I406266

(19)中華民國智慧財產局

(12)發明說明書公告本 (11)證書號數：TW I406266B1

(45)公告日： 中華民國 102 (2013) 年 08 月 21 日

(21)申請案號：100119556 (22)申請日： 中華民國 100 (2011) 年 06 月 03 日

(51)Int. Cl.： *G10L15/02 (2006.01)*

(71)申請人：國立交通大學 (中華民國) NATIONAL CHIAO TUNG UNIVERSITY (TW)
新竹市大學路 1001 號

(72)發明人：胡竹生 HU, JWU SHENG (TW) ；李明唐 LEE, MING TANG (TW) ；王庭昭 WANG, TING CHAO (TW) ；楊佳興 YANG, CHIA HSIN (TW)

(74)代理人：林火泉

(56)參考文獻：

TW I262433 TW 200916812A

TW 200916813A

Michael L. Seltzer, et al., "Likelihood-Maximizing Beamforming for Robust Hands-Free Speech Recognition," IEEE Transactions on Speech and Audio Processing, Vol. 12, No. 5, Sep. 2004, pp. 489-498.

審查人員：黃衍勳

申請專利範圍項數：33 項 圖式數：10 共 0 頁

(54)名稱

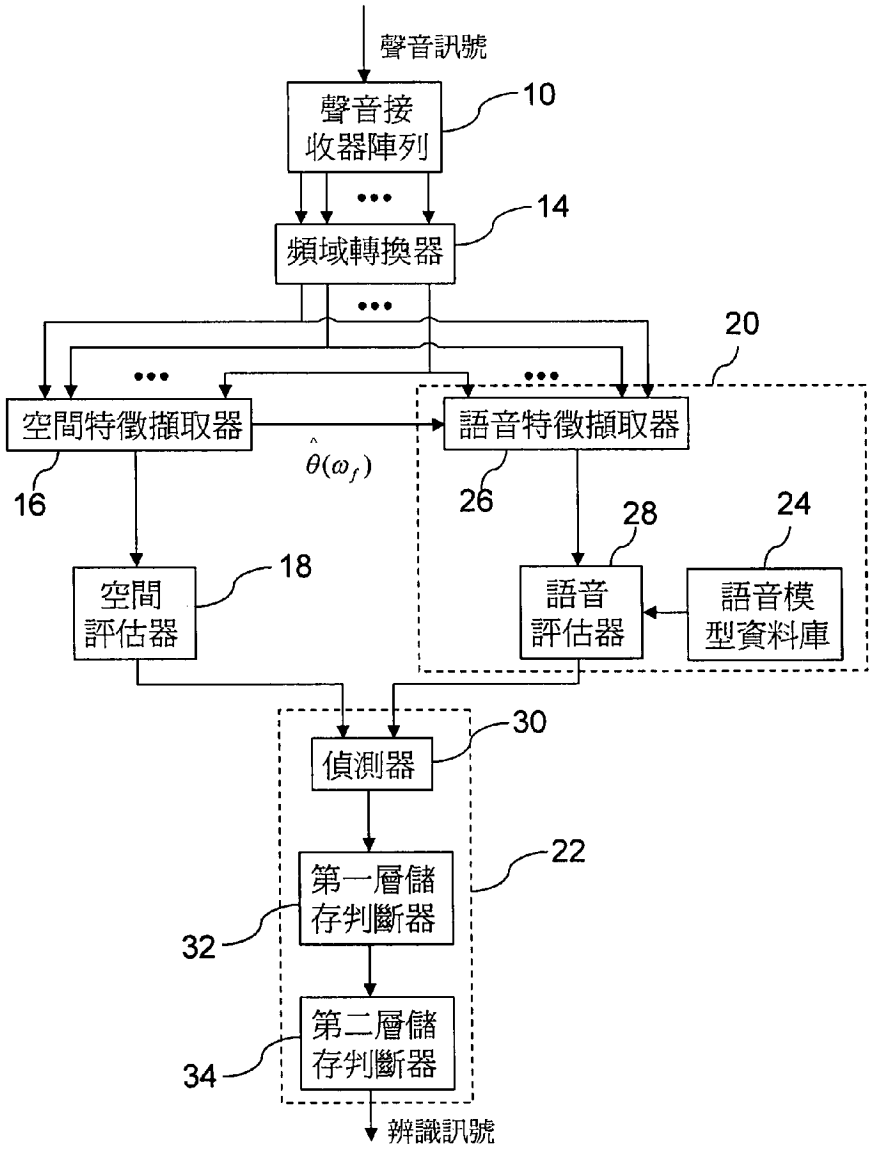
語音辨識裝置及其辨識方法

SPEECH RECOGNITION DEVICE AND A SPEECH RECOGNITION METHOD THEREOF

(57)摘要

本發明係揭露一種語音辨識裝置及其語音辨識方法，其係接收至少一關鍵詞以進行辨識，此關鍵詞包括至少一關鍵字。首先，接收關鍵字之一聲源訊號，以產生複數聲音訊號，進而將其轉換至頻域，形成複數聲頻訊號。接著，藉聲頻訊號擷取出一空間頻譜及其角度估測值。再來，利用空間頻譜定義至少一空間特徵參數輸出之，並依據角度估測值與聲頻訊號進行擷取與評估，輸出一巴塔恰里雅距離。最後，接收上述特徵與距離，以利用與其對應之檢測門檻值判斷關鍵詞之正確性，如此便可在極低訊噪比之環境下，仍達成相當強健的語音辨識率。

A speech recognition device and a speech recognition method thereof are disclosed. In the speech recognition method, at least one key word is received to be recognized, wherein the key word comprises at least one letter. Firstly, a sound source signal of the keyword is received to generate a plurality of sound signals. Then, the sound signals are converted into a plurality of sound frequency signals in the frequency domain. Next, a spatial spectrum and an angle estimation thereof are retrieved by the sound frequency signals. Next, at least one spatial feature defined by the spatial spectrum is outputted. A retrieval and estimation are executed by the angle estimation and the sound frequency signals to output the Bhattacharyya distance. Finally, the feature and the distance are received and detection thresholds corresponding to them are used to determine the correctness of the key word. As a result, a high recognition rate is achieved under very low signal-noise ratio (SNR) conditions.



第 1 圖

10・・・聲音接收器陣列

14・・・頻域轉換器

16・・・空間特徵擷取器

18・・・空間評估器

20・・・語音特徵擷取評估裝置

22・・・偵測裝置

24・・・語音模型資料庫

26・・・語音特徵擷取器

28・・・語音評估器

30・・・偵測器

32・・・第一層儲存判斷器

34・・・第二層儲存判斷器

發明專利說明書

公告本

（本說明書格式、順序，請勿任意更動，※記號部分請勿填寫）

※申請案號：100119556

※申請日：100. 6. 03　　※IPC 分類：G10L 15/02 (2006.01)

一、發明名稱：（中文/英文）

語音辨識裝置及其辨識方法 / speech recognition device and a speech recognition method thereof

二、中文發明摘要：

本發明係揭露一種語音辨識裝置及其語音辨識方法，其係接收至少一關鍵詞以進行辨識，此關鍵詞包括至少一關鍵字。首先，接收關鍵字之一聲源訊號，以產生複數聲音訊號，進而將其轉換至頻域，形成複數聲頻訊號。接著，藉聲頻訊號擷取出一空間頻譜及其角度估測值。再來，利用空間頻譜定義至少一空間特徵參數輸出之，並依據角度估測值與聲頻訊號進行擷取與評估，輸出一巴塔恰里雅距離。最後，接收上述特徵與距離，以利用與其對應之檢測門檻值判斷關鍵詞之正確性，如此便可在極低訊噪比之環境下，仍達成相當強健的語音辨識率。

三、英文發明摘要：

A speech recognition device and a speech recognition method thereof are disclosed. In the speech recognition method, at least one key word is received to be recognized, wherein the key word comprises at least one letter. Firstly, a sound source signal of the keyword is received to generate a plurality of sound signals. Then, the sound signals are converted into a plurality of sound frequency signals in the frequency domain. Next, a spatial spectrum and an angle estimation thereof are retrieved by the sound frequency signals. Next, at least one spatial feature defined by the spatial spectrum is outputted. A retrieval and estimation are executed by the angle estimation and the sound frequency signals to output the Bhattacharyya distance. Finally, the feature and the distance are received and detection thresholds corresponding to them are used to determine the correctness of the key word. As a result, a high recognition rate is achieved under very low signal-noise ratio (SNR) conditions.

四、指定代表圖：

(一)本案指定代表圖為：第（ 1 ）圖。

(二)本代表圖之元件符號簡單說明：

10	聲音接收器陣列	14	頻域轉換器
16	空間特徵擷取器	18	空間評估器
20	語音特徵擷取評估裝置	22	偵測裝置
24	語音模型資料庫	26	語音特徵擷取器
28	語音評估器	30	偵測器
32	第一層儲存判斷器	34	第二層儲存判斷器

五、本案若有化學式時，請揭示最能顯示發明特徵的化學式：

六、發明說明：

【發明所屬之技術領域】

本發明係有關一種辨識技術，特別是關於一種語音辨識裝置及其語音辨識方法。

【先前技術】

在一般語音訊號處理上，關鍵字偵測或是擷取(Keyword Detection or Spotting)是語音辨識(Speech Recognition)相當重要的一環，辨識步驟主要為先擷取語音特徵參數、為語音特徵建出模型及設定特徵參數比對方法(計算距離或相似度)。儘管語音辨識技術已經發展多年，在訊噪比高的情形下對大型詞彙庫的辨識率已經相當不錯，然而面對環境的雜訊干擾或是多人同時發聲的情況，即使是單一關鍵字的辨識率，也大多很難維持一定的水準。在現實環境中，各種不同的聲音干擾是無法避免的。

於現有技術之自動語音辨識系統中(Automatic Speech Recognition，ASR)，何時可以開始進行辨識是其中一項重要的功能，該功能通常稱作按鈕(push button)或是喚醒(wake-up)。Wake-up 功能運用得宜可以大量降低辨識錯誤率。一般在如電腦或手機的介面中往往以觸控或按鈕來實現，但是這個前提是所面對的裝置或機器需要在使用者的手邊。如果與使用者有一段距離，使用者往往必須配戴一無線裝置以提供可靠的 wake-up 訊號，在許多實際應用上這仍有其障礙。例如要命令智慧型居家服務機器人提供服務，若使用者必須一直配戴一無線裝置，在居家的情境中幾乎是不可行。因此，如何能夠在無需配戴任何裝置的情形下有效的實現 wake-up 功能，就成為一個實用且富挑戰性的研發題目。因為使用者不能配戴任何裝置，

且提供語音辨識介面的機器很可能不在視野範圍內，因此無可避免的必須回歸到以語音來執行 wake-up 的功能。簡單來說，這即是單一關鍵字的辨識問題，但是其所面臨的問題是語者可能距離相當遠，或位於吵雜的環境中，因此訊噪比通常很差。其次是如同按鈕或觸控，以語音關鍵字實現 wake-up 也必須有幾乎 100%的偵測率(detection rate)以及接近於 0 的偽陽性比率(false positive rate)，否則將產生誤動作或反應遲鈍。

因此，本發明係在針對上述之困擾，提出一種語音辨識裝置及其語音辨識方法，以解決習知所產生的問題。

【發明內容】

本發明之主要目的，在於提供一種語音辨識裝置及其語音辨識方法，其係運用聲源於聲音接收器陣列的特徵空間一致性，以及關鍵字語音特徵相似度，同時利用偵測裝置結合各別特徵辨識結果以計算出所指定之語音關鍵字是否有存在的機制。經大量的語料測試，此語音辨識技術可在-3.82 分貝(dB)的訊噪比之下達成 100%的偵測率(detection rate)以及 10.32%的偽陽性比率(false positive rate)。

為達上述目的，本發明提供一種語音辨識裝置，其係接收至少一關鍵詞以進行辨識，此關鍵詞包括至少一關鍵字，語音辨識裝置包括一聲音接收器陣列，用以接收關鍵字之一聲源訊號，以產生複數聲音訊號。聲音接收器陣列連接一頻域轉換器，其係接收聲音訊號，並將其轉換至頻域，形成複數聲頻訊號。頻域轉換器連接一空間特徵擷取器與一語音特徵擷取評估裝置，且空間特徵擷取器與語音特徵擷取評估裝置相互連接。空間特徵擷取器接收聲頻訊號，以藉此擷取出一空間頻譜及其角度估測值，另語音

特徵擷取評估裝置接收角度估測值與聲頻訊號，以據此進行擷取與評估，輸出一巴塔恰里雅(Bhattacharyya)距離。空間特徵擷取器更連接一空間評估器，其係接收空間頻譜，以定義至少一空間特徵參數輸出之。空間評估器與語音特徵擷取評估裝置皆連接一偵測裝置，其係預設有與空間特徵參數與巴塔恰里雅距離對應之檢測門檻值，此偵測裝置接收空間特徵參數與巴塔恰里雅距離，並利用檢測門檻值判斷關鍵詞之正確性。

本發明亦提供一種語音辨識方法，其係接收至少一關鍵詞以進行辨識，此關鍵詞包括至少一關鍵字。首先，接收關鍵字之一聲源訊號，以產生複數聲音訊號，進而將其轉換至頻域，形成複數聲頻訊號。接著，接收聲頻訊號，以藉此擷取出一空間頻譜及其角度估測值。再來，接收空間頻譜、角度估測值與聲頻訊號，以利用空間頻譜定義至少一空間特徵參數輸出之，並依據角度估測值與聲頻訊號進行擷取與評估，輸出一巴塔恰里雅距離。最後，接收空間特徵參數與巴塔恰里雅距離，以利用與其對應之檢測門檻值判斷關鍵詞之正確性。

茲為使　貴審查委員對本發明之結構特徵及所達成之功效更有進一步之瞭解與認識，謹佐以較佳之實施例圖及配合詳細之說明，說明如後：

【實施方式】

本發明的要點乃基於目標關鍵字語音共鳴曲線相似度(Resonant Curve Similarity)的波形特徵，同時亦須具備空間特徵一致性(Spatial Eigenspace Consistency)，例如某一關鍵詞可能包含三個關鍵字，以特定之先後順序組合而成。因此這三個關鍵字必須有同樣的特徵空間，若是以聲音傳遞到麥克風陣列的情況而言，代表這三個字必須為同一聲源來向。其次，一旦這

5

個關係符合，其所擷取的特徵空間訊號便可以用來進一步計算其與目標關鍵字的波型相似度，此一優點為特徵空間的訊號較不易受到環境干擾，因而可以大量提升其語音辨識度，換言之，此技術可用在遠距關鍵字語音偵測或者在吵雜的環境下，作為關鍵字語音喚醒機制。實施方式如下所述：

請參閱第 1 圖與第 2 圖，本發明之語音辨識裝置，係接收至少一關鍵詞以進行辨識，關鍵詞包括至少一關鍵字，且關鍵字具有複數音框。語音辨識裝置包括一聲音接收器陣列 10，如麥克風陣列，聲音接收器陣列 10 係包括複數個接收器 12，並呈環形排列，並位於一平面上，此環形排列具有一中心點，中心點與環形排列之周邊相距 R，此平面以中心點為原點，定義出互相垂直之橫軸 X、直軸 Y 與縱軸 Z。聲音接收器陣列 10 係連續接收關鍵字之每一音框之一聲源訊號，以產生複數聲音訊號。聲源訊號之聲源點與上述中心點及平面係形成一垂直上述平面之三角面，此三角面之底邊與橫軸 X 夾有 φ 角，斜邊與縱軸 Z 夾有 θ 角。聲音接收器陣列 10 連接一頻域轉換器 14，如快速傅立葉轉換(FFT)器或離散餘弦轉換(discrete cosine transform, DCT)器，頻域轉換器 14 係接收聲音訊號，並將其轉換至頻域，形成複數聲頻訊號。

頻域轉換器 14 連接一空間特徵擷取器 16，且空間特徵擷取器 16 與一空間評估器 18 連接，空間特徵擷取器 16 接收聲頻訊號，以藉此擷取出一空間頻譜及其角度估測值 $\hat{\theta}(\omega_f) = \arg\max_{\theta} S(\theta, \omega_f)$，其中空間頻譜可表示為 $S(\theta, \omega_f) = \frac{1}{a^H(\theta, \omega_f) P_N(\omega_f) a(\theta, \omega_f)}$，其中 $f = 1 \ldots F$，F 代表快速傅立葉轉換尺寸(FFT size)，ω_f 為頻率，$a(\theta, \omega_f)$ 與 $a^H(\theta, \omega_f)$ 分別為進行轉置(tranpose)

及共軛(conjugate)之 $a^T(\theta,\omega_f)$ ， $a^T(\theta,\omega_f)$

$=[1,e^{j\omega_f * R\sin t\cos\theta/c},e^{j\omega_f * R\sin\varphi\cos(\theta-2\pi/M)/c},\cdots,e^{j\omega_f * R\sin\varphi\cos(\theta-2(M-1)\pi/M)/c}]$

， t 為時間， c 為光速， M 為聲音接收器陣列 10 之接收器 12 數量，

$P_N(\omega_f)=\sum_{i=2}^{M}V(\omega_f)_i V(\omega_f)_i^H$ ， $V(\omega_f)_i$ 、 $V(\omega_f)_i^H$ 為利用聲頻訊號得到之資料相關矩陣 R_{XX} 之特徵向量；資料相關矩陣可表示為 $R_{XX}(\omega_f,k)=E(X(\omega_f,k),X(\omega_f,k)^H)$ ， $X(\omega_f,k)$ 為聲頻訊號，且資料相關矩陣 $R_{XX}(\omega_f)=\sum_{i=1}^{M}\lambda_i(\omega_f)V_i(\omega_f)V_i^H(\omega_f)$ ， $\lambda_i(\omega_f)$ 為資料相關矩陣之特徵值。另外，空間評估器 18 則接收空間頻譜，以定義至少一空間特徵參數輸出之。此外，空間評估器 18 更可同時接收空間頻譜與角度估測值，以分別定義二空間特徵參數輸出之，且在此實施例中，係以此種方式為例，其中空間頻譜定義之空間特徵參數為角度估測量值 $x_1=\max_{\theta}[\frac{\sum_{f\in F'}S(\theta,\omega_f)}{D}]$ ， D 為正規化因子(normalized factor)， F' 為共振峰對應之頻帶(formant frequency bands)；角度估測值定義之空間特徵參數為角度估測變異數 $x_2=\mathrm{var}(\hat{\theta}(\omega_f))$ ，且 $f\in F'$ 。第 3(a)圖與第 3(b)圖分別表示關鍵字與非關鍵字之角度估測量值之統計分佈，由於角度估測量值為分佈圖中的峰值處，所以若偵測字為關鍵字時，角度估測量值較大，為非關鍵字時，角度估測量值較小。另第 4(a)圖與第 4(b)圖分別表示關鍵字與非關鍵字之角度估測變異數之統計分佈，由於角度估測變異數為分佈圖中的離散程度，因此若偵測字為關鍵字時，角度估測變異數較小，為非關鍵字時，角度估測變異數較大。換言之，由角度估測量值與角度估測變異數可以驗證聲源訊號之特徵空間一致性。

繼續如第 1 圖所示，空間特徵擷取器 16 與頻域轉換器 14 連接一語音特徵擷取評估裝置 20，其係接收角度估測值與聲頻訊號，以據此進行擷取與評估，輸出一巴塔恰里雅(Bhattacharyya)距離。空間評估器 18 與語音特徵擷取評估裝置 20 更連接一偵測裝置 22，其係預設有與空間特徵參數與巴塔恰里雅距離對應之檢測門檻值，偵測裝置 22 接收空間特徵參數與巴塔恰里雅距離，並利用檢測門檻值判斷關鍵詞之正確性。

語音特徵擷取評估裝置 20 更包括一語音模型資料庫 24、一語音特徵擷取器 26 與一語音評估器 28，語音模型資料庫 24 係存有複數種語音共鳴模型資料。語音特徵擷取器 26 連接空間特徵擷取器 16 與頻域轉換器 14，並接收角度估測值與聲頻訊號，以據此擷取一語音特徵資料。上述語音共鳴模型資料可分別為語音共鳴模型曲線峰值或語音共鳴模型曲線兩種，為了對應此兩種模型資料，則語音特徵資料亦分別為語音特徵數值或語音特徵曲線。語音評估器 28 連接語音特徵擷取器 26 與語音模型資料庫 24，並接收語音特徵資料，語音評估器 28 從語音模型資料庫 24 取出與語音特徵資料對應之語音共鳴模型資料，以對語音特徵資料進行評估後，輸出巴塔恰里雅距離。在此實施例中，語音共鳴模型資料與語音特徵資料分別以語音共鳴模型曲線及語音特徵曲線為例，如第 5(a)圖與第 5(b)圖所示，在此兩張圖中，皆分別有兩條曲線，其一在上，為語音共鳴模型曲線，另一在下，為語音特徵曲線，語音評估器 28 會將兩種曲線進行評估，以輸出巴塔恰里雅距離，其可表示為 $BC(p,q)=\int\sqrt{p(x)q(x)}dx$，$p(x)$ 為語音特徵曲線，$q(x)$ 為語音共鳴模型曲線。此外，當語音共鳴模型資料與語音特徵資料分別為語音共鳴模型曲線峰值及語音特徵數值時，語音共鳴模型曲線峰值及語音特徵

數值分別代表語音共鳴模型曲線及語音特徵曲線之波峰處的強度。

請繼續參閱第 1 圖，偵測裝置 22 更包括作為一偵測器 30 之串聯式偵測器、一第一層、第二層儲存判斷器 32、34。偵測器 30 連接空間評估器 18 與語音特徵擷取評估裝置 20 之語音評估器 28，並接收每一音框對應之空間特徵參數與巴塔恰里雅距離，偵測器 30 預設有檢測門檻值，並據此對空間特徵參數與巴塔恰里雅距離進行檢測，以輸出分別代表正確與錯誤之音框的一第一層正確旗標或一第一層錯誤旗標。舉例來說，偵測器 30 在空間特徵參數與巴塔恰里雅距離中，至少其中一者小於或等於對應之檢測門檻值時，則輸出第一層錯誤旗標，在空間特徵參數與巴塔恰里雅距離皆大於對應之該檢測門檻值時，則輸出第一層正確旗標。

第一層儲存判斷器 32 連接偵測器 30，並接收每一音框之第一層正確旗標或第一層錯誤旗標，以儲存之，在關鍵字對應之所有第一層正確旗標及第一層錯誤旗標儲存完後，則據其數量輸出分別代表正確與錯誤之關鍵字的一第二層正確旗標或一第二層錯誤旗標。舉例來說，第一層儲存判斷器 32 預設有一第一層檢測值，在關鍵字對應之所有第一層正確旗標及第一層錯誤旗標中，第一層儲存判斷器 32 利用第一層檢測值檢測第一層正確旗標之比例，當此比例大於第一層檢測值時，第一層儲存判斷器 32 輸出第二層正確旗標，當比例小於或等於第一層檢測值時，第一層儲存判斷器 32 輸出第二層錯誤旗標。

第二層儲存判斷器 34 連接第一層儲存判斷器 32，在關鍵字數量為二以上時，第二層儲存判斷器 34 接收每一關鍵字之第二層正確旗標或第二層錯誤旗標，以儲存之，並在關鍵詞對應之所有第二層正確旗標及第二層錯誤

旗標儲存完後，則據其數量輸出分別代表正確與錯誤之關鍵詞的一正確辨識訊號或一錯誤辨識訊號。舉例來說，第二層儲存判斷器 34 預設有一第二層檢測值，在關鍵詞對應之所有第二層正確旗標及第二層錯誤旗標中，第二層儲存判斷器 34 利用第二層檢測值檢測第二層正確旗標之比例，當此比例大於第二層檢測值時，第二層儲存判斷器 34 輸出正確辨識訊號，當比例小於或等於第二層檢測值時，第二層儲存判斷器 34 輸出錯誤辨識訊號。由於關鍵字都會來自同一聲源方向，因此，本發明運用聲源於聲音接收器陣列的特徵空間一致性，以及關鍵字語音特徵相似度，同時利用偵測裝置結合各別特徵辨識結果以計算出所指定之語音關鍵字是否有存在的機制。經大量的語料測試，此語音辨識技術可在-3.82 分貝(dB)的訊噪比之下達成 100%的偵測率(detection rate)以及 10.32%的偽陽性比率(false positive rate)，換言之，本發明之技術在極低訊噪比之環境下，仍達成相當強健的語音辨識率，因而可以適用在遠距關鍵字語音偵測或者在吵雜的環境下，作為關鍵字語音喚醒機制。

請同時參閱第 6 圖，以下介紹語音辨識方法之流程。

首先，如步驟 S10 所示，聲音接收器陣列 10 連續接收關鍵字之每一音框之聲源訊號，以產生聲音訊號。接著，如步驟 S12 所示，頻域轉換器 14 接收聲音訊號，並將其轉換至頻域，以形成聲頻訊號。再來，如步驟 S14 所示，空間特徵擷取器 16 接收聲頻訊號，以藉此擷取出空間頻譜及其角度估測值。接續之，如步驟 S16 所示，空間評估器 18 接收空間頻譜，以利用空間頻譜定義至少一空間特徵參數輸出之，同時，語音特徵擷取評估裝置 20 接收角度估測值與聲頻訊號，並依據角度估測值與聲頻訊號進行擷取與

評估，輸出巴塔恰里雅距離，其中擷取的方式可採用採用線性預估編碼(Linear Predictive Coding, LPC)法或梅爾倒頻譜係數(Mel-scale Frequency Cepstral Coefficients, MFCC)法。此外，在步驟 S16 中，空間評估器 18 除了接收空間頻譜，以定義至少一空間特徵參數輸出之之外，亦可同時接收空間頻譜與角度估測值，以定義至少二空間特徵參數輸出之，且在此實施例中，係以此種方式為例，其中由空間頻譜定義之空間特徵參數為角度估測量值；由角度估測值定義之空間特徵參數為角度估測變異數。最後，如步驟 S18 所示，偵測裝置 22 接收空間特徵參數與巴塔恰里雅距離，以利用與其對應之檢測門檻值判斷關鍵詞之正確性。

在上述流程中，其中語音特徵擷取評估裝置 20 在依據角度估測值與聲頻訊號進行擷取與評估，以輸出巴塔恰里雅距離之步驟中，更可以下列步驟實施之。首先，語音特徵擷取器 26 係依據角度估測值與聲頻訊號擷取一語音特徵資料。接著，語音評估器 28 接收語音特徵資料，並從存於語音模型資料庫 24 裡的複數種語音共鳴模型資料中，取出與語音特徵資料對應之語音共鳴模型資料，以對語音特徵資料進行評估後，輸出巴塔恰里雅距離。

另外，偵測裝置 22 利用檢測門檻值判斷關鍵詞之正確性之步驟更可以下列步驟實施之。請同時參閱第 1 圖與第 7 圖。首先，如步驟 S20 所示，偵測器 30 利用檢測門檻值對每一音框對應之空間特徵參數與巴塔恰里雅距離進行判斷，以輸出分別代表正確與錯誤之音框的第一層正確旗標或第一層錯誤旗標。此步驟 S20 可以下列子步驟實施之，例如，偵測器 30 在空間特徵參數與巴塔恰里雅距離中，判斷是否至少其中一者小於或等於對應之檢測門檻值，若是，輸出第一層錯誤旗標；若否，輸出第一層正確旗標。

接著，如步驟 S22 所示，第一層儲存判斷器 32 接收每一音框之第一層正確旗標或第一層錯誤旗標，以儲存之，直到關鍵字對應之所有第一層正確旗標及第一層錯誤旗標儲存完後，進行步驟 S24。在步驟 S24 中，第一層儲存判斷器 32 係根據所有第一層正確旗標及第一層錯誤旗標之數量，輸出分別代表正確與錯誤之關鍵字的第二層正確旗標或第二層錯誤旗標。此步驟 S24 可以下列子步驟實施之，例如，由於第一層儲存判斷器 32 預設第一層檢測值，因此，第一層儲存判斷器 32 於所有第一層正確旗標及第一層錯誤旗標中，判斷第一層正確旗標所佔之比例，是否大於第一層檢測值，若是，輸出第二層正確旗標；若否，輸出第二層錯誤旗標。

當關鍵字數量為二以上時，於步驟 S24 後，更可進行下列步驟，首先如步驟 S26 所示，第二層儲存判斷器 34 接收每一關鍵字之第二層正確旗標或第二層錯誤旗標，以儲存之，並在關鍵詞對應之所有第二層正確旗標及第二層錯誤旗標儲存完後，進行步驟 S28。在步驟 S28 中，第二層儲存判斷器 34 根據所有第二層正確旗標及第二層錯誤旗標之數量，輸出分別代表正確與錯誤之關鍵詞的正確辨識訊號或錯誤辨識訊號。此步驟 S28 可以下列子步驟實施之，例如，由於第二層儲存判斷器 34 預設第二層檢測值，因此，第二層儲存判斷器 34 於所有第二層正確旗標及第二層錯誤旗標中，判斷第二層正確旗標所佔之比例，是否大於第二層檢測值，若是，輸出正確辨識訊號；若否，輸出錯誤辨識訊號。

綜上所述，本發明結合語音關鍵字的空間一致性判斷與關鍵字語音共鳴曲線相似度的判斷，以提升關鍵字偵測的強健性。

以上所述者，僅為本發明一較佳實施例而已，並非用來限定本發明實

施之範圍，故舉凡依本發明申請專利範圍所述之形狀、構造、特徵及精神所為之均等變化與修飾，均應包括於本發明之申請專利範圍內。

【圖式簡單說明】

第 1 圖為本發明之裝置方塊圖。

第 2 圖為本發明之聲音接收器陣列與聲源點之相關位置立體圖。

第 3(a)圖為本發明之關鍵字之角度估測量值統計分佈圖。

第 3(b)圖為本發明之非關鍵字之角度估測量值統計分佈圖。

第 4(a)圖為本發明之關鍵字之角度估測變異數統計分佈圖。

第 4(b)圖為本發明之非關鍵字之角度估測變異數統計分佈圖。

第 5(a)圖與第 5(b)圖分別為本發明之不同關鍵字之語音共鳴模型曲線與語音特徵曲線波形圖。

第 6 圖為本發明之語音辨識方法流程圖。

第 7 圖為本發明之判斷關鍵詞之正確性之流程圖。

【主要元件符號說明】

10 聲音接收器陣列	12 接收器
14 頻域轉換器	16 空間特徵擷取器
18 空間評估器	20 語音特徵擷取評估裝置
22 偵測裝置	24 語音模型資料庫
26 語音特徵擷取器	28 語音評估器
30 偵測器	32 第一層儲存判斷器
34 第二層儲存判斷器	

七、申請專利範圍：

1. 一種語音辨識裝置，其係接收至少一關鍵詞以進行辨識，該關鍵詞包括至少一關鍵字，該語音辨識裝置包括：

一聲音接收器陣列，接收該關鍵字之一聲源訊號，以產生複數聲音訊號；

一頻域轉換器，連接該聲音接收器陣列，以接收該些聲音訊號，並將其轉換至頻域，形成複數聲頻訊號；

一空間特徵擷取器，連接該頻域轉換器，並接收該些聲頻訊號，以藉此擷取出一空間頻譜及其角度估測值；

一空間評估器，連接該空間特徵擷取器，並接收該空間頻譜，以定義至少一空間特徵參數輸出之；

一語音特徵擷取評估裝置，其係連接該空間特徵擷取器與該頻域轉換器，並接收該角度估測值與該些聲頻訊號，以據此進行擷取與評估，輸出一巴塔恰里雅(Bhattacharyya)距離；以及

一偵測裝置，連接該空間評估器與該語音特徵擷取評估裝置，並預設有與該空間特徵參數與該巴塔恰里雅距離對應之檢測門檻值，該偵測裝置接收該空間特徵參數與該巴塔恰里雅距離，並利用該檢測門檻值判斷該關鍵詞之正確性。

2. 如請求項 1 所述之語音辨識裝置，其中該聲音接收器陣列呈環形排列，並位於一平面上，該環形排列具有一中心點，該中心點與該環形排列之周邊相距 R，該平面以該中心點為原點，定義出互相垂直之橫軸 X、直軸 Y 與縱軸 Z，該聲源訊號之聲源點與該中心點及該平面係形成一垂直該平面之三角面，該三角面之底邊與該橫軸夾有 φ 角，斜邊與該縱軸夾

有 θ 角；該空間頻譜可表示為 $S(\theta,\omega_f)=\frac{1}{a^H(\theta,\omega_f)P_N(\omega_f)a(\theta,\omega_f)}$，其中 $f=1\ldots F$，F 代表快速傅立葉轉換尺寸(FFT size)，ω_f 為頻率，$a(\theta,\omega_f)$ 與 $a^H(\theta,\omega_f)$ 分別為進行轉置(tranpose)及共軛(conjugate)之 $a^T(\theta,\omega_f)$，$a^T(\theta,\omega_f)$

$=[1,e^{j\omega_f*R\sin t\cos\theta/c},e^{j\omega_f*R\sin\varphi\cos(\theta-2\pi/M)/c},\cdots,e^{j\omega_f*R\sin\varphi\cos(\theta-2(M-1)\pi/M)/c}]$

，t 為時間，c 為光速，M 為該聲音接收器陣列之接收器數量，$P_N(\omega_f)=\sum_{i=2}^{M}V(\omega_f)_iV(\omega_f)_i^H$，$V(\omega_f)_i$、$V(\omega_f)_i^H$ 為利用該些聲頻訊號得到之資料相關矩陣 R_{XX} 之特徵向量；該資料相關矩陣可表示為 $R_{XX}(\omega_f,k)=E(X(\omega_f,k),X(\omega_f,k)^H)$，$X(\omega_f,k)$ 為該聲頻訊號，且該資料相關矩陣 $R_{XX}(\omega_f)=\sum_{i=1}^{M}\lambda_i(\omega_f)V_i(\omega_f)V_i^H(\omega_f)$，$\lambda_i(\omega_f)$ 為該資料相關矩陣之特徵值。

3. 如請求項 2 所述之語音辨識裝置，其中該角度估測值 $\hat{\theta}(\omega_f)=\arg\max_{\theta}S(\theta,\omega_f)$。

4. 如請求項 2 所述之語音辨識裝置，其中該空間評估器更同時接收該空間頻譜與該角度估測值，以分別定義二該空間特徵參數輸出之。

5. 如請求項 4 所述之語音辨識裝置，其中該空間頻譜定義之該空間特徵參數為角度估測量值 $x_1=\max_{\theta}[\frac{\sum_{f\in F'}S(\theta,\omega_f)}{D}]$，D 為正規化因子(normalized factor)，F' 為共振峰對應之頻帶(formant frequency bands)；以及該角度估測值定義之該空間特徵參數為角度估測變異數 $x_2=\mathrm{var}(\hat{\theta}(\omega_f))$，且 $f\in F'$。

6. 如請求項 1 所述之語音辨識裝置，其中該語音特徵擷取評估裝置更包括：

一語音模型資料庫，存有複數種語音共鳴模型資料；

一語音特徵擷取器，連接該空間特徵擷取器與該頻域轉換器，並接收該角度估測值與該些聲頻訊號，以據此擷取一語音特徵資料；以及

一語音評估器，連接該語音特徵擷取器與該語音模型資料庫，並接收該語音特徵資料，該語音評估器從該語音模型資料庫取出與該語音特徵資料對應之該語音共鳴模型資料，以對該語音特徵資料進行評估後，輸出該巴塔恰里雅距離。

7. 如請求項 6 所述之語音辨識裝置，其中該語音共鳴模型資料分別為語音共鳴模型曲線峰值或語音共鳴模型曲線時，該語音特徵資料分別為語音特徵數值或語音特徵曲線。

8. 如請求項 7 所述之語音辨識裝置，其中該巴塔恰里雅距離 $BC(p,q)=\int\sqrt{p(x)q(x)}dx$，$p(x)$ 為該語音特徵曲線，$q(x)$ 為該語音共鳴模型曲線。

9. 如請求項 1 所述之語音辨識裝置，其中該關鍵字具有複數音框，該聲音接收器陣列連續接收每一該音框之該聲源訊號，以供該頻域轉換器、該空間特徵擷取器、該空間評估器、該語音特徵擷取評估裝置及該偵測裝置運作之。

10. 如請求項 9 所述之語音辨識裝置，其中該偵測裝置更包括：

一偵測器，連接該空間評估器與該語音特徵擷取評估裝置，並接收每一該音框對應之該空間特徵參數與該巴塔恰里雅距離，該偵測器預設有該檢測門檻值，並據此對該空間特徵參數與該巴塔恰里雅距離進行檢

測，以輸出分別代表正確與錯誤之該音框的一第一層正確旗標或一第一層錯誤旗標；以及

一第一層儲存判斷器，連接該偵測器，並接收每一該音框之該第一層正確旗標或該第一層錯誤旗標，以儲存之，在該關鍵字對應之所有該第一層正確旗標及該第一層錯誤旗標儲存完後，則據其數量輸出分別代表正確與錯誤之該關鍵字的一第二層正確旗標或一第二層錯誤旗標。

11.如請求項 10 所述之語音辨識裝置，其中該偵測器在該空間特徵參數與該巴塔恰里雅距離中，至少其中一者小於或等於對應之該檢測門檻值時，則輸出該第一層錯誤旗標，在該空間特徵參數與該巴塔恰里雅距離皆大於對應之該檢測門檻值時，則輸出該第一層正確旗標。

12.如請求項 10 所述之語音辨識裝置，其中該第一層儲存判斷器預設有一第一層檢測值，在該關鍵字對應之所有該第一層正確旗標及該第一層錯誤旗標中，該第一層儲存判斷器利用該第一層檢測值檢測該第一層正確旗標之比例，該比例大於該第一層檢測值時，該第一層儲存判斷器輸出該第二層正確旗標，該比例小於或等於該第一層檢測值時，該第一層儲存判斷器輸出該第二層錯誤旗標。

13.如請求項 10 所述之語音辨識裝置，更包括一第二層儲存判斷器，其係連接該第一層儲存判斷器，在該關鍵字數量為二以上時，該第二層儲存判斷器接收每一該關鍵字之該第二層正確旗標或該第二層錯誤旗標，以儲存之，並在該關鍵詞對應之所有該第二層正確旗標及該第二層錯誤旗標儲存完後，則據其數量輸出分別代表正確與錯誤之該關鍵詞的一正確辨識訊號或一錯誤辨識訊號。

14.如請求項 13 所述之語音辨識裝置，其中該第二層儲存判斷器預設有一第二層檢測值，在該關鍵詞對應之所有該第二層正確旗標及該第二層錯誤旗標中，該第二層儲存判斷器利用該第二層檢測值檢測該第二層正確旗標之比例，該比例大於該第二層檢測值時，該第二層儲存判斷器輸出該正確辨識訊號，該比例小於或等於該第二層檢測值時，該第二層儲存判斷器輸出該錯誤辨識訊號。

15.如請求項 10 所述之語音辨識裝置，其中該偵測器為串聯式偵測器。

16.如請求項 1 所述之語音辨識裝置，其中該語音特徵擷取評估裝置採用線性預估編碼(Linear Predictive Coding, LPC)法或梅爾倒頻譜係數(Mel-scale Frequency Cepstral Coefficients, MFCC)法，以根據該角度估測值與該些聲頻訊號進行擷取。

17.如請求項 1 所述之語音辨識裝置，其中該聲音接收器陣列為麥克風陣列。

18.如請求項 1 所述之語音辨識裝置，其中該頻域轉換器為快速傅立葉轉換(FFT)器或離散餘弦轉換(discrete cosine transform, DCT)器。

19.一種語音辨識方法，其係接收至少一關鍵詞以進行辨識，該關鍵詞包括至少一關鍵字，該語音辨識方法包括下列步驟：

接收該關鍵字之一聲源訊號，以產生複數聲音訊號；

轉換該些聲音訊號至頻域，形成複數聲頻訊號；

接收該些聲頻訊號，以藉此擷取出一空間頻譜及其角度估測值；

接收該空間頻譜、該角度估測值與該些聲頻訊號，以利用該空間頻譜定義至少一空間特徵參數輸出之，並依據該角度估測值與該些聲頻訊號進行擷取與評估，輸出一巴塔恰里雅(Bhattacharyya)距離；以及

接收該空間特徵參數與該巴塔恰里雅距離，以利用與其對應之檢測門檻值判斷該關鍵詞之正確性。

20.如請求項19所述之語音辨識方法，其中該些聲音訊號由一聲音接收器陣列接收之，該聲音接收器陣列呈環形排列，並位於一平面上，該環形排列具有一中心點，該中心點與該環形排列之周邊相距R，該平面以該中心點為原點，定義出互相垂直之橫軸X、直軸Y與縱軸Z，該聲源訊號之聲源點與該中心點及該平面係形成一垂直該平面之三角面，該三角面之底邊與該橫軸夾有φ角，斜邊與該縱軸夾有θ角；該空間頻譜可表示為 $S(\theta,\omega_f)=\frac{1}{a^H(\theta,\omega_f)P_N(\omega_f)a(\theta,\omega_f)}$ ，其中 $f=1\ldots F$ ，F 代表快速傅立葉轉換尺寸(FFT size)，ω_f 為頻率，$a(\theta,\omega_f)$ 與 $a^H(\theta,\omega_f)$ 分別為進行轉置(tranpose)及共軛(conjugate)之 $a^T(\theta,\omega_f)$ ，$a^T(\theta,\omega_f)$ $=[1,e^{j\omega_f*R\sin t\cos\theta/c},e^{j\omega_f*R\sin\varphi\cos(\theta-2\pi/M)/c},\cdots,e^{j\omega_f*R\sin\varphi\cos(\theta-2(M-1)\pi/M)/c}]$ ，t 為時間，c 為光速，M為該聲音接收器陣列之接收器數量，$P_N(\omega_f)=\sum_{i=2}^{M}V(\omega_f)_iV(\omega_f)_i^H$ ，$V(\omega_f)_i$ 、$V(\omega_f)_i^H$ 為利用該些聲頻訊號得到之資料相關矩陣 R_{XX} 之特徵向量；該資料相關矩陣可表示為 $R_{XX}(\omega_f,k)=E(X(\omega_f,k),X(\omega_f,k)^H)$ ，$X(\omega_f,k)$ 為該聲頻訊號，且該資料相關矩陣 $R_{XX}(\omega_f)=\sum_{i=1}^{M}\lambda_i(\omega_f)V_i(\omega_f)V_i^H(\omega_f)$ ，$\lambda_i(\omega_f)$ 為該資料相關矩陣之特徵值。

21.如請求項20所述之語音辨識方法，其中該角度估測值 $\hat{\theta}(\omega_f)=\arg\max_{\theta}S(\theta,\omega_f)$ 。

22.如請求項 20 所述之語音辨識方法，其中在利用該空間頻譜定義該空間特徵參數輸出之之步驟中，係同時利用該空間頻譜與該角度估測值，定義二該空間特徵參數輸出之。

23.如請求項 22 所述之語音辨識方法，其中該空間頻譜定義之該空間特徵參數為角度估測量值 $x_1 = \max_{\theta}[\frac{\sum_{f\in F'} S(\theta,\omega_f)}{D}]$，D 為正規化因子(normalized factor)，F' 為共振峰對應之頻帶(formant frequency bands)；以及該角度估測值定義之該空間特徵參數為角度估測變異數 $x_2 = \mathrm{var}(\hat{\theta}(\omega_f))$，且 $f \in F'$。

24.如請求項 19 所述之語音辨識方法，其中在依據該角度估測值與該些聲頻訊號進行擷取與評估，以輸出該巴塔恰里雅距離之步驟中，更包括下列步驟：

依據該角度估測值與該些聲頻訊號擷取一語音特徵資料；以及

接收該語音特徵資料，並從複數種語音共鳴模型資料中取出與該語音特徵資料對應之該語音共鳴模型資料，以對該語音特徵資料進行評估後，輸出該巴塔恰里雅距離。

25.如請求項 24 所述之語音辨識方法，其中該語音共鳴模型資料分別為語音共鳴模型曲線峰值或語音共鳴模型曲線時，該語音特徵資料分別為語音特徵數值或語音特徵曲線。

26.如請求項 25 所述之語音辨識方法，其中該巴塔恰里雅距離 $BC(p,q) = \int\sqrt{p(x)q(x)}dx$，$p(x)$ 為該語音特徵曲線，$q(x)$ 為該語音共鳴模型曲線。

27.如請求項 19 所述之語音辨識方法，其中該關鍵字具有複數音框，在接收

該聲源訊號之步驟中，係連續接收每一該音框之該聲源訊號，以供後續所有步驟運作之。

28.如請求項27所述之語音辨識方法，其中利用該檢測門檻值判斷該正確性之步驟更包括下列步驟：

利用該檢測門檻值對每一該音框對應之該空間特徵參數與該巴塔恰里雅距離進行判斷，以輸出分別代表正確與錯誤之該音框的一第一層正確旗標或一第一層錯誤旗標；

接收每一該音框之該第一層正確旗標或該第一層錯誤旗標，以儲存之，直到該關鍵字對應之所有該第一層正確旗標及該第一層錯誤旗標儲存完後，進行下一步驟；以及

根據該所有該第一層正確旗標及該第一層錯誤旗標之數量，輸出分別代表正確與錯誤之該關鍵字的一第二層正確旗標或一第二層錯誤旗標。

29.如請求項28所述之語音辨識方法，其中在利用該檢測門檻值對每一該音框對應之該空間特徵參數與該巴塔恰里雅距離進行判斷，以輸出該第一層正確旗標或該第一層錯誤旗標之步驟更包括下列步驟：

在該空間特徵參數與該巴塔恰里雅距離中，判斷是否至少其中一者小於或等於對應之該檢測門檻值；

若是，輸出該第一層錯誤旗標；以及

若否，輸出該第一層正確旗標。

30.如請求項28所述之語音辨識方法，其中在根據該所有該第一層正確旗標及該第一層錯誤旗標之該數量，輸出該第二層正確旗標或該第二層錯誤旗標之步驟更包括下列步驟：

於該所有該第一層正確旗標及該第一層錯誤旗標中，判斷該第一層正確旗標所佔之比例，是否大於一第一層檢測值；

若是，輸出該第二層正確旗標；以及

若否，輸出該第二層錯誤旗標。

31.如請求項28所述之語音辨識方法，其中該關鍵字數量為二以上時，於輸出該第二層正確旗標或該第二層錯誤旗標後，更可進行下列步驟：

接收每一該關鍵字之該第二層正確旗標或該第二層錯誤旗標，以儲存之，並在該關鍵詞對應之所有該第二層正確旗標及該第二層錯誤旗標儲存完後，進行下一步驟；以及

根據該所有該第二層正確旗標及該第二層錯誤旗標之數量，輸出分別代表正確與錯誤之該關鍵詞的一正確辨識訊號或一錯誤辨識訊號。

32.如請求項31所述之語音辨識方法，其中在根據該所有該第二層正確旗標及該第二層錯誤旗標之該數量，輸出該正確辨識訊號或該錯誤辨識訊號之步驟更包括下列步驟：

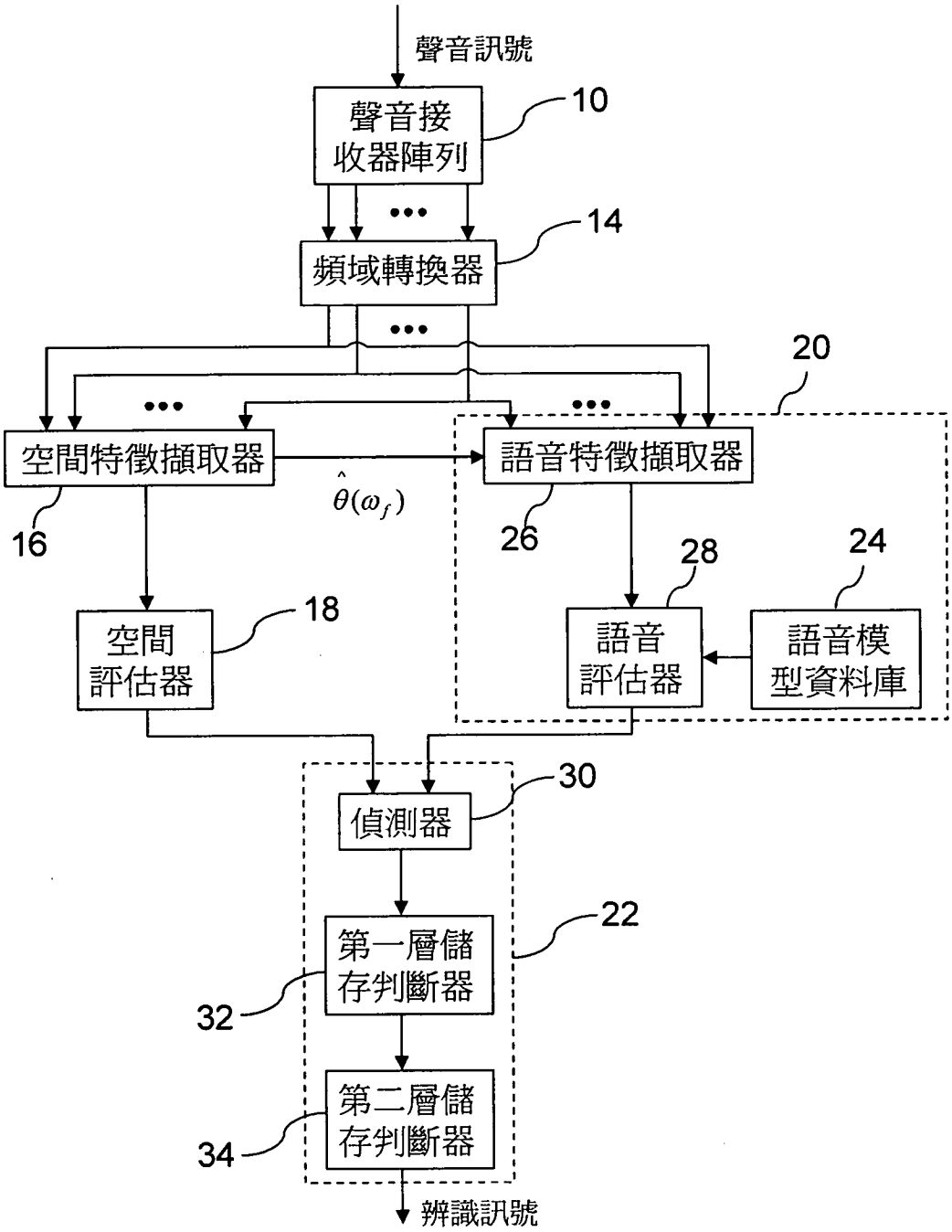
於該所有該第二層正確旗標及該第二層錯誤旗標中，判斷該第二層正確旗標所佔之比例，是否大於一第二層檢測值；

若是，輸出該正確辨識訊號；以及

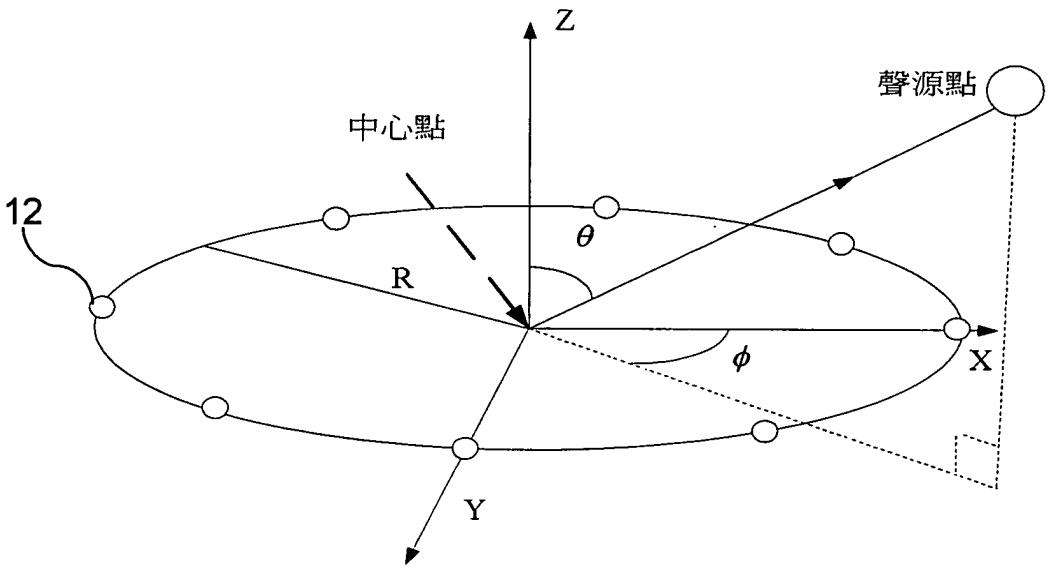
若否，輸出該錯誤辨識訊號。

33.如請求項19所述之語音辨識方法，其中在依據該角度估測值與該些聲頻訊號進行擷取之步驟中，係採用線性預估編碼(Linear Predictive Coding, LPC)法或梅爾倒頻譜係數(Mel-scale Frequency Cepstral Coefficients, MFCC)法，進行之。

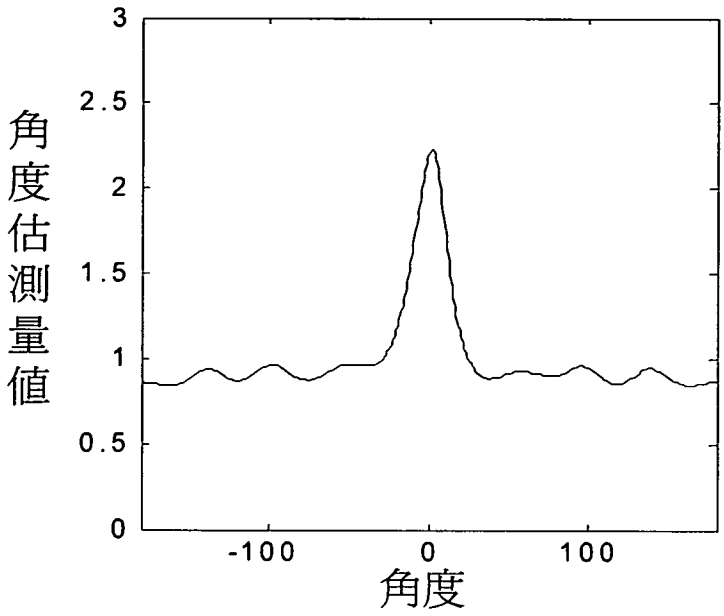
八、圖式：


聲音訊號
聲音接收器陣列
10
頻域轉換器
14
20
空間特徵擷取器
語音特徵擷取器
$\hat{\theta}(\omega_f)$
16
26
28
24
空間評估器
18
語音評估器
語音模型資料庫
偵測器
30
第一層儲存判斷器
22
32
第二層儲存判斷器
34
辨識訊號

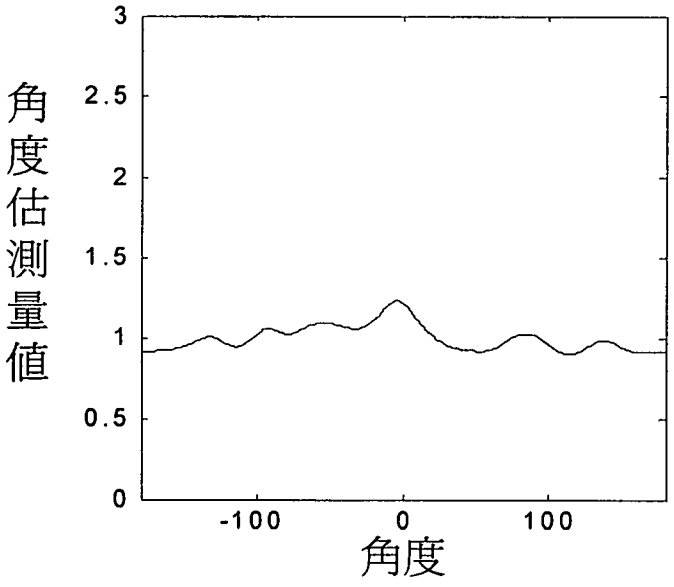
第 1 圖



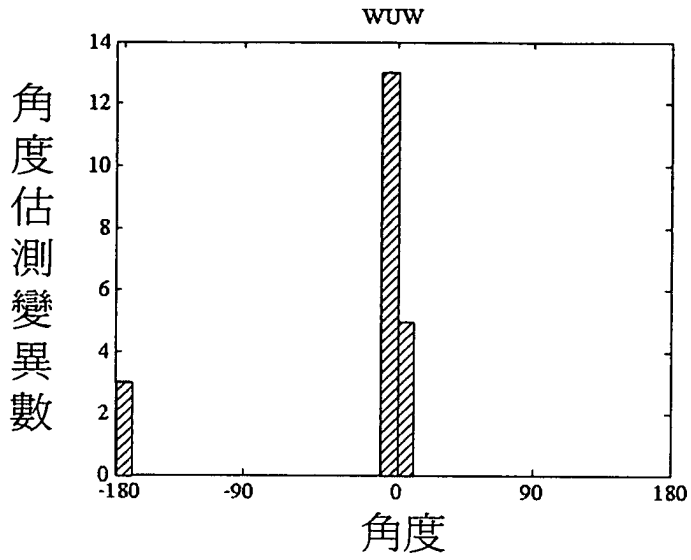
第 2 圖


3
2.5
2
1.5
1
0.5
0
-100
0
100
角度
角度估測量值

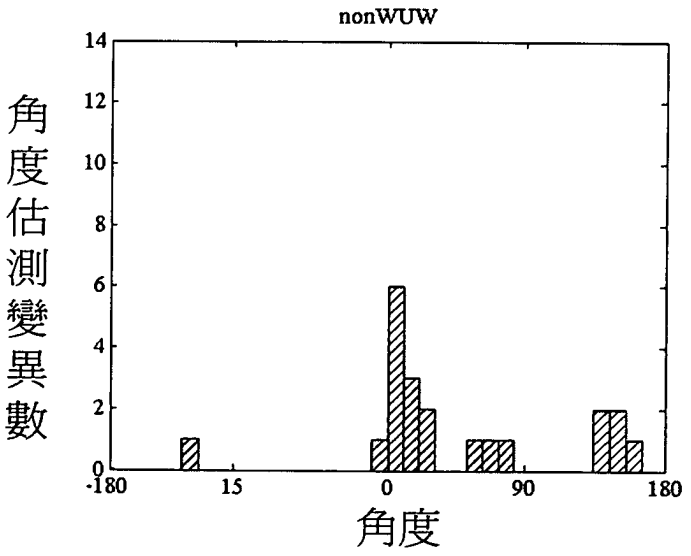
第 3(a) 圖


3
2.5
2
1.5
1
0.5
0
-100
0
100
角度
角度估測量值

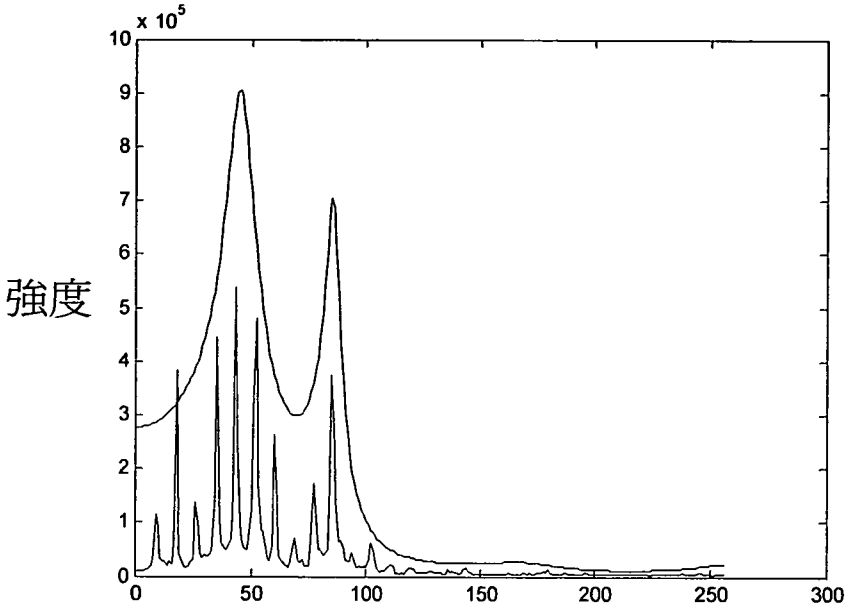
第 3(b) 圖


WUW
角度估測變異數
14
12
10
8
6
4
2
0
-180
-90
0
90
180
角度

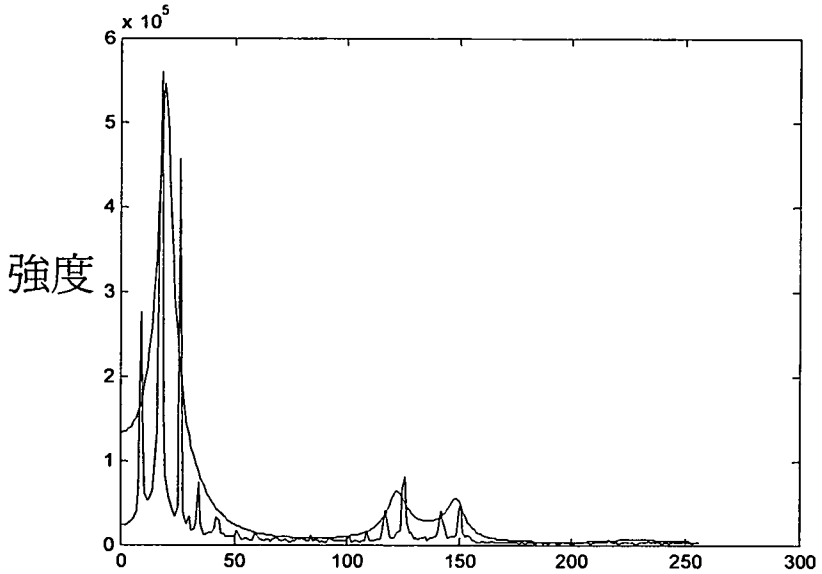
第 4(a) 圖


nonWUW
角度估測變異數
14
12
10
8
6
4
2
0
-180
15
0
90
180
角度

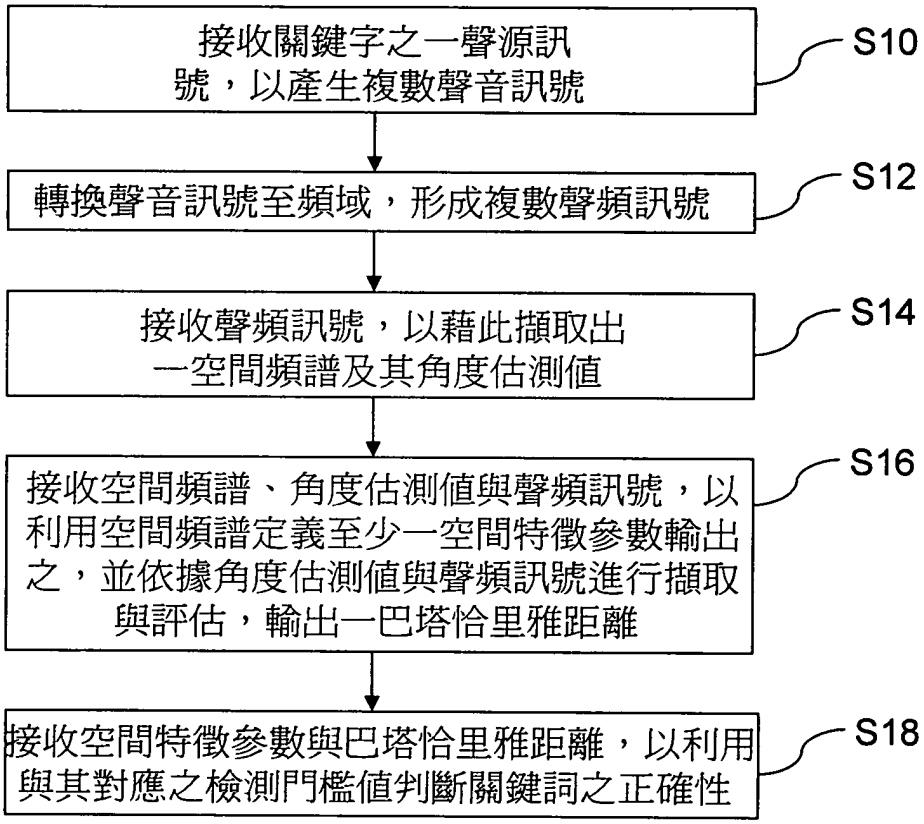
第 4(b) 圖


x 10^5
10
9
8
7
6
5
4
3
2
1
0
強度
0
50
100
150
200
250
300
頻率

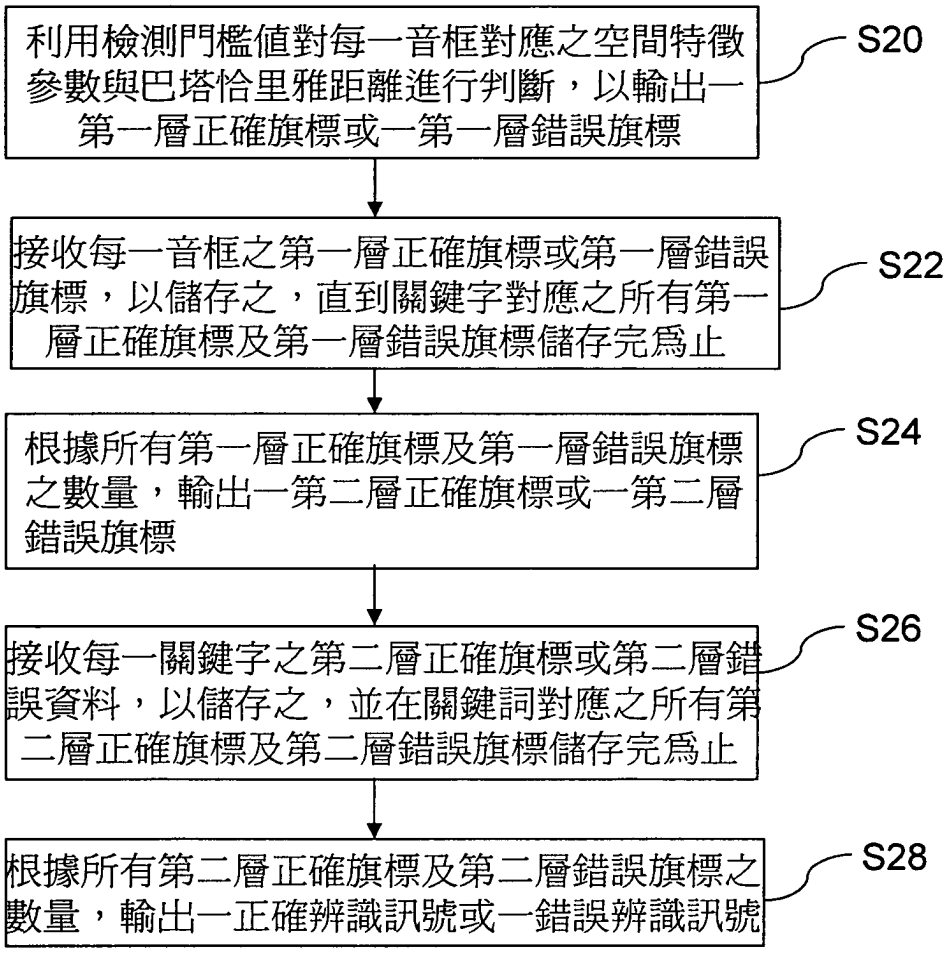
第 5(a) 圖


x 10^5
6
5
4
3
2
1
0
強度
0
50
100
150
200
250
300
頻率

第 5(b) 圖


接收關鍵字之一聲源訊號，以產生複數聲音訊號
S10
轉換聲音訊號至頻域，形成複數聲頻訊號
S12
接收聲頻訊號，以藉此擷取出一空間頻譜及其角度估測值
S14
接收空間頻譜、角度估測值與聲頻訊號，以利用空間頻譜定義至少一空間特徵參數輸出之，並依據角度估測值與聲頻訊號進行擷取與評估，輸出一巴塔恰里雅距離
S16
接收空間特徵參數與巴塔恰里雅距離，以利用與其對應之檢測門檻值判斷關鍵詞之正確性
S18

第 6 圖


利用檢測門檻值對每一音框對應之空間特徵參數與巴塔恰里雅距離進行判斷，以輸出一第一層正確旗標或一第一層錯誤旗標
S20
接收每一音框之第一層正確旗標或第一層錯誤旗標，以儲存之，直到關鍵字對應之所有第一層正確旗標及第一層錯誤旗標儲存完為止
S22
根據所有第一層正確旗標及第一層錯誤旗標之數量，輸出一第二層正確旗標或一第二層錯誤旗標
S24
接收每一關鍵字之第二層正確旗標或第二層錯誤資料，以儲存之，並在關鍵詞對應之所有第二層正確旗標及第二層錯誤旗標儲存完為止
S26
根據所有第二層正確旗標及第二層錯誤旗標之數量，輸出一正確辨識訊號或一錯誤辨識訊號
S28

第 7 圖