I398853

(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 I398853B1**

(45)公告日： 中華民國 102 (2013) 年 06 月 11 日

(21)申請案號：099114811 (22)申請日： 中華民國 99 (2010) 年 05 月 10 日

(51)Int. Cl.： *G10L15/25 (2013.01)*

(71)申請人：國立成功大學 (中華民國) NATIONAL CHENG-KUNG UNIVERSITY (TW)
臺南市東區大學路 1 號

(72)發明人：王駿發 WANG, JHING FA (TW)；施伯宜 SHIH, PO YI (TW)；陳宗佑 CHEN, ZONG YOU (TW)

(74)代理人：詹銘文；蕭錫清

(56)參考文獻：

JP 2008-126329A US 5586215

US 2004/0243413A1

Tze Fen Li and Shui-Ching Chang, "Classification on defective items using unidentified samples," Pattern Recognition, vol. 38, pp. 51-58, 2005.

審查人員：涂淑惠

申請專利範圍項數：18 項 圖式數：8 共 0 頁

(54)名稱

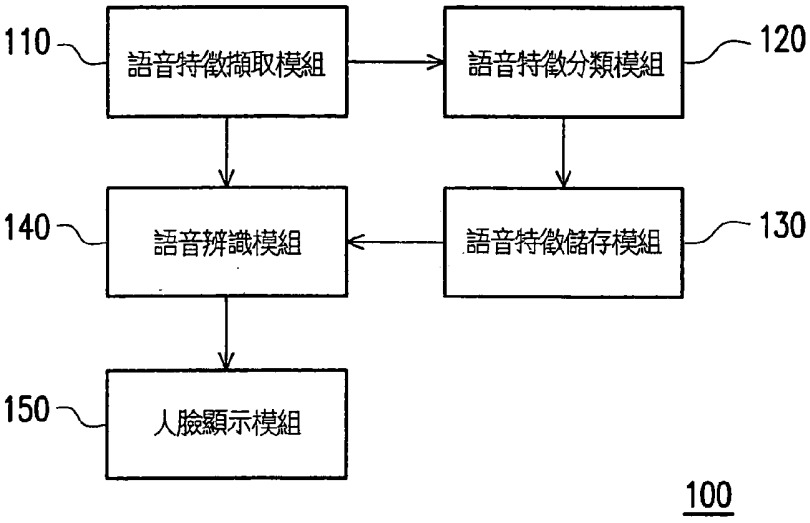
人臉說話模擬系統及方法

SYSTEM AND METHOD FOR SIMULATING HUMAN SPEAKING

(57)摘要

一種人臉說話模擬系統及方法，此方法係擷取樣本語音訊號中的多個語音特徵，並轉換為對應的特徵向量。接著，將這些特徵向量分類為多個語音類別，然後將兩兩語音類別的特徵向量導入一個支援向量機，以求取可區分兩兩語音類別之特徵向量的最佳分割超平面。據此，當接收到使用者輸入的語音訊號時，即可擷取此語音訊號中的語音特徵，並與所求取之最佳分割超平面比對，以判定所屬的語音類別。最後，依據各個語音特徵所屬的語音類別，依序在人臉影像上顯示對應的嘴形圖片，以模擬人臉說話。

A system and a method for simulating human speaking are provided. In the present method, a plurality of voice features of a sample voice signal are captured and transformed into corresponding feature vectors. These feature vectors are then classified into a plurality of voice types and the feature vectors of each two voice types are input into a support vector machine to obtain an optimal separating hyperplane for separating the feature vectors of the two voice types. Accordingly, when receiving a voice single input by a user, the voice features of the voice signal are captured and compared with the previously obtained optimal separating hyperplanes, so as to determine the voice types of the voice features. Finally, according to the voice types of the voice features, a plurality of mouth pictures corresponding to the voice types are sequentially displayed on a human face image, so as to simulate human speaking.



100・・・人臉說話模擬系統

110・・・語音特徵擷取模組

120・・・語音特徵分類模組

130・・・語音特徵儲存模組

140・・・語音辨識模組

150・・・人臉顯示模組

圖 1

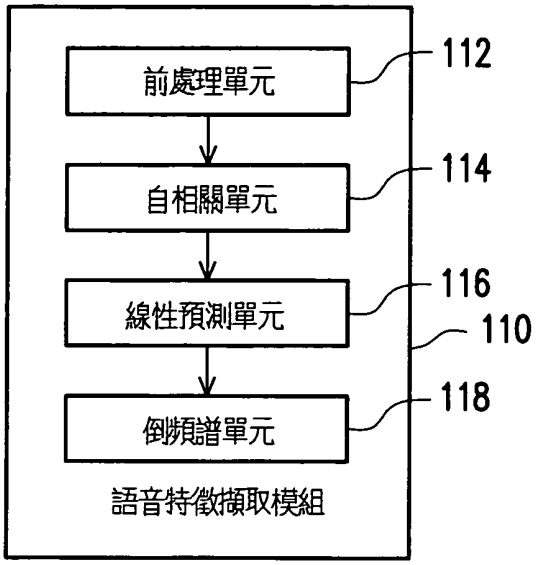

前處理單元
112
自相關單元
114
線性預測單元
116
110
倒頻譜單元
118
語音特徵擷取模組

圖 2

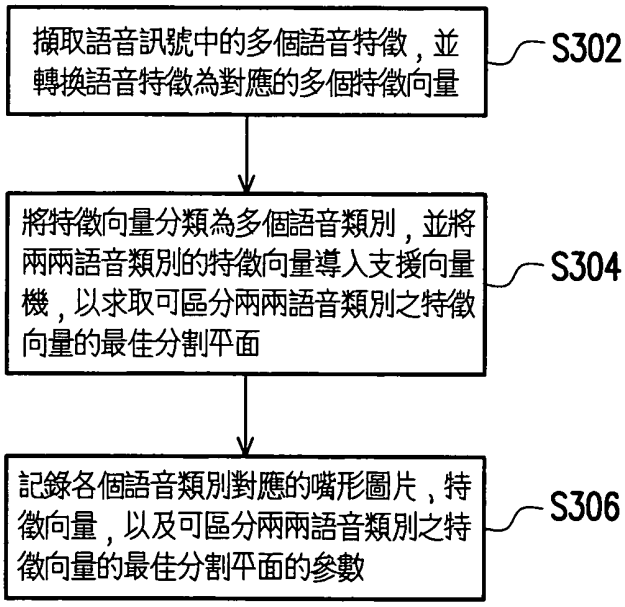

擷取語音訊號中的多個語音特徵，並轉換語音特徵為對應的多個特徵向量
S302
將特徵向量分類為多個語音類別，並將兩兩語音類別的特徵向量導入支援向量機，以求取可區分兩兩語音類別之特徵向量的最佳分割平面
S304
記錄各個語音類別對應的嘴形圖片，特徵向量，以及可區分兩兩語音類別之特徵向量的最佳分割平面的參數
S306

圖 3

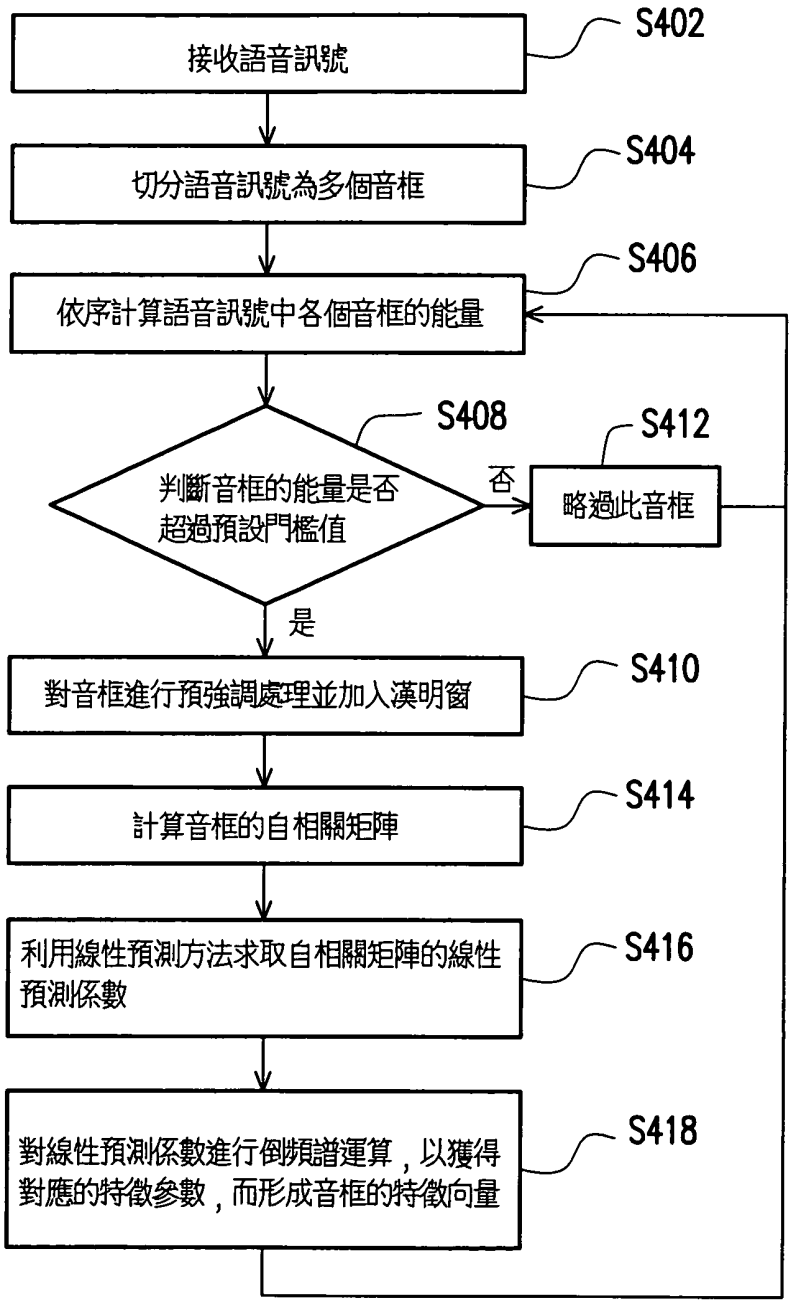
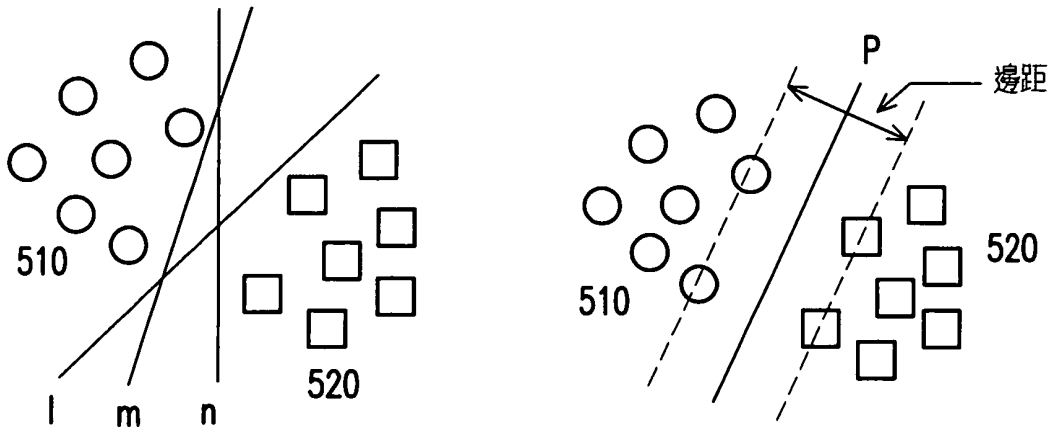

接收語音訊號
S402
切分語音訊號為多個音框
S404
依序計算語音訊號中各個音框的能量
S406
判斷音框的能量是否超過預設門檻值
S408
否
略過此音框
S412
是
對音框進行預強調處理並加入漢明窗
S410
計算音框的自相關矩陣
S414
利用線性預測方法求取自相關矩陣的線性預測係數
S416
對線性預測係數進行倒頻譜運算，以獲得對應的特徵參數，而形成音框的特徵向量
S418

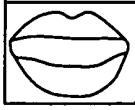
圖 4



(a) (b)

圖 5

34207TW_J

	ㄚ	ㄛ	ㄜ	ㄝ	ㄞ	ㄟ	ㄠ	ㄡ
ㄚ	0	15.0741	19.0442	13.4894	9.1595	10.7216	12.8073	15.9302
ㄛ	15.0741	0	9.8609	12.0444	13.7439	13.6265	5.1153	4.8561
ㄜ	19.0442	9.8609	0	13.7486	17.0691	16.3756	11.1739	7.8477
ㄝ	13.4894	12.0444	13.7486	0	7.7906	6.1886	9.9078	12.4639
ㄞ	9.1595	13.7439	17.0691	7.7906	0	3.9085	11.6632	14.5923
ㄟ	10.7216	13.6265	16.3756	6.1886	3.9085	0	11.3828	14.2116
ㄠ	12.8073	5.1153	11.1739	9.9078	11.6632	11.3828	0	6.5968
ㄡ	15.9302	4.8561	7.8477	12.4639	14.5923	14.2116	6.5968	0
ㄢ	7.2408	13.5677	17.1718	9.592	4.983	6.7236	11.3126	14.3785
ㄣ	17.6195	10.7784	5.854	12.2202	15.3183	14.5832	11.407	8.768
ㄤ	4.0735	16.4539	20.2058	14.1981	9.5736	11.3608	14.386	17.3555
ㄥ	16.8203	7.2798	5.544	13.0477	15.5414	15.1075	8.5858	4.9804
ㄦ	16.1026	7.0348	5.7276	10.7932	13.965	13.3924	7.571	5.5982
ㄧ	19.8457	14.4329	12.2662	11.8169	15.8925	14.7167	14.5519	13.3641
ㄨ	24.6084	18.4423	12.9097	21.1956	23.0373	22.6559	19.8366	16.97
ㄩ	25.0929	18.3472	12.9362	21.372	23.3927	22.9999	19.8164	16.8175
								

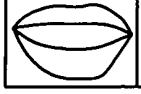
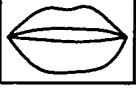
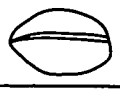
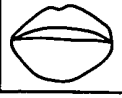
	ㄢ	ㄣ	ㄤ	ㄥ	ㄦ	ㄧ	ㄨ	ㄩ
ㄚ	7.2408	17.6195	4.0735	16.8203	16.1026	19.8457	24.6084	25.0929
ㄛ	13.5677	10.7784	16.4539	7.2798	7.0348	14.4329	18.4423	18.3472
ㄜ	17.1718	5.854	20.2058	5.544	5.7276	12.2662	12.9097	12.9362
ㄝ	9.592	12.2202	14.1981	13.0477	10.7932	11.8169	21.1956	21.372
ㄞ	4.983	15.3183	9.5736	15.5414	13.965	15.8925	23.0373	23.3927
ㄟ	6.7236	14.5832	11.3608	15.1075	13.3924	14.7167	22.6559	22.9999
ㄠ	11.3126	11.407	14.386	8.5858	7.571	14.5519	19.8366	19.8164
ㄡ	14.3785	8.768	17.3555	4.9804	5.5982	13.3641	16.97	16.8175
ㄢ	0	15.406	8.0092	15.2623	13.877	16.9292	23.2871	23.601
ㄣ	15.406	0	18.6361	4.4678	6.4506	9.6816	14.9313	15.113
ㄤ	8.0092	18.6361	0	18.1141	17.2721	20.3067	25.3811	25.9601
ㄥ	15.2623	4.4678	18.1141	0	5.406	12.9031	14.9042	14.8756
ㄦ	13.877	6.4506	17.2721	5.406	0	11.1761	15.9566	16.0363
ㄧ	16.9292	9.6816	20.3067	12.9031	11.1761	0	19.9616	20.2808
ㄨ	23.2871	14.9313	25.3811	14.9042	15.9566	19.9616	0	4.8243
ㄩ	23.601	15.113	25.9601	14.8756	16.0363	20.2808	4.8243	0
								

圖 6

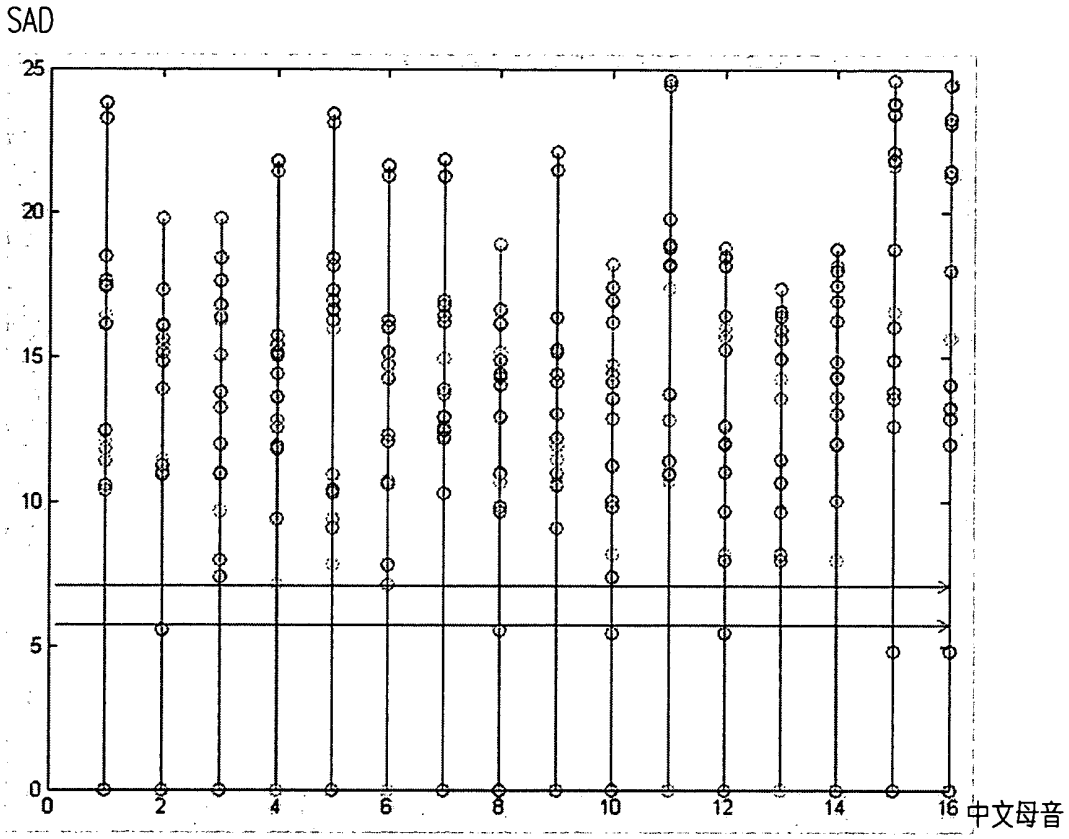


圖 7

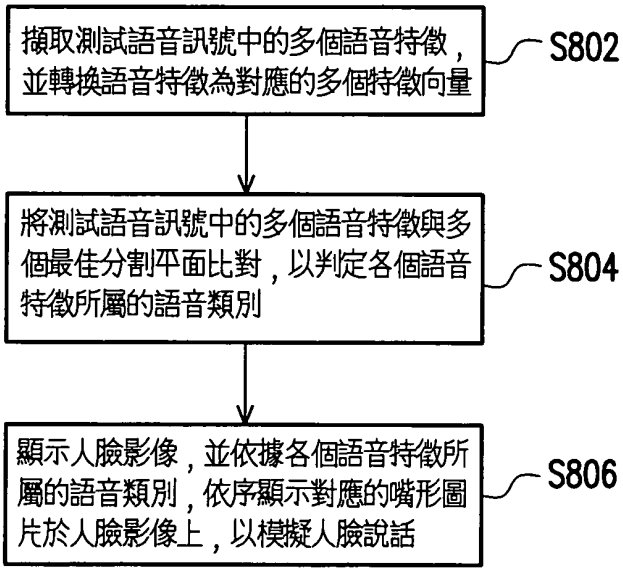

擷取測試語音訊號中的多個語音特徵，並轉換語音特徵為對應的多個特徵向量
S802
將測試語音訊號中的多個語音特徵與多個最佳分割平面比對，以判定各個語音特徵所屬的語音類別
S804
顯示人臉影像，並依據各個語音特徵所屬的語音類別，依序顯示對應的嘴形圖片於人臉影像上，以模擬人臉說話
S806

圖 8

102年 3月 6日修正替換頁(本)

102-3-6

六、發明說明：

【發明所屬之技術領域】

本發明是有關於一種視訊通話系統及方法，且特別是有關於一種人臉說話模擬系統及方法。

【先前技術】

近年來，隨著通訊技術的快速發展，通訊網路已遍及生活周遭，而通訊網路頻寬的增加則使得通訊裝置的功能由基本的語音通話、收發簡訊、電子郵件、瀏覽網頁，擴展到可同時傳輸語音及影像資料的視訊通話。

最近新發展出來的第三代（3G）行動通訊協定即支援視訊通話的功能，其提供了語音資料和非語音資料的進階服務，使用者只需透過支援此第三代行動通訊協定的通訊裝置撥打視訊電話，即可在進行語音通話的同時，透過配置在通訊裝置上的鏡頭擷取自身影像並傳送給對方，而實現視訊通話。

然而，由於視訊通話的資料傳輸量相當大，也需佔用較多的網路頻寬，在網路頻寬有限的情況下，視訊影像的解析度及傳輸速度將會受到影響，結果往往造成視訊影像不夠清晰、產生延遲或出現馬賽克的情況，進而影響視訊通話的品質。

因此，如何能夠在網路頻寬有限的情況下，提供高解析度的視訊影像，並解決影像延遲的問題，已然成為本領域技術的一大課題。

【發明內容】

本發明提供一種人臉說話模擬系統，以對應於語者說話的嘴形圖片取代真人影像，可解決視訊通話中影像延遲的問題。

本發明提供一種人臉說話模擬方法，藉由分辨語音訊號中的語音特徵，並據以顯示對應的嘴形圖片，可模擬真人說話。

本發明提出一種人臉說話模擬系統，其包括語音特徵擷取模組、語音特徵分類模組、語音特徵儲存模組、語音辨識模組及人臉顯示模組。其中，語音特徵擷取模組係用以擷取樣本語音訊號中的多個語音特徵，並將各個語音特徵轉換為對應的特徵向量；語音特徵分類模組係用以將語音特徵對應的特徵向量分類為多個語音類別，並將兩兩語音類別的特徵向量導入支援向量機（Support vector machine，SVM），以求取可區分兩兩語音類別之特徵向量的最佳分割超平面；語音特徵儲存模組係用以記錄各個語音類別對應的嘴形圖片、特徵向量，以及可區分兩兩語音類別之特徵向量的最佳分割超平面；語音辨識模組係用以將輸入語音訊號中各個語音特徵對應的特徵向量與最佳分割超平面比對，以判定此些特徵向量所屬的語音類別，其中所述的特徵向量係透過語音特徵擷取模組擷取及轉換；人臉顯示模組係用以顯示一人臉影像，並依據各個語音特徵所屬的語音類別，依序顯示對應的嘴形圖片於此人臉影像上，以模擬人臉說話。

在本發明之一實施例中，上述之語音特徵擷取模組包括前處理單元、自相關單元、線性預測單元及倒頻譜單元。其中，前處理單元係用以將語音訊號切分為多個音框，以對各個音框進行預強調處理並加入漢明窗；自相關單元係用以對前處理單元處理後的音框進行自相關運算，以取得這些音框的自相關矩陣；線性預測單元係利用線性預測方法求取自相關矩陣的多個線性預測係數（Linear Predictive Coefficient，LPC）；倒頻譜單元係用以對上述的線性預測係數進行倒頻譜運算，以獲得對應的多個特徵參數，而這些特徵參數即形成所述的特徵向量。上述的線性預測方法例如是 Levinson-Durbin 遞回演算法。

在本發明之一實施例中，上述之前處理單元更包括判斷所切分之音框中每一個音框的能量是否超過一個預設門檻值，其中若音框的能量超過預設門檻值，即對此音框進行預強調處理及加入漢明窗，並記錄此音框以供自相關單元進行自相關運算。

在本發明之一實施例中，上述之人臉說話模擬系統更包括圖片擷取模組及圖片分類模組。其中，圖片擷取模組係用以擷取各個語音分類所對應的多張嘴形圖片；圖片分類模組則用以計算這些語音分類中兩兩語音分類所對應之嘴形圖片的差異，據以對這些嘴形圖片進行分類。所述的差異例如是兩兩語音分類所對應之嘴形圖片中對應像素之像素值的絕對差值總和（Sum of Absolute Differences，SAD）。

在本發明之一實施例中，上述之嘴形圖片分類模組包括判斷兩兩語音分類所對應之嘴形圖片的差異是否低於一個門檻值，其中若此差異低於門檻值，則判斷這兩個語音分類的嘴形圖片相似，而使用同一張嘴形圖片做為這兩個語音分類的嘴形圖片。

在本發明之一實施例中，上述之語音辨識模組包括依照特徵向量位於各個最佳分割超平面兩邊的比例，判定這些特徵向量所屬的語音類別。

在本發明之一實施例中，上述之人臉顯示模組更包括計算所要顯示之相鄰語音特徵的特徵向量所佔之權重，並用以加乘相鄰語音特徵對應的嘴形圖片，以顯示混合嘴形圖片。

本發明提出一種人臉說話模擬方法，其包括訓練步驟及模擬步驟。其中，訓練步驟包括接收樣本語音訊號，並擷取此樣本語音訊號中的多個語音特徵，而將這些語音特徵轉換為對應的特徵向量。接著，將這些語音特徵對應的特徵向量分類為多個語音類別，然後將兩兩語音類別的特徵向量導入一個支援向量機，以求取可區分兩兩語音類別之特徵向量的最佳分割超平面。最後，記錄各個語音類別對應的嘴形圖片、特徵向量，以及可區分兩兩語音類別之特徵向量的最佳分割超平面的多個參數。另一方面，模擬步驟包括接收輸入語音訊號，並擷取此輸入語音訊號中的語音特徵，而將這些語音特徵轉換為對應的特徵向量。接著，將這些特徵向量與所記錄之最佳分割超平面比對，以

判定這些特徵向量所屬的語音類別。最後，顯示一張人臉影像，並依據各個語音特徵所屬的語音類別，依序在此人臉影像上顯示對應的嘴形圖片，以模擬人臉說話。

在本發明之一實施例中，上述擷取樣本語音訊號中的語音特徵，並將語音特徵轉換為對應之特徵向量的步驟包括將此語音訊號為多個音框，以對各個音框進行預強調處理並加入漢明窗，接著對這些音框進行自相關運算，以取得這些音框的自相關矩陣，然後利用線性預測方法求取此自相關矩陣的多個線性預測係數，最後則對這些線性預測係數進行倒頻譜運算，以獲得對應的多個特徵參數，而這些特徵參數即形成特徵向量。上述的線性預測方法例如是Levinson-Durbin 遞回演算法。

在本發明之一實施例中，上述的訓練步驟更包括判斷所切分之音框中每一個音框的能量是否超過一個預設門檻值，其中若音框的能量超過預設門檻值，即對此音框進行預強調處理及加入漢明窗，並記錄此音框以進行自相關運算。

在本發明之一實施例中，上述的訓練步驟更包括擷取各個語音分類所對應的多張嘴形圖片，並計算這些語音分類中兩兩語音分類所對應之嘴形圖片的差異，據以對這些嘴形圖片進行分類。

在本發明之一實施例中，上述計算兩兩語音分類所對應之嘴形圖片的差異，據以對嘴形圖片進行分類的步驟包括判斷兩兩語音分類所對應之嘴形圖片的差異是否低於一

個門檻值，其中若此差異低於門檻值，則判斷這兩個語音分類的嘴形圖片相似，而使用同一張嘴形圖片做為這兩個語音分類的嘴形圖片。上述的差異例如是兩兩語音分類所對應之嘴形圖片中對應像素之像素值的絕對差值總和。

在本發明之一實施例中，上述將特徵向量與所記錄之最佳分割超平面比對，以判定特徵向量所屬的語音類別的步驟包括依照特徵向量位於各個最佳分割超平面兩邊的比例，判定這些特徵向量所屬的語音類別。

在本發明之一實施例中，上述依據各個語音特徵所屬的語音類別，依序在人臉影像上顯示對應的嘴形圖片的步驟包括計算所要顯示之相鄰語音特徵的特徵向量所佔之權重，用以加乘相鄰語音特徵對應的嘴形圖片，而顯示混合嘴形圖片。

基於上述，本發明之人臉說話模擬系統及方法係利用語音訊號中多種語音特徵的特徵向量訓練語音模型，而可用以分辨真人說話中多個語音特徵的類型，以顯示對應的嘴形圖片，可達到模擬真人說話的功效，並解決視訊通話中影像延遲的問題。

為讓本發明之上述特徵和優點能更明顯易懂，下文特舉實施例，並配合所附圖式作詳細說明如下。

【實施方式】

圖 1 是依照本發明一實施例所繪示之人臉說話模擬系統的方塊圖，圖 3 則是依照本發明一實施例所繪示之人臉

說話模擬方法的流程圖。請同時參照圖 1 及圖 3，本實施例的人臉說話模擬方法包括訓練步驟及模擬步驟兩部分，其中訓練步驟係訓練可區分不同語音特徵的語音模型，而模擬步驟則是利用訓練步驟所訓練的語音模型來區分輸入語音訊號中各個語音特徵的語音類別，據以顯示對應的嘴形圖片，而模擬人臉說話。

本實施例之模擬系統 100 包括語音特徵擷取模組 110、語音特徵分類模組 120、語音特徵儲存模組 130、語音辨識模組 140 及人臉顯示模組 150。其中，語音特徵擷取模組 110、語音特徵分類模組 120 及語音特徵儲存模組 130 適用於上述的訓練步驟，而用以訓練語音模型；語音特徵擷取模組 110、語音辨識模組 140 及人臉顯示模組 150 則適用於上述的模擬步驟，而用以模擬人臉說話。以下即搭配上述模擬系統 100 中的各個元件說明本實施例之人臉說話模擬方法的詳細步驟。

在訓練階段，首先提供樣本語音訊號至語音特徵擷取模組 110，而由語音特徵擷取模組 110 擷取此樣本語音訊號中的多個語音特徵，並將這些語音特徵分別轉換為對應的特徵向量（步驟 S302）。其中，所述的樣本語音訊號例如是由使用者所唸出的多個中文母音，而語音特徵擷取模組 110 即擷取語音訊號中對應於這些中文母音的語音特徵。

詳細地說，圖 2 是依照本發明一實施例所繪示之語音特徵擷取模組的方塊圖，圖 4 則是依照本發明一實施例所

繪示之語音特徵擷取方法的流程圖。請參照圖 2，本實施例係將上述的語音特徵擷取模組 110 再細分為前處理單元 112、自相關單元 114、線性預測單元 116 及倒頻譜單元 118。以下即搭配上述語音特徵擷取模組 110 中的各個元件說明本實施例之語音特徵擷取方法的詳細步驟。

每當語音特徵擷取模組 110 接收到語音訊號（步驟 S402）時，即由前處理單元 112 將其切分為多個音框（步驟 S404），並依序計算各個音框的能量（步驟 S406），而判斷這些音框的能量是否超過預設門檻值（步驟 S408）。其中，若音框的能量超過預設門檻值，前處理單元 112 即判定此音框屬於有用的音框，此時前處理單元 112 除了將此音框儲存起來以進行後續的處理外，還會對此音框進行預強調處理及加入漢明窗（步驟 S410）；反之，若音框的能量未超過預設門檻值，前處理單元 112 則會略過此音框（步驟 S412），而繼續處理下個音框（步驟 S406）。

在經由前處理單元 112 的預強調處理及加入漢明窗之後，接著則由自相關單元 114 對處理後的音框進行自相關運算，以取得這些音框的自相關矩陣（步驟 S414）。然後，由線性預測單元 116 利用線性預測方法來求取此自相關矩陣對應的線性預測係數（步驟 S416）。所述的線性預測方法例如是 Levinson-Durbin 遞回演算法，而藉由此演算法的遞回來求解，即可得到一組線性預測係數。最後，由倒頻譜單元 118 對這些線性預測係數進行倒頻譜運算，以獲得對應的多個特徵參數，這些特徵參數即可集合形成特徵向

量，以作為後續分類語音特徵的依據（步驟 S418）。

回到圖 3，在語音特徵擷取模組 110 取得各個語音特徵對應的特徵向量後，即將此資料輸入語音特徵分類模組 120，而由語音特徵分類模組 120 將其分類為多個語音類別。其中，語音特徵分類模組 120 例如是將兩兩語音類別的特徵向量導入支援向量機（support vector machine，SVM），以求取可區分兩兩語音類別之特徵向量的最佳分割超平面（optimal separating hyperplane，OSH）（步驟 S304）。

舉例來說，假設目前有兩組特徵向量，其對應於不同的語音類別，若將每一個特徵向量均視為空間中的一個點，則可繪示出如圖 5(a)所示的特徵向量分佈圖 500。此分佈圖 500 中的圓形座標點 510 及方形座標點 520 即分別代表兩種語音類別的特徵向量，而分割線 l、m、n 則為可區分這兩類資料的分割線。需注意此分割線在高維度空間中不再是以直線的形式存在，而是以超平面（hyperplane）的形式存在，本實施例所繪示的直線僅為舉例說明。本實施例即求取一個可區別兩類資料的超平面（如圖 5(b)所示的超平面 p），使得這個超平面到兩類資料的距離為最短，而此最短距離稱為邊距（margin）。

支援向量機的特性就是可以根據兩類資料的特徵向量，找出一個與兩類資料之距離為最短的超平面作為最佳分割超平面。本實施例在訓練語音模型時，就是將不同語音類別的特徵向量兩兩送入支援向量機，以求取最佳分割

超平面。

在求取最佳分割超平面之後，接著則由語音特徵儲存模組 130 記錄各個語音類別對應的嘴形圖片、多個特徵向量，以及由語音特徵分類模組 120 所求出可區分兩兩語音類別之特徵向量的最佳分割超平面（步驟 S306），而完成語音模型的訓練步驟。

詳細地說，本實施例在訓練階段中，就會將特徵向量分門別類儲存好，以作為後續辨識語者的依據。舉例來說，若中文母音"ㄚ"的音檔有兩筆，其中音檔 1 包括 100 個音框的母音"ㄚ"，音檔 2 包括 150 個音框的母音"ㄚ"，則音檔 1 經過特徵擷取後有 100 個特徵向量，音檔 2 則有 150 個特徵向量。本實施例即將此母音"ㄚ"的 250 個特徵向量用來訓練語音模型。同理，任何一類的語音資料也都會先分門別類轉換成特徵向量並儲存好。

需注意的是，在人們發出中文母音"ㄛ"和"ㄡ"時，由於這兩個母音的音調相似，故在分類上很容易會導致辨識錯誤，例如在模擬多個音框的母音"ㄛ"時，錯將部分音框辨識為母音"ㄡ"， 因此模擬母音"ㄛ"的嘴形圖片中會摻雜母音"ㄡ"的嘴形圖片，結果則導致模擬母音"ㄛ"的嘴形圖片會產生些許的顫動。

為了解決上述問題，本實施例之模擬系統 500 還可額外配置圖片擷取模組及圖片分類模組（未繪示）。藉由圖片擷取模組擷取各個語音分類所對應的嘴形圖片，並由圖片分類模組計算這些語音分類中兩兩語音分類所對應之嘴

形圖片的差異，而對這些嘴形圖片進行分類。詳細地說，嘴形圖片分類模組例如會判斷兩兩語音分類所對應之嘴形圖片的差異是否低於門檻值。其中，若所此差異低於門檻值，則判斷這兩種語音分類的嘴形圖片相似，而使用同一張嘴形圖片來做為這兩種語音分類的嘴形圖片；反之，則使用各自的嘴形圖片。上述的差異例如是兩種語音分類所對應之嘴形圖片中對應像素之像素值的絕對差值總和（Sum of Absolute Differences，SAD）或其他可區分圖片差異的參數值，本實施例不限制其範圍。

舉例來說，圖 6 是依照本發明一實施例所繪示之嘴形圖片分類圖。請參照圖 6，本實施例係針對一個語者唸出 16 個中文母音時的嘴形圖片，計算兩兩中文母音之嘴形圖片中對應像素之像素值的絕對差值總和。而藉由這些絕對差值總和的資料，即可判斷出哪些母音的嘴形圖片相類似。

圖 7 進一步繪示 16 個中文母音與其他中文母音之嘴形圖片的絕對差值總和分佈圖。由經驗值可知，正常用以區分嘴形圖片的門檻值大約分佈在 5~10 之間，因此本實施例即在 5~10 之間找一個最大且其中沒有絕對差值總和分佈的區間，而取此區間的中點作為判斷嘴形圖片是否相似的門檻值。

在完成上述的語音模型訓練之後，則可進行模擬步驟。圖 8 是依照本發明一實施例所繪示之人臉說話模擬方法的流程圖。請同時參照圖 1 及圖 8，本實施例的模擬方法例如是接續在圖 3 所示的模擬方法之後，而利用其所訓

練之語音模型進行人臉說話的模擬，其詳細步驟分述如下：

首先，由使用者將輸入語音訊號輸入語音特徵擷取模組 110，而由語音特徵擷取模組 110 擷取此輸入語音訊號中的多個語音特徵，並將這些語音特徵分別轉換為對應的特徵向量（步驟 S802）。其中，所述的輸入語音訊號例如是由使用者對著語音特徵擷取模組 110 說話而產生，而語音特徵擷取模組 110 即擷取語音訊號中對應於多個中文母音的語音特徵。

接著，語音辨識模組 140 即會將此輸入語音訊號中各個語音特徵對應的特徵向量與語音特徵儲存模組 130 中記錄的最佳分割超平面比對，以判定這些特徵向量所屬的語音類別（步驟 S804）。其中，語音辨識模組 140 例如是依照這些特徵向量位於各個最佳分割超平面兩邊的比例，而判定這些特徵向量所屬的語音類別。

舉例來說，假設目前只有兩類語音資料，其中一類在最佳分割超平面的左邊，作為+1 類；另一類在最佳分割超平面的右邊，作為-1 類。本實施例即由語音特徵儲存模組 130 取出可區分這兩類語音資料的最佳分割超平面，然後將每個音框求取出來的特徵向量都與這個超平面做比對。其中，若特徵向量落在超平面的左邊，則標記為+1；反之，則標記為-1。在完成每個特徵向量的標記後，即可將此語音資料中所有音框的分數加總起來，而用以判斷此語音資料所述的類別。其中，若分數小於零，則可判定此語音資料屬於-1 類；反之，則判定此語音資料屬於+1 類，如此即

可達到分類的效果。

最後，人臉顯示模組150例如是在電子裝置的螢幕上顯示一張人臉影像，並依據語音辨識模組140所辨識之各個語音特徵所屬的語音類別，依序在此人臉影像上顯示對應的嘴形圖片，以模擬人臉說話（步驟S806）。

需注意的是，為了增加嘴形圖片顯示的平順度，本實施例的人臉顯示模組150更包括在顯示嘴形圖片時，計算所要顯示之相鄰語音特徵的特徵向量所佔之權重，並將相鄰兩個語音特徵的權重加乘對應的嘴形圖片，而以混合嘴形圖片的形式作為語音特徵轉換期間的嘴形圖片顯示。

詳細地說，本實施例例如是藉由調整目的圖片（後一張嘴形圖片）之透明度來混合來源圖片（前一張嘴形圖片）與目的圖片之像素。其中，本實施例例如是採用 Alpha Blending 等圖片混合技術，在語音特徵的權重 alpha=0 時，將目的圖片完全透明化，以使顯示出來的圖片為來源圖片；在語音特徵的權重 alpha=0.5 時，將來源圖片與目的圖片之像素混合，使得顯示出來的圖片為來源圖片與目的圖片各占一半的混合圖片，而產生重疊的效果。利用上述方式改變來源圖片與目的圖片的權重(0~1)，即可達到將兩張相異圖片平滑化的需求。

綜上所述，本發明人臉說話模擬系統及方法係對即時輸入的語音訊號辨識其中的語音特徵，並根據預先訓練好的語音模型找出各個語音特徵對應的嘴形圖片以顯示於人臉影像中的嘴部區域，而達到模擬真人說話的功效。本發

明技術只需使用低位元流的語音訊號即可模擬真人說話，而可解決傳統視訊通話中影像延遲的問題。

雖然本發明已以實施例揭露如上，然其並非用以限定本發明，任何所屬技術領域中具有通常知識者，在不脫離本發明之精神和範圍內，當可作些許之更動與潤飾，故本發明之保護範圍當視後附之申請專利範圍所界定者為準。

【圖式簡單說明】

圖 1 是依照本發明一實施例所繪示之人臉說話模擬系統的方塊圖。

圖 2 是依照本發明一實施例所繪示之語音特徵擷取模組的方塊圖。

圖 3 是依照本發明一實施例所繪示之人臉說話模擬方法的流程圖。

圖 4 是依照本發明一實施例所繪示之語音特徵擷取方法的流程圖。

圖 5(a)及圖 5(b)是依照本發明一實施例所繪示之特徵向量分佈圖。

圖 6 是依照本發明一實施例所繪示之嘴形圖片分類圖。

圖 7 是依照本發明一實施例所繪示之嘴形圖片的絕對差值總和分佈圖。

圖 8 是依照本發明一實施例所繪示之人臉說話模擬方法的流程圖。

【主要元件符號說明】

100：人臉說話模擬系統

110：語音特徵擷取模組

112：前處理單元

114：自相關單元

116：線性預測單元

118：倒頻譜單元

120：語音特徵分類模組

130：語音特徵儲存模組

140：語音辨識模組

150：人臉顯示模組

500：特徵向量分佈圖

510：圓形座標點

520：方形座標點

S302~S306：本發明一實施例之人臉說話模擬方法的步驟

S402~S418：本發明一實施例之語音特徵擷取方法的步驟

S802~S806：本發明一實施例之人臉說話模擬方法的步驟

I398853

發明專利說明書

（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：99114811

※申請日：99.5.10　　※IPC分類：G10L 15/25 (2013.01)

一、發明名稱：

人臉說話模擬系統及方法 / SYSTEM AND METHOD FOR SIMULATING HUMAN SPEAKING

二、中文發明摘要：

一種人臉說話模擬系統及方法，此方法係擷取樣本語音訊號中的多個語音特徵，並轉換為對應的特徵向量。接著，將這些特徵向量分類為多個語音類別，然後將兩兩語音類別的特徵向量導入一個支援向量機，以求取可區分兩兩語音類別之特徵向量的最佳分割超平面。據此，當接收到使用者輸入的語音訊號時，即可擷取此語音訊號中的語音特徵，並與所求取之最佳分割超平面比對，以判定所屬的語音類別。最後，依據各個語音特徵所屬的語音類別，依序在人臉影像上顯示對應的嘴形圖片，以模擬人臉說話。

三、英文發明摘要：

A system and a method for simulating human speaking are provided. In the present method, a plurality of voice features of a sample voice signal are captured and transformed into corresponding feature vectors. These feature vectors are then classified into a plurality of voice

types and the feature vectors of each two voice types are input into a support vector machine to obtain an optimal separating hyperplane for separating the feature vectors of the two voice types. Accordingly, when receiving a voice single input by a user, the voice features of the voice signal are captured and compared with the previously obtained optimal separating hyperplanes, so as to determine the voice types of the voice features. Finally, according to the voice types of the voice features, a plurality of mouth pictures corresponding to the voice types are sequentially displayed on a human face image, so as to simulate human speaking.

四、指定代表圖：

（一）本案之指定代表圖：圖 1

（二）本代表圖之元件符號簡單說明：

100：人臉說話模擬系統

110：語音特徵擷取模組

120：語音特徵分類模組

130：語音特徵儲存模組

140：語音辨識模組

150：人臉顯示模組

五、本案若有化學式時，請揭示最能顯示發明特徵的化學式：無

七、申請專利範圍：

1.一種人臉說話模擬系統，包括：

一語音特徵擷取模組，擷取一樣本語音訊號中的多個語音特徵，並轉換各該些語音特徵為對應的一特徵向量；

一語音特徵分類模組，將該些語音特徵對應的該些特徵向量分類為多個語音類別，並將兩兩語音類別的特徵向量導入一支援向量機，以求取可區分兩兩語音類別之特徵向量的一最佳分割超平面；

一語音特徵儲存模組，記錄各該些語音類別對應的一嘴形圖片、該些特徵向量，以及可區分兩兩語音類別之特徵向量的該最佳分割超平面；

一語音辨識模組，將一輸入語音訊號中各該些語音特徵對應的該些特徵向量與該些最佳分割超平面比對，以判定該些特徵向量所屬的一語音類別，其中該特徵向量係透過該語音特徵擷取模組擷取及轉換；以及

一人臉顯示模組，顯示一人臉影像，並依據各該些語音特徵所屬的該語音類別，依序顯示對應的一嘴形圖片於該人臉影像上，以模擬人臉說話。

2.如申請專利範圍第 1 項所述之人臉說話模擬系統，其中該語音特徵擷取模組包括：

一前處理單元，切分該語音訊號為多個音框，以對各該些音框進行一預強調處理並加入一漢明窗；

一自相關單元，對該前處理單元處理後的該些音框進行一自相關運算，以取得該些音框的一自相關矩陣；

一線性預測單元，利用一線性預測方法求取該自相關矩陣的多個線性預測係數（Linear Predictive Coefficient，LPC）；以及

一倒頻譜單元，對該些線性預測係數進行一倒頻譜運算，以獲得對應的多個特徵參數，其中該些特徵參數形成該些特徵向量。

3.如申請專利範圍第 2 項所述之人臉說話模擬系統，其中該線性預測方法為 Levinson-Durbin 遞回演算法。

4.如申請專利範圍第 2 項所述之人臉說話模擬系統，其中該前處理單元更包括判斷所切分之該些音框中每一個音框的一能量是否超過一預設門檻值，其中

若該音框的該能量超過該預設門檻值，對該音框進行該預強調處理及加入該漢明窗，並記錄該音框以供該自相關單元進行該自相關運算。

5.如申請專利範圍第 1 項所述之人臉說話模擬系統，更包括：

一圖片擷取模組，擷取各該些語音分類所對應的多張嘴形圖片；以及

一圖片分類模組，計算該些語音分類中兩兩語音分類所對應之嘴形圖片的一差異，據以分類該些嘴形圖片。

6.如申請專利範圍第 5 項所述之人臉說話模擬系統，其中該嘴形圖片分類模組包括判斷兩兩語音分類所對應之嘴形圖片的該差異是否低於一門檻值，其中

若該差異低於該門檻值，則判斷該兩個語音分類的嘴

形圖片相似，而使用同一嘴形圖片做為該兩個語音分類的嘴形圖片。

7.如申請專利範圍第 5 項所述之人臉說話模擬系統，其中該圖片分類模組所計算的該差異為兩兩語音分類所對應之嘴形圖片中對應像素之像素值的絕對差值總和（Sum of Absolute Differences，SAD）。

8.如申請專利範圍第 1 項所述之人臉說話模擬系統，其中該語音辨識模組包括依照該些特徵向量位於各該些最佳分割超平面兩邊的一比例，判定該些特徵向量所屬的該語音類別。

9.如申請專利範圍第 1 項所述之人臉說話模擬系統，其中該人臉顯示模組更包括計算所要顯示之相鄰語音特徵的該些特徵向量所佔之一權重，加乘該相鄰語音特徵對應的嘴形圖片，以顯示一混合嘴形圖片。

10.一種人臉說話模擬方法，包括下列步驟：

一訓練步驟，包括：

接收一樣本語音訊號；

擷取該樣本語音訊號中的多個語音特徵，並轉換各該些語音特徵為對應的一特徵向量；

分類該些語音特徵對應的該些特徵向量為多個語音類別；

將兩兩語音類別的特徵向量導入一支援向量機，以求取可區分兩兩語音類別之特徵向量的一最佳分割超平面；以及

記錄各該些語音類別對應的一嘴形圖片、該些特徵向量，以及可區分兩兩語音類別之特徵向量的該最佳分割超平面的多個參數；以及

一模擬步驟，包括：

接收一輸入語音訊號；

擷取該輸入語音訊號中的語音特徵，並轉換各該些語音特徵為對應的特徵向量；

將該些特徵向量與所記錄之該些最佳分割超平面比對，以判定該些特徵向量所屬的一語音類別；以及

顯示一人臉影像，並依據各該些語音特徵所屬的該語音類別，依序顯示對應的一嘴形圖片於該人臉影像上，以模擬人臉說話。

11.如申請專利範圍第 10 項所述之人臉說話模擬方法，其中擷取該樣本語音訊號中的該些語音特徵，並轉換各該些語音特徵為對應的該特徵向量的步驟包括：

切分該語音訊號為多個音框，以對各該些音框進行一預強調處理並加入一漢明窗；

對該些音框進行一自相關運算，以取得該些音框的一自相關矩陣；

利用一線性預測方法求取該自相關矩陣的多個線性預測係數；以及

對該些線性預測係數進行一倒頻譜運算，以獲得對應的多個特徵參數，其中該些特徵參數形成該些特徵向量。

12.如申請專利範圍第 11 項所述之人臉說話模擬方

法，其中該線性預測方法為 Levinson-Durbin 遞回演算法。

13.如申請專利範圍第 11 項所述之人臉說話模擬方法，其中該訓練步驟更包括：

判斷所切分之該些音框中每一個音框的一能量是否超過一預設門檻值；以及

若該音框的該能量超過該預設門檻值，對該音框進行該預強調處理及加入該漢明窗，並記錄該音框以進行該自相關運算。

14.如申請專利範圍第 10 項所述之人臉說話模擬方法，其中該訓練步驟更包括：

擷取各該些語音分類所對應的多張嘴形圖片；以及

計算該些語音分類中兩兩語音分類所對應之嘴形圖片的一差異，據以分類該些嘴形圖片。

15.如申請專利範圍第 14 項所述之人臉說話模擬方法，其中計算該些語音分類中兩兩語音分類所對應之嘴形圖片的該差異，據以分類該些嘴形圖片的步驟包括：

判斷兩兩語音分類所對應之嘴形圖片的該差異是否低於一門檻值；以及

若該差異低於該門檻值，則判斷該兩個語音分類的嘴形圖片相似，而使用同一嘴形圖片做為該兩個語音分類的嘴形圖片。

16.如申請專利範圍第 14 項所述之人臉說話模擬方法，其中該差異為兩兩語音分類所對應之該些嘴形圖片中對應像素之像素值的絕對差值總和。

17.如申請專利範圍第 10 項所述之人臉說話模擬方法，其中將該些特徵向量與所記錄之該些最佳分割超平面比對，以判定該些特徵向量所屬的該語音類別的步驟包括：

依照該些特徵向量位於各該些最佳分割超平面兩邊的一比例，判定該些特徵向量所屬的該語音類別。

18.如申請專利範圍第 10 項所述之人臉說話模擬方法，其中依據各該些語音特徵所屬的該語音類別，依序顯示對應的該嘴形圖片於該人臉影像上的步驟包括：

計算所要顯示之相鄰語音特徵的該些特徵向量所佔之一權重，加乘該相鄰語音特徵對應的嘴形圖片，以顯示一混合嘴形圖片。

八、圖式：

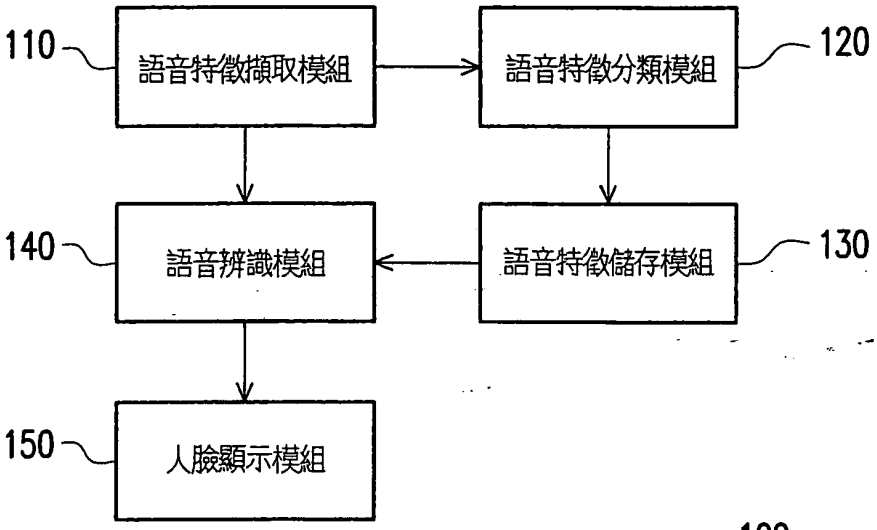


圖 1

types and the feature vectors of each two voice types are input into a support vector machine to obtain an optimal separating hyperplane for separating the feature vectors of the two voice types. Accordingly, when receiving a voice single input by a user, the voice features of the voice signal are captured and compared with the previously obtained optimal separating hyperplanes, so as to determine the voice types of the voice features. Finally, according to the voice types of the voice features, a plurality of mouth pictures corresponding to the voice types are sequentially displayed on a human face image, so as to simulate human speaking.

四、指定代表圖：

（一）本案之指定代表圖：圖 1

（二）本代表圖之元件符號簡單說明：

100：人臉說話模擬系統

110：語音特徵擷取模組

120：語音特徵分類模組

130：語音特徵儲存模組

140：語音辨識模組

150：人臉顯示模組

五、本案若有化學式時，請揭示最能顯示發明特徵的化學式：無