沃夫模型(HMM)模型230、說話者獨立動態時間歪曲(DTW)模型232、以及說話者相依(SD)聲學模型234。熟習此項技術者將會了解到的是，SI聲學模型的不同結合可用於另一種具體實施例中。例如，遠端台202可能只包括SIHMM聲學模型230及SD聲學模型234，而忽略SIDTW聲學模型232。另一種為，遠端台202可能包括單一SIHMM聲學模型230、SD聲學模型234及兩種不同的SIDTW聲學模型232。除此之外，熟習此項技術者將會了解到的是，SD聲學模型234可能為HMM型式或DTW型式或這兩種的結合。在一示範性具體實施例中，SD聲學模型234為DTW聲學模型。

如上所述，VR引擎220係執行樣本匹配，以決定聲學特徵向量與一種或更多種聲學模型230、232、以及234的內容之間的匹配程度。在一示範性具體實施例中，VR引擎220會產生基於聲學特徵向量與在聲學模型230、232、以及234

五、發明說明（ 8 ）

之每一種中的不同聲學樣板匹配的匹配分數。例如，VR引擎220會產生基於一組聲學特徵向量與在SIHMM聲學模型230中的多種HMM樣板匹配的HMM匹配分數。同樣地，VR引擎220會產生基於聲學特徵向量與在SIDTW聲學模型232中的多種DTW樣板匹配的DTW匹配分數。VR引擎220會產生基於聲學特徵向量與在SD聲學模型234中的樣板匹配的匹配分數。

如上所述，在聲學模型中的每一種樣板係相關於聲階。在一示範性具體實施例中，VR引擎220會將相關於相同聲階的樣板之分數結合，以產生結合的匹配分數，其係用於未監督訓練中。例如，VR引擎220會結合從互相關聯的一輸入組之聲學特徵向量獲得之SIHMM及SIDTW分數，以產生結合的SI分數。基於結合的匹配分數，VR引擎220會決定是否將輸入組之聲學特徵向量儲存為SD聲學模型234中的SD樣板。在一示範性具體實施例中，執行更新SD聲學模型234的未監督訓練係使用獨有的SI匹配分數。這樣可避免可能在其他方面，起因於將推斷出的SD聲學模型234來用於本身的未監督訓練所產生的附加錯誤。執行此未監督訓練的一種示範性方法詳細說明如下。

除了未監督訓練之外，在測試期間，VR引擎220會使用各種不同的聲學模型(230，232，234)。在一示範性具體實施例中，VR引擎220會從聲學模型(230，232，234)中得到匹配分數，並且會產生用於每種聲階之結合的匹配分數。結合的匹配分數會用來選擇與輸入語音最佳匹配之聲階。

五、發明說明（ 9 ）

當需要辨識整個談話或詞組時，VR引擎220會把連續的聲階聚集在一起。然後，VR引擎220會提供關於辨識過的談話或詞組之資訊到控制處理器222，其使用此資訊來決定語音資訊或命令之合適的回應。例如，回應於辨識過的談話或詞組，控制處理器222會經由顯示器或其他的使用者介面而提供回授給使用者。在另一個例子中，控制處理器222會經由無線數據機(modem)218及天線224而傳送訊息至無線網路(圖中未顯示)，開始打行動電話到相關於已說出名字且已完全辨識的個人之目的地電話號碼。

無線數據機218可以經由包括CDMA、TDMA、或FDMA之多種的無線頻道型式的任何一種來傳送訊號。除此之外，無線數據機218可以不違反已敘述的具體實施例之範圍之能於非無線頻道上傳輸的通訊介面型式取代。例如，遠端台202可經由包括地線(land-line)數據機、T1/E1、ISDN、DSL、乙太網路，或甚至印刷電路板(printed circuit board，簡稱PCB)上的走線(trace)之多種型式的通訊頻道之任何一種來傳送訊號資訊。

圖3顯示的是執行未監督訓練之一示範性方法的流程圖。在步驟302，類比語音資料會於類比至數位轉換器(ADC)中取樣(圖2中的212)。然後，數位樣本流在步驟304，會使用預先強調(PE)濾波器(圖2中的214)來濾波。在步驟306，在聲學特徵選取(AFE)單元(圖2中的216)中，會從濾波過的樣本中選取聲學特徵向量。VR引擎(圖2中的220)會接收來自AFE單元216的輸入聲學特徵向量，並且會對SI聲學模型

五、發明說明（ 10 ）

(圖2中的230及232)的內容執行輸入聲學特徵向量的樣本匹配。在步驟308，VR引擎220會從樣本匹配的結果中產生匹配分數。VR引擎220會藉由匹配具有SIHMM聲學模型230的輸入聲學特徵向量而產生SIHMM匹配分數，並且會藉由匹配具有SIDTW聲學模型232的輸入聲學特徵向量而產生SIDTW匹配分數。在SIHMM及SIDTW聲學模型(230及232)中的每種聲學樣板係與特定的聲階相關。在步驟310，會結合SIHMM及SIDTW分數，以組成結合匹配分數。

圖4顯示的是產生用於未監督訓練的結合匹配分數。在所顯示之示範性具體實施例中，用於特定聲階的說話者獨立結合匹配分數S_{COMB_SI}為根據如所顯示的EQN.1之權重和，其中：

$SIHMM_T$為用於目標聲階的SIHMM匹配分數；

$SIHMM_{NT}$為用於相關於非目標聲階(為目標聲階之外的一種聲階)之SIHMM聲學模型中的樣板之下一個最佳匹配分數；

$SIHMM_G$為用於「無用(garbage)」聲階的SIHMM匹配分數；

$SIDTW_T$為用於目標聲階的SIDTW匹配分數；

$SIDTW_{NT}$為用於相關於非目標聲階之SIDTW聲學模型中的樣板之下一個最佳匹配分數；以及

$SIDTW_G$為用於「無用」聲階的SIDTW匹配分數。

各種不同的個別匹配分數$SIHMM_n$及$SIDTW_n$可視為表示出一連串的輸入聲學特徵向量與聲學模型中的樣板之間的距離值。在輸入聲學特徵向量與樣板之間的距離愈大，匹配

五、發明說明（ 11 ）

分數愈大。在樣板與輸入聲學特徵向量之間的緊密匹配會產生非常低的匹配分數。如果將一連串的輸入聲學特徵向量與相關於不同聲階的兩種樣板比較而產生幾乎相同的兩個匹配分數，則VR系統會不能辨識哪一個為「正確」聲階。

$SIHMM_G$及$SIDTW_G$為用於「無用」聲階的SIDTW匹配分數。相關於無用聲階的樣板或多種樣板稱為無用樣板，並且沒有對應於特定的談話或詞組。由於這個原因，其與全部的輸入語音同樣沒有關聯。無用匹配分數可用來當作VR系統中的一種雜訊層測量。一般而言，在聲階可確信地辨識之前，一連串的輸入聲學特徵向量與相關於目標聲階的樣板之匹配應該比相關於無用樣板的匹配具有更佳的程度。

在VR系統可確信地辨識聲階為「正確」聲階之前，輸入聲學特徵向量與相關於此聲階的樣板之匹配，應該比相關於無用樣板或相關於其他聲階的樣板之匹配具有更高的程度。從多種聲學模型中產生的結合匹配分數比只基於一種聲學模型的匹配分數可在聲階之間，做更確信的辨識。在一示範性具體實施例中，VR系統係使用這樣的結合匹配分數，以決定是否取代具有從新組的輸入聲學特徵向量中得到的一個輸入聲學特徵向量之SD聲學模型中的樣板(圖2中的234)。

會選擇權重因子($W_1 \ldots W_6$)，以提供全部聲學環境之最佳的訓練效能。在一示範性具體實施例中，權重因子($W_1 \ldots W_6$)為用於全部聲階的常數。換句話說，用來產生用

五、發明說明（ 12 ）

於第一目標聲階的結合匹配分數之W_n與用來產生用於另一種目標聲階的結合匹配分數之W_n值是相同的。在另一種具體實施例中，權重因子會基於目標聲階而變化。圖4中所顯示之其他方式的結合對於熟習此項技術者將是顯然可知的，並且視為在此所敘述之具體實施例的範圍內。例如，也可以使用超過六個或少於六個的權重輸入。另一種顯然可知的變化將會產生基於一種型式的聲學模型之結合匹配分數。例如，結合匹配分數可基於$SIHMM_T$、$SIHMM_{NT}$、以及$SIHMM_G$而產生。或者，結合匹配分數可基於$SIDTW_T$、$SIDTW_{NT}$、以及$SIDTW_G$而產生。

在一示範性具體實施例中，W_1及W_4為負數，而S_{COMB}之較大(或較小之負的)值係表示在目標聲階與一連串的輸入聲學特徵向量之間的匹配(較小距離)具有較大的程度。熟習此項技術者將會顯然可知，權重因子的符號可輕易地重新安排，以致於對應於較小值的匹配具有較大的程度，而不會違反所揭露的具體實施例之範圍。

回到圖3，在步驟310，會產生用於相關於HMM及DTW聲學模型(230及232)中的樣板之聲階的結合匹配分數。在一示範性具體實施例中，只會產生用於相關於最佳n SIHMM匹配分數的聲階及用於相關於最佳m SIDTW匹配分數的聲階之結合匹配分數。這種限制可適合用來節省計算資源，即使當產生個別的匹配分數時，會消耗非常大量的計算電源。例如，如果n＝m＝3，會產生用於相關於頂端三個SIHMM的聲階及相關於頂端三個SIDTW匹配分數的聲階之

五、發明說明（ 13 ）

結合匹配分數。取決於相關於頂端三個SIHMM匹配分數的聲階與相關於頂端三個SIDTW匹配分數的聲階是否相同，這種方法將會產生三種到六種不同的結合匹配分數。

在步驟312，遠端台202會將結合匹配分數與以相對應的樣板(相關於相同的聲階)儲存於SD聲學模型中的結合匹配分數做比較。如果新的一連串的輸入聲學特徵向量比用於相同聲階之儲存於SD模型中之較舊樣板之一連串的輸入聲學特徵向量具有較大的匹配程度，則會從新的一連串的輸入聲學特徵向量中產生新的SD樣板。在一具體實施例中，其中SD聲學模型為DTW聲學模型，一連串的輸入聲學特徵向量本身會產生新的SD樣板。然後，較舊的樣板會以新的樣板取代，並且相關於新的樣板之結合匹配分數會儲存於SD聲學模型中，以在未來的比較中使用。

在另一種具體實施例中，未監督訓練係用來更新說話者相依隱藏式馬克沃夫模型(SDHMM)聲學模型中的一種或更多種的樣板。SDHMM聲學模型可用來取代SDDTW模型或除了SD聲學模型234內的SDDTW聲學模型。

在一示範性具體實施例中，在步驟312的比較也包括將未來新的SD樣板與常數訓練臨界值(threshold)的結合匹配分數做比較。即使還沒有任何樣板為了特定聲階儲存於SD聲學模型之中，新樣板將不會儲存於SD聲學模型之中，除非其具有的結合匹配分數比訓練臨界值更佳(表示較大的匹配程度)。

在另一種具體實施例中，在已取代SD聲學模型中的任一

五、發明說明（ 14 ）

種樣板之前，預設的SD聲學模型係具有來自SI聲學模型的樣板。這樣的初始值可提供另一種方法，以確定使用SD聲學模型的VR效能與只使用SI聲學模型的VR效能至少開始時是一樣好的。當愈來愈多之SD聲學模型中的樣板已更新，使用SD聲學模型的VR效能將會超越只使用SI聲學模型的VR效能。

在另一種具體實施例中，VR系統允許使用者執行監督訓練。使用者必須在執行這樣的監督訓練之前，將VR系統放入監督訓練模式之中。在監督訓練期間，VR系統具有一種正確聲階的先驗知識。如果輸入語音的結合匹配分數比此聲階先前所儲存之SD樣板的結合匹配分數更佳，則會使用輸入語音來組成一種取代SD樣板。在另一種具體實施例中，VR系統允許使用者在監督訓練期間，強制取代存在的SD樣板。

SD聲學模型可設計成對於單一聲階，具有多種(兩種或更多種)樣板的空間。在一示範性具體實施例中，對於每種聲階，SD聲學模型中會儲存兩種樣板。因此，在步驟312的比較，對於相同的聲階需要將以新樣板所得到的匹配分數與以SD聲學模型中的兩種樣板所得到的匹配分數做比較。如果新樣板比SD聲學模型中的任一種舊樣板具有更佳的匹配分數，則在步驟314，具有最差匹配分數的SD聲學模型樣板會以新樣板來取代。如果新樣板的匹配分數並沒有比任一種舊樣板更佳，則會跳過步驟314。此外，在步驟312，以新樣板所得到的匹配分數會與匹配分數臨界做比較。所

五、發明說明（ 15 ）

以，直到新樣板具有比儲存於SD聲學模型中的臨界更佳之匹配分數，在其用來覆寫SD聲學模型的先前內容之前，新樣板會與臨界值做比較。顯然可知的變化，如根據結合匹配分數及只比較新匹配分數與最低匹配分數之以排序順序所儲存的SD聲學模型樣板，可預期及視為在此所揭露的具體實施例之範圍內。對於每種聲階，儲存於聲學模型中之樣板的數目之顯然可知的變化也可以預期。例如，SD聲學模型對於每種聲階，可包含超過兩種樣板，或者對於不同聲階，可包含不同數目的樣板。

圖5顯示的是執行使用SI及SD聲學模型的結合之VR測試之示範性方法的流程圖。步驟302、304、306、以及308與圖3中所敘述的相同。示範性方法在步驟510係不同於圖3中所顯示的方法。在步驟510，VR引擎220會產生基於輸入聲學特徵向量與SD聲學模型中的樣板之比較的SD匹配分數。在一示範性具體實施例中，所產生的SD匹配分數只會用於相關於最佳n SIHMM匹配分數及最佳m SIDTW匹配分數之聲階。在一示範性具體實施例中，n＝m＝3。取決於兩組聲階之間的重疊程度，這將會導致產生用於三種到六種聲階的SD匹配分數。如上所討論的，對於單一聲階，SD聲學模型可包含多種樣板。在步驟512，VR引擎220會產生用於VR測試之混合結合匹配分數。在一示範性具體實施例中，這些混合結合匹配分數係基於個別的SI及個別的SD匹配分數。在步驟514，會選擇具有最佳結合匹配分數之談話或話語，並且會與測試臨界值做比較。如果一種聲階的結合匹配

五、發明說明（ 16 ）

分數超過此測試臨界，才會視為辨識到此種聲階。在一示範性具體實施例中，用來產生訓練的結合分數之權重[W_1... W_6]（如圖4所顯示）係與用來產生測試的結合分數之權重值[W_1... W_6]（如圖6所顯示）相同，但是訓練臨界值與測試臨界並不相同。

圖6顯示的是執行步驟512而產生的混合結合匹配分數。所顯示之示範性具體實施例之運作係與圖4所顯示的結合相同，除了權重因子W_4係用於DTW_T，來取代$SIDTW_T$，以及權重因子W_5係用於DTW_{NT}，來取代$SIDTW_{NT}$之外。DTW_T（用於目標聲階的動態時間歪曲匹配分數）係選自相關於目標聲階之最佳的SIDTW及SDDTW分數。同樣地，DTW_{NT}（用於非目標聲階的動態時間歪曲匹配分數）係選自相關於非目標聲階之最佳的SIDTW及SDDTW分數。

用於特定聲階的SI/SD混合分數S_{COMB_H}為根據如所顯示之EQN.2之權重和，其中$SIHMM_T$、$SIHMM_{NT}$、$SIHMM_G$、以及$SIDTW_G$與EQN.1相同。特別而言，在EQN.2中：

$SIHMM_T$為用於目標聲階的SIHMM匹配分數；

$SIHMM_{NT}$為用於相關於非目標聲階(為目標聲階之外的一種聲階)之SIHMM聲學模型中的樣板之下一個最佳匹配分數；

$SIHMM_G$為用於「無用」聲階的SIHMM匹配分數；

DTW_T為對應於目標聲階之用於SI及SD樣板之最佳DTW匹配分數；

DTW_{NT}為對應於非目標聲階之用於SI及SD樣板之最佳DTW匹配分數；以及

五、發明說明（ 17 ）

$SIDTW_G$為用於「無用」聲階的SIDTW匹配分數。

因此，SI/SD混合分數S_{COMB_H}為個別的SI及SD匹配分數之結合。所產生的結合匹配分數不完全依賴SI或SD聲學模型。如果匹配分數$SIDTW_T$比任何的$SDDTW_T$分數更佳，則會從較佳的$SIDTW_T$分數中計算SI/SD混合分數。同樣地，如果匹配分數$SDDTW_T$比任何的$SIDTW_T$更佳，則會從較佳的$SDDTW_T$分數中計算SI/SD混合分數。因此，如果SD聲學模型中的樣板產生差的匹配分數，VR系統仍然會基於SI/SD混合分數的SI部分而辨識出輸入語音。這種差的SD匹配分數可能有多種原因，包括在訓練及測試期間，聲學環境之間的差異，或者也許是用於訓練的輸入品質很差。

在另一種具體實施例中，SI分數的權重與SD分數比較起來係非常的小，或者甚至可以完全忽視。例如，DTW_T係選自相關於目標聲階之最佳的SDDTW分數，而忽視用於目標聲階的SIDTW分數。再者，DTW_{NT}可選自相關於非目標聲階之最佳的SIDTW或SDDTW分數，來取代使用兩組的分數。

雖然所敘述之示範性具體實施例係只使用說話者相依模型化的SDDTW聲學模型，但是在此所敘述的混合方法同樣可應用到使用SDHMM聲學模型的VR系統或甚至SDDTW及SDHMM聲學模型的結合。例如，藉由修改圖6所顯示的方法，權重因子W_1可用於選自最佳的$SIHMM_T$及$SDHMM_T$之匹配分數。權重因子W_2可用於選自最佳的$SIHMM_{NT}$及$SDHMM_{NT}$之匹配分數。

因此，在此所揭露的為使用SI及SD聲學模型的結合之VR

五、發明說明（ 18 ）

方法及裝置，用以改善在未監督訓練及測試期間的VR效能。熟習此項技術者將會了解到的是，資訊及訊號可使用任何變化的不同科技及技術來表示。例如，以上全部敘述會提及到的資料、指令、命令、資訊、訊號、位元、符號、以及晶片可藉由電壓、電流、電磁波、磁場或粒子、光學場或微粒、或者是其任何的結合來表示。再者，雖然具體實施例主要是就動態時間歪曲(DTW)或隱藏式馬克沃夫模型(HMM)聲學模型的方面來做說明，但是所敘述的技術可應用於如神經網路聲學模型之其他型式的聲學模型。

熟習者將會進一步了解到，配合在此揭露的具體實施例而敘述的各種顯示的邏輯方塊、模組、電路、以及演算法步驟可以電子硬體、電腦軟體、或兩者的結合來實施。為了清楚表示硬體及軟體的可交換性，就其功能而言，各種顯示的元件、方塊、模組、電路、以及步驟已敘述如上。這樣的功能是否以硬體或軟體來實施係取決於加諸於全部系統的特定應用及設計限制。熟習的技術者對於每種特定的應用，可以變化的方式來實施所敘述的功能，但是這樣的實施決定應該不能解釋為產生違反本發明的範圍。

配合在此揭露的具體實施例而敘述的各種顯示的邏輯方塊、模組、以及電路可以一般用途的處理器、數位訊號處理器(digital signal processor，簡稱DSP)、特定應用積體電路(application specific integrated circuit，簡稱ASIC)、場域可程式閘極陣列(field programmable gate array，簡稱FPGA)或其他可程式的邏輯裝置、分離閘或電晶體邏輯、分離硬體元

五、發明說明（ 19 ）

件、或其設計用來執行功能的任意結合來實施或執行。一般用途的處理器可以是微處理器，但是另一方面，處理器可以是任何傳統的處理器、控制器、微控制器、或狀態機器。處理器也可以實施為計算裝置的結合，例如，DSP及微處理器、多個微處理器、結合DSP核心之一個或更多個微處理器、或者是這樣配置之結合。

配合在此揭露的具體實施例而敘述的方法或演算法之步驟可直接以硬體、以藉由處理器來執行的軟體模組、或以兩者的結合來具體實施。軟體模組可存在於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、可抽取磁碟、CD-ROM、或在此技術中所熟知之任何其他型式的儲存媒體。一種示範性儲存媒體係耦接至處理器，以致於處理器可自儲存媒體中讀取資訊，以及寫入資訊到儲存媒體。在另一方面，可整合儲存媒體到處理器。處理器及儲存媒體可存在於ASIC中。在另一方面，處理器及儲存媒體可存在於使用者終端中，當作分離元件。

揭露的具體實施例之先前的說明係用以使熟習此項技術的任何人能實施或使用本發明。這些具體實施例的各種修改對於熟習此項技術者，將能立即顯然可知，並且在此所定義的通則可應用於其他的具體實施例，而不違反本發明的精神或範圍。因此，本發明並不是意圖限制在此所顯示的具體實施例，而是使本發明的之最廣的範圍與在此所揭露的原理及新穎性符合。

A5
B5

四、中文發明摘要（發明之名稱：　使用內隱說話者匹配的聲音辨識系統　）

本發明所揭露的是聲音辨識(VR)系統，其使用說話者獨立(SI)(230及232)及說話者相依(SD)(234)聲學模型的結合。使用至少一種SI聲學模型(230及232)與至少一種SD聲學模型(234)結合，以提供一種層級的語音辨識效能，其至少與純粹的SI聲學模型之語音辨識效能相同。所揭露的混合SI/SD VR系統係連續使用未監督訓練，以更新一種或更多種SD聲學模型(234)中的聲學樣板。然後，混合VR系統會使用更新過的SD聲學模型(234)與至少一種SI聲學模型(230及232)結合，用以在VR測試期間，提供改善的VR效能。

英文發明摘要（發明之名稱：　"VOICE RECOGNITION SYSTEM USING IMPLICIT SPEAKER ADAPTATION"　）

A voice recognition (VR) system is disclosed that utilizes a combination of speaker independent (SI) (**230** and **232**) and speaker dependent (SD) (**234**) acoustic models. At least one SI acoustic model (**230** and **232**) is used in combination with at least one SD acoustic model (**234**) to provide a level of speech recognition performance that at least equals that of a purely SI acoustic model. The disclosed hybrid SI/SD VR system continually uses unsupervised training to update the acoustic templates in the one or more SD acoustic models (**234**). The hybrid VR system then uses the updated SD acoustic models (**234**) in combination with the at least one SI acoustic model (**230** and **232**) to provide improved VR performance during VR testing

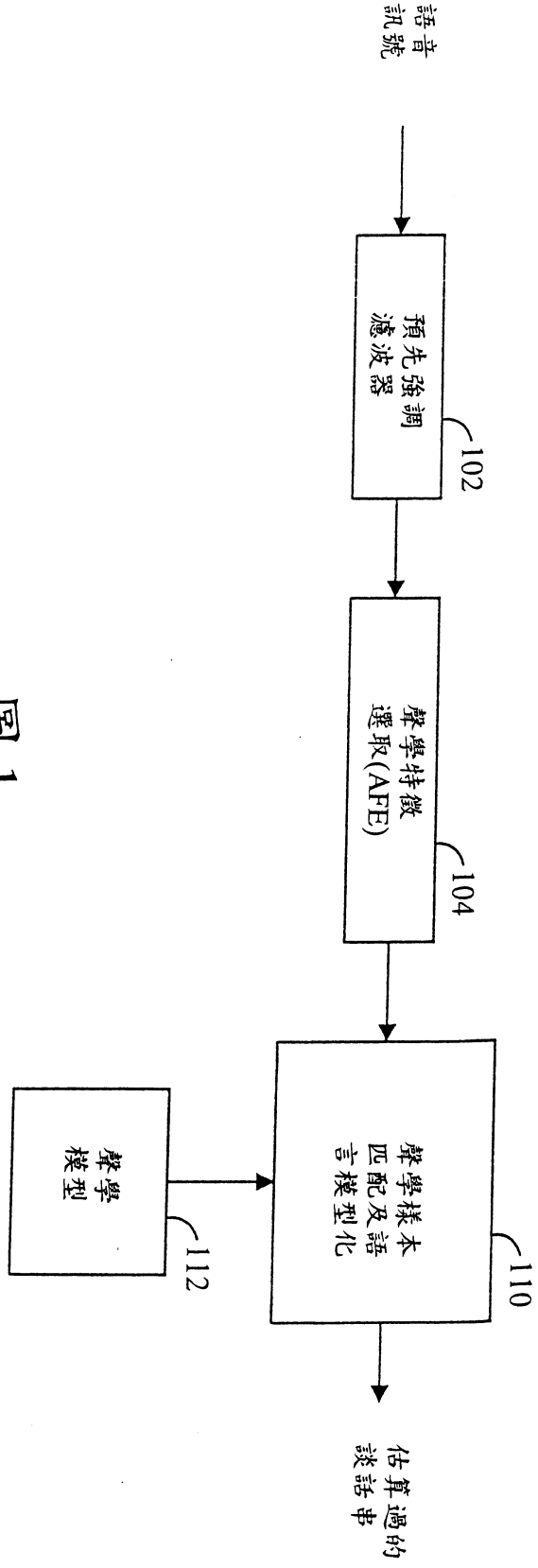

語音訊號
預先強調濾波器
102
聲學特徵選取(AFE)
104
聲學樣本匹配及語言模型化
110
聲學模型
112
估算過的談話串

圖 1

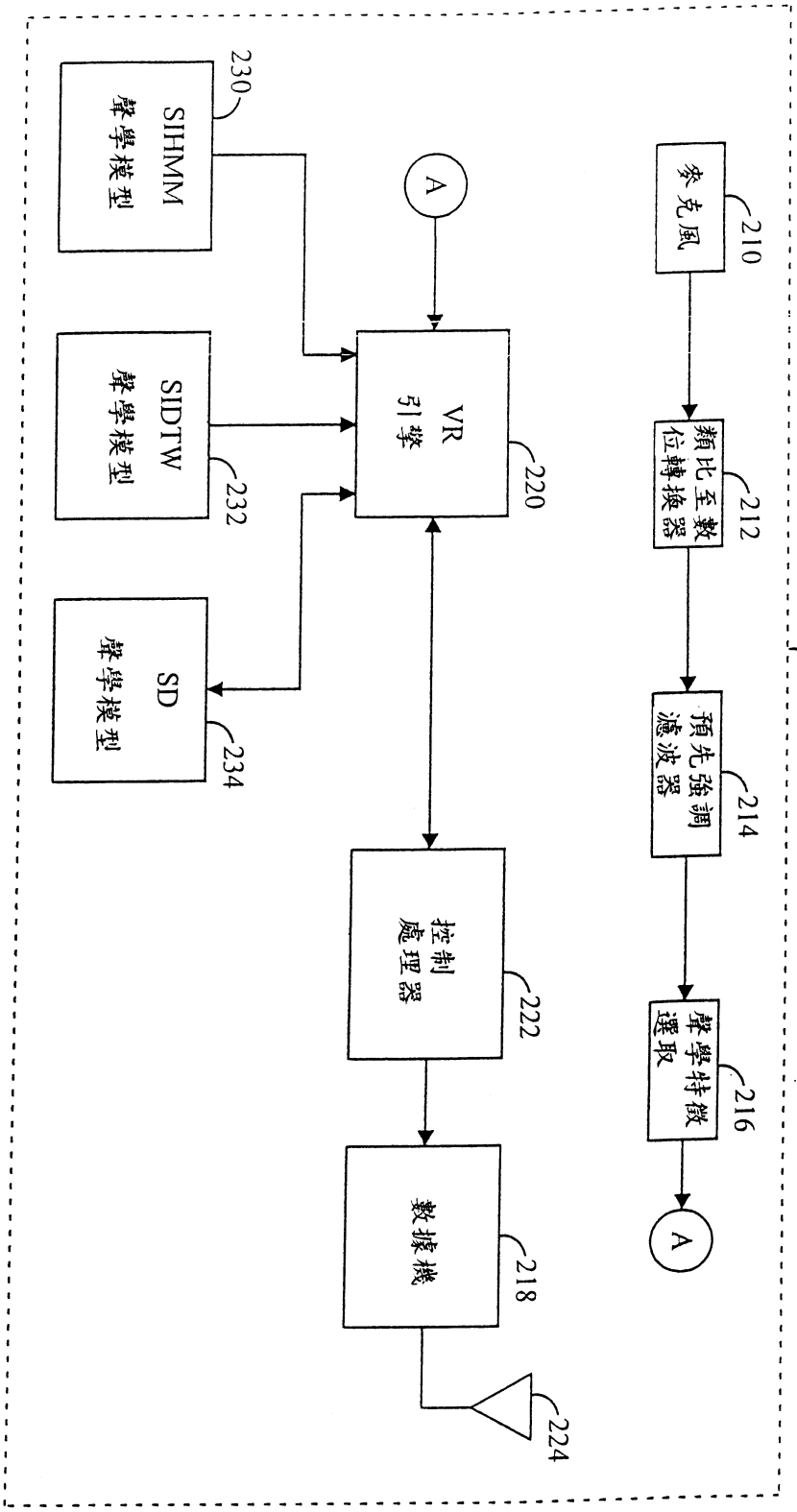

202
麥克風
210
類比至數位轉換器
212
預先強調濾波器
214
聲學特徵選取
216
A
A
VR引擎
220
控制處理器
222
數據機
218
224
SIHMM聲學模型
230
SIDTW聲學模型
232
SD聲學模型
234

圖 2

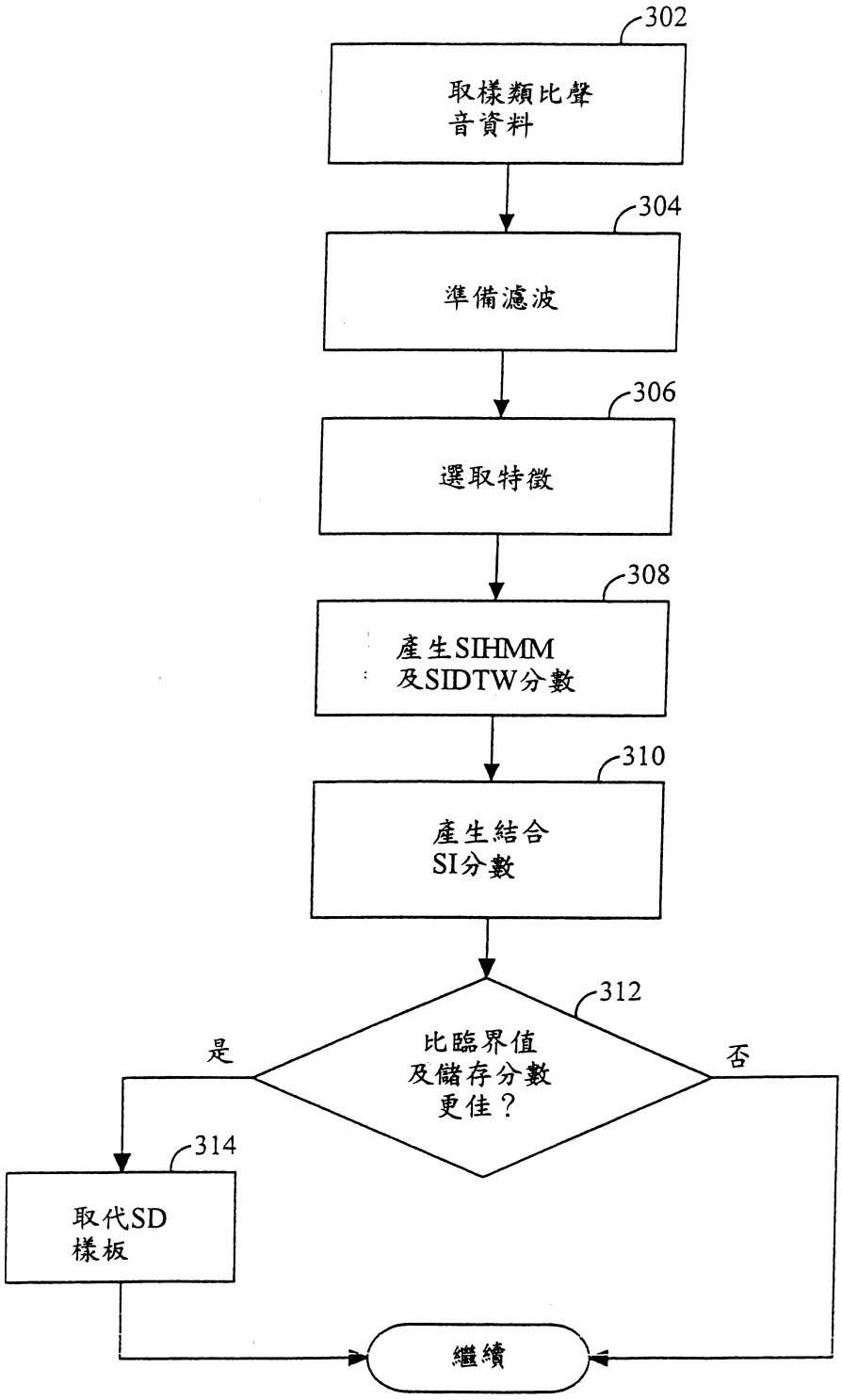
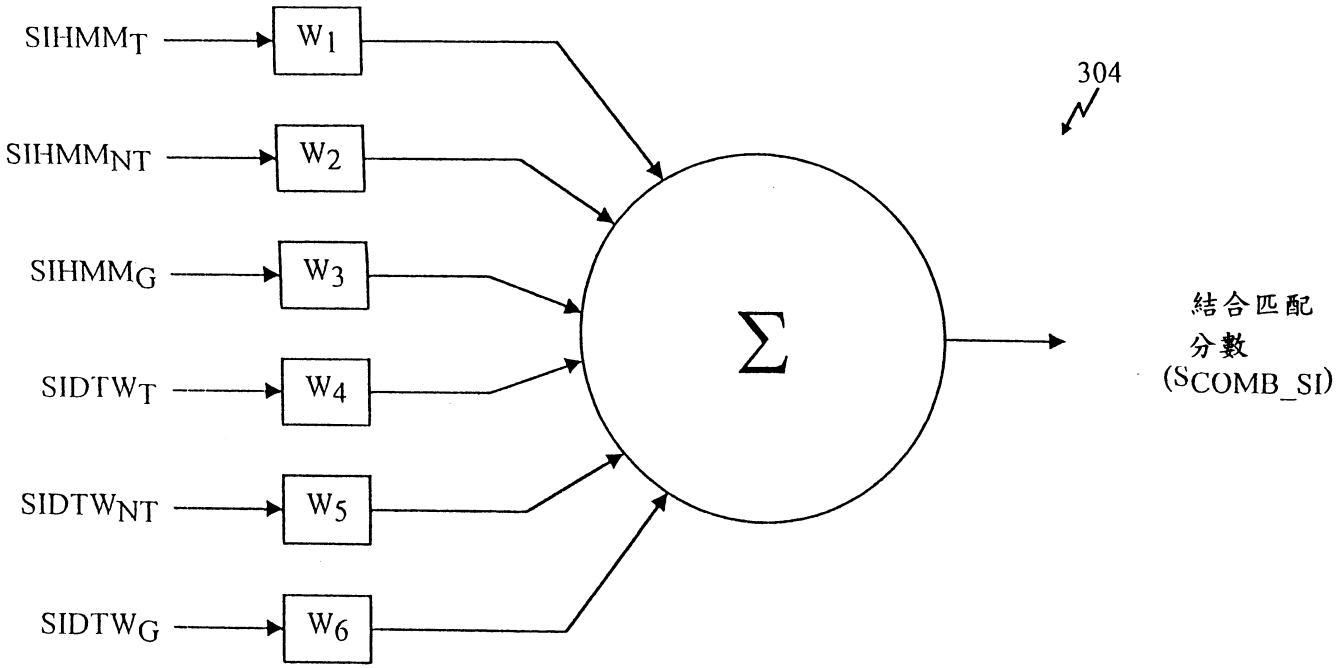

302
取樣類比聲音資料
304
準備濾波
306
選取特徵
308
產生SIHMM及SIDTW分數
310
產生結合SI分數
312
比臨界值及儲存分數更佳？
是
否
314
取代SD樣板
繼續

圖 3



EQN. 1 $S_{COMB\_SI} = (SIHMM_T * W_1)+(SIHMM_{NT} * W_2)+(SIHMM_G * W_3)+(SIDTW_T * W_4)+(SIDTW_{NT} * W_5)+(SIDTW_G * W_6)$

圖 4

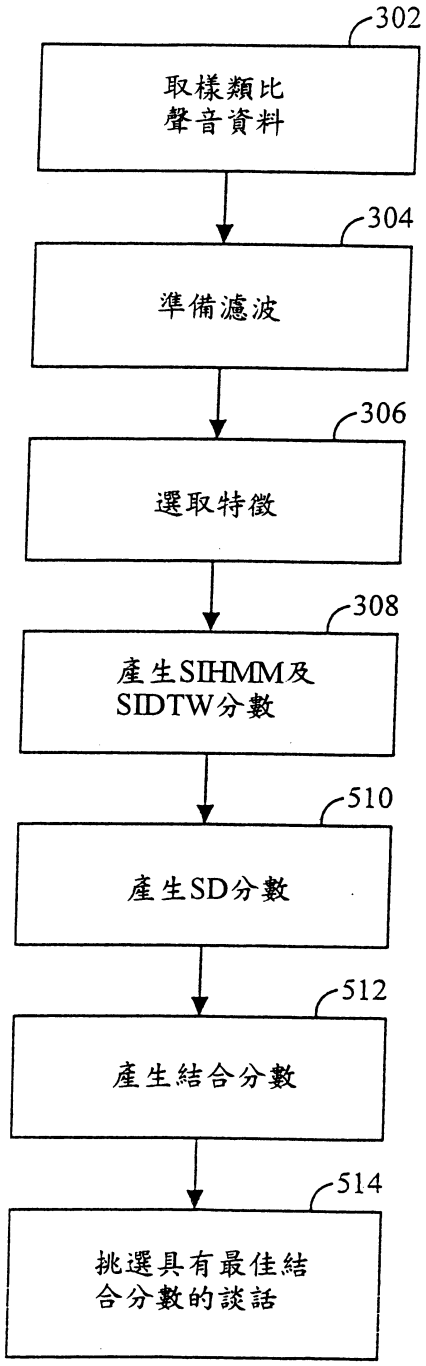
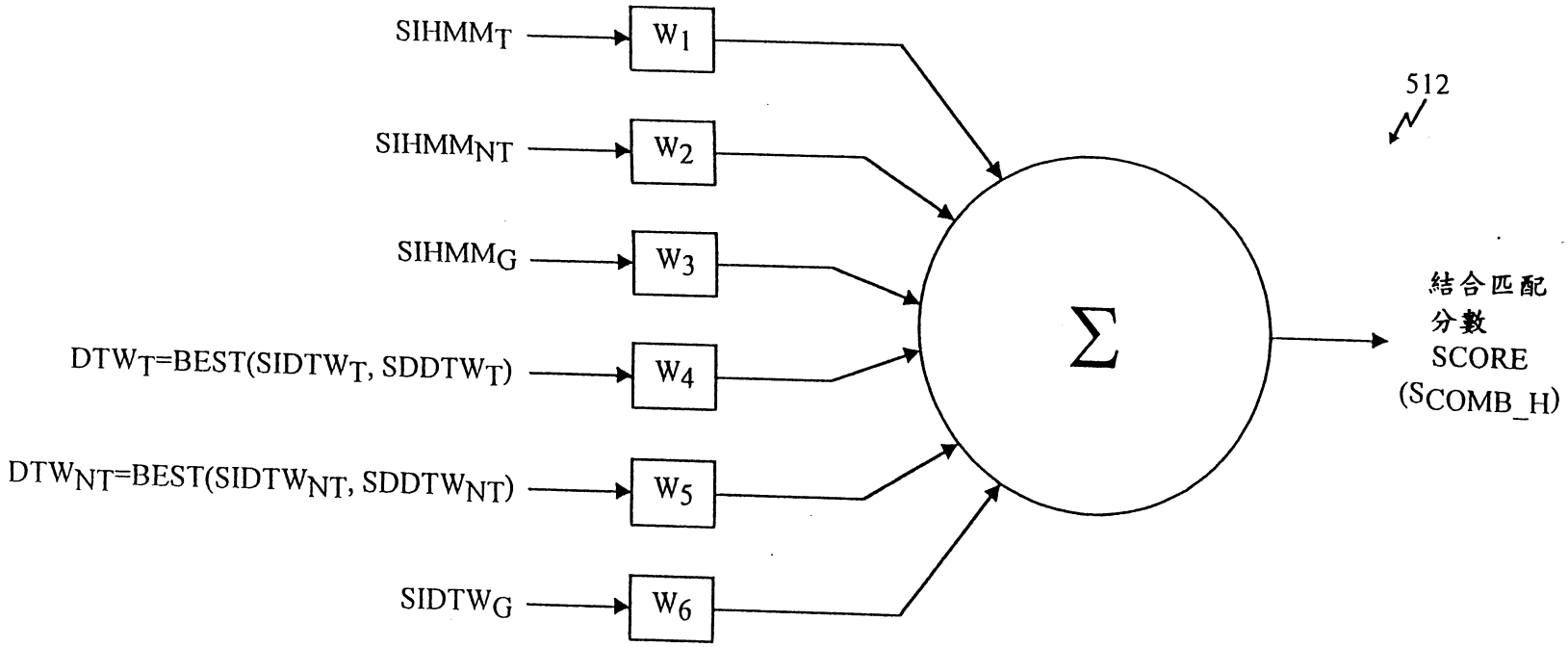

302
取樣類比
聲音資料
304
準備濾波
306
選取特徵
308
產生SIHMM及
SIDTW分數
510
產生SD分數
512
產生結合分數
514
挑選具有最佳結
合分數的談話

圖 5



EQN. 2 $S_{COMB\_H} = (SIHMM_T*W_1)+(SIHMM_{NT}*W_2)+(SIHMM_G*W_3)+(DTW_T*W_4)+(DTW_{NT}*W_5)+(SIDTW_G*W_6)$

圖 6

五、發明說明（ 20 ）

圖式元件符號說明

102 預先強調濾波器

104 聲學特徵選取單元

110 樣本匹配引擎

112 VR 聲學模型

202 無線遠端台

210 麥克風

212 類比至數位轉換器

214 預先強調濾波器

216 聲學特徵選取單元

218 數據機

220 VR 引擎

222 控制處理器

224 天線

230 聲學模型

232 聲學模型

234 聲學模型

六、申請專利範圍

1. 一種聲音辨識裝置，包括：

一說話者獨立聲學模型；

一說話者相依聲學模型；

一聲音辨識引擎；以及

一電腦可讀媒體，用以具體實施執行未監督聲音辨識訓練及測試的一方法，該方法包括執行具有該說話者獨立聲學模型的內容之輸入語音的樣本匹配，以產生說話者獨立樣本匹配分數、將說話者獨立樣本匹配分數與相關於儲存於該說話者相依聲學模型中的樣板之分數做比較、以及基於該比較的結果而更新該說話者相依聲學模型中的至少一樣板。

2. 如申請專利範圍第1項之聲音辨識裝置，其中該說話者獨立聲學模型包括至少一隱藏式馬克沃夫模型(HMM)聲學模型。

3. 如申請專利範圍第1項之聲音辨識裝置，其中該說話者獨立聲學模型包括至少一動態時間歪曲(DTW)聲學模型。

4. 如申請專利範圍第1項之聲音辨識裝置，其中該說話者獨立聲學模型包括至少一隱藏式馬克沃夫模型(HMM)聲學模型及至少一動態時間歪曲(DTW)聲學模型。

5. 如申請專利範圍第1項之聲音辨識裝置，其中該說話者獨立聲學模型包括至少一無用樣板，其中該比較包括比較該至少一無用樣板的輸入語音。

6. 如申請專利範圍第1項之聲音辨識裝置，其中該說話者相

A8
B8
C8
D8

六、申請專利範圍

依聲學模型包括至少一動態時間歪曲(DTW)聲學模型。

7. 一種聲音辨識裝置，包括：

一說話者獨立聲學模型；

一說話者相依聲學模型；

一聲音辨識引擎；以及

一電腦可讀媒體，用以具體實施執行未監督聲音辨識訓練及測試的一方法，該方法包括執行具有該說話者獨立聲學模型的內容之一第一輸入語音片段之樣本匹配，以產生說話者獨立樣本匹配分數；將說話者獨立樣本匹配分數與相關於儲存於該說話者相依聲學模型中的樣本之分數做比較；基於該比較的結果而更新該說話者相依聲學模型中的至少一樣板；配置該聲音辨識引擎，用以將一第二輸入語音片段與該說話者獨立聲學模型及該說話者相依聲學模型的內容做比較，以產生至少一結合的說話者相依及說話者獨立匹配分數；以及識別具有最佳結合的說話者相依及說話者獨立匹配分數的一聲階。

8. 如申請專利範圍第7項之聲音辨識裝置，其中該說話者獨立聲學模型包括至少一隱藏式馬克沃夫模型(HMM)聲學模型。

9. 如申請專利範圍第7項之聲音辨識裝置，其中該說話者獨立聲學模型包括至少一動態時間歪曲(DTW)聲學模型。

10. 如申請專利範圍第7項之聲音辨識裝置，其中該說話者獨立聲學模型包括至少一隱藏式馬克沃夫模型(HMM)聲

六、申請專利範圍

學模型及至少一動態時間歪曲(DTW)聲學模型。

11. 如申請專利範圍第7項之聲音辨識裝置，其中該說話者相依聲學模型包括至少一動態時間歪曲(DTW)聲學模型。

12. 一種聲音辨識裝置，包括：

一說話者獨立聲學模型；

一說話者相依聲學模型；以及

一聲音辨識引擎，用以執行具有該說話者獨立聲學模型的內容之輸入語音的樣本匹配，以產生說話者獨立樣本匹配分數，以及用以執行具有該說話者相依聲學模型的內容之輸入語音的樣本匹配，以產生說話者相依樣本匹配分數，以及用以產生基於該些說話者獨立樣本匹配分數及該些說話者相依樣本匹配分數之複數個聲階結合匹配分數。

13. 如申請專利範圍第12項之聲音辨識裝置，其中該說話者獨立聲學模型包括至少一隱藏式馬克沃夫模型(HMM)聲學模型。

14. 如申請專利範圍第12項之聲音辨識裝置，其中該說話者獨立聲學模型包括至少一動態時間歪曲(DTW)聲學模型。

15. 如申請專利範圍第12項之聲音辨識裝置，其中該說話者獨立聲學模型包括至少一隱藏式馬克沃夫模型(HMM)聲學模型及至少一動態時間歪曲(DTW)聲學模型。

16. 如申請專利範圍第12項之聲音辨識裝置，其中該說話

六、申請專利範圍

者相依聲學模型包括至少一動態時間歪曲(DTW)聲學模型。

17.一種執行聲音辨識的方法，包括：

執行具有至少一說話者獨立聲學樣板之一第一輸入語音片段的樣本匹配，以產生至少一輸入樣本匹配分數；

將該至少一輸入樣本匹配分數與相關於一儲存聲學樣板的一儲存分數做比較；以及

基於該比較的結果而取代該儲存聲學樣板。

18.如申請專利範圍第17項之方法，其中該執行樣本匹配進一步包括：

執行具有至少一HMM樣板之該第一輸入語音片段的隱藏式馬克沃夫模型(HMM)樣板樣本匹配，以產生至少一HMM匹配分數；

執行具有至少一DTW樣板之該第一輸入語音片段的動態時間歪曲(DTW)樣本匹配，以產生至少一DTW匹配分數；以及

執行該至少一HMM匹配分數及該至少一DTW匹配分數之至少一權重和，以產生該至少一輸入樣本匹配分數。

19.如申請專利範圍第17項之方法，進一步包括：

執行具有至少一說話者獨立聲學樣板之一第二輸入語音片段的樣本匹配，以產生至少一說話者獨立匹配分數；

執行具有該儲存聲學樣板之該第二輸入語音片段的

A8
B8
C8
D8

六、申請專利範圍

樣本匹配，以產生一說話者相依匹配分數；以及

將該至少一說話者獨立匹配分數與該說話者相依匹配分數結合，以產生至少一結合匹配分數。

20. 如申請專利範圍第19項之方法，進一步包括識別一聲階，其係關於最佳的至少一結合匹配分數。

21. 一種執行聲音辨識的方法，包括：

執行具有至少一說話者獨立聲學樣板之一輸入語音片段的樣本匹配，以產生至少一說話者獨立匹配分數；

執行具有一說話者相依聲學樣板之該輸入語音片段的樣本匹配，以產生至少一說話者相依匹配分數；以及

將該至少一說話者獨立匹配分數與該至少一說話者相依匹配分數結合，以產生至少一結合匹配分數。

22. 一種執行聲音辨識的方法，包括：

將一組輸入聲學特徵向量與在一說話者獨立聲學模型中的一說話者獨立樣板做比較，以產生一說話者獨立樣本匹配分數，其中該說話者獨立樣板係相關於一第一聲階；

將該組輸入聲學特徵向量與在一說話者相依聲學模型中的至少一說話者相依樣板做比較，以產生一說話者相依樣本匹配分數，其中該說話者相依樣板係相關於該第一聲階；

將該說話者獨立樣本匹配分數與該說話者相依樣本匹配分數結合，以產生一結合樣本匹配分數；以及

六、申請專利範圍

將該結合樣本匹配分數與相關於一第二聲階之至少一其他的結合樣本匹配分數做比較。

23.一種執行聲音辨識的裝置，包括：

執行具有至少一說話者獨立聲學樣板之一第一輸入語音片段的樣本匹配，以產生至少一輸入樣本匹配分數之裝置；

將該至少一輸入樣本匹配分數與相關於一儲存聲學樣板的一儲存分數做比較之裝置；以及

基於該比較的結果而取代該儲存聲學樣板之裝置。

24.一種執行聲音辨識的裝置，包括：

執行具有至少一說話者獨立聲學樣板之一輸入語音片段的樣本匹配，以產生至少一說話者獨立匹配分數之裝置；

執行具有一說話者相依聲學樣板之該輸入語音片段的樣本匹配，以產生至少一說話者相依匹配分數之裝置；以及

將該至少一說話者獨立匹配分數與該至少一說話者相依匹配分數結合，以產生至少一結合匹配分數之裝置。