I471854

(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 I471854 B**

(45)公告日： 中華民國 104 (2015) 年 02 月 01 日

(21)申請案號：101138742 (22)申請日： 中華民國 101 (2012) 年 10 月 19 日

(51)Int. Cl.： *G10L13/02 (2013.01)*

(71)申請人：財團法人工業技術研究院 (中華民國) INDUSTRIAL TECHNOLOGY RESEARCH INSTITUTE (TW)

新竹縣竹東鎮中興路 4 段 195 號

(72)發明人：林政源 LIN, CHENG YUAN (TW) ；林政賢 LIN, CHENG HSIEN (TW) ；郭志忠 KUO, CHIH CHUNG (TW)

(74)代理人：洪堯順

(56)參考文獻：

TW 200741645A US 7402745B2

US 2010/0324901A1 US 2012/0116766A1

Yannis Stylianou, "A Simple and Fast Way of Generating a Harmonic Signal", IEEE Signal Processing Letters, Vol. 7, No. 5, May 2000

審查人員：涂淑惠

申請專利範圍項數：34 項 圖式數：13 共 46 頁

(54)名稱

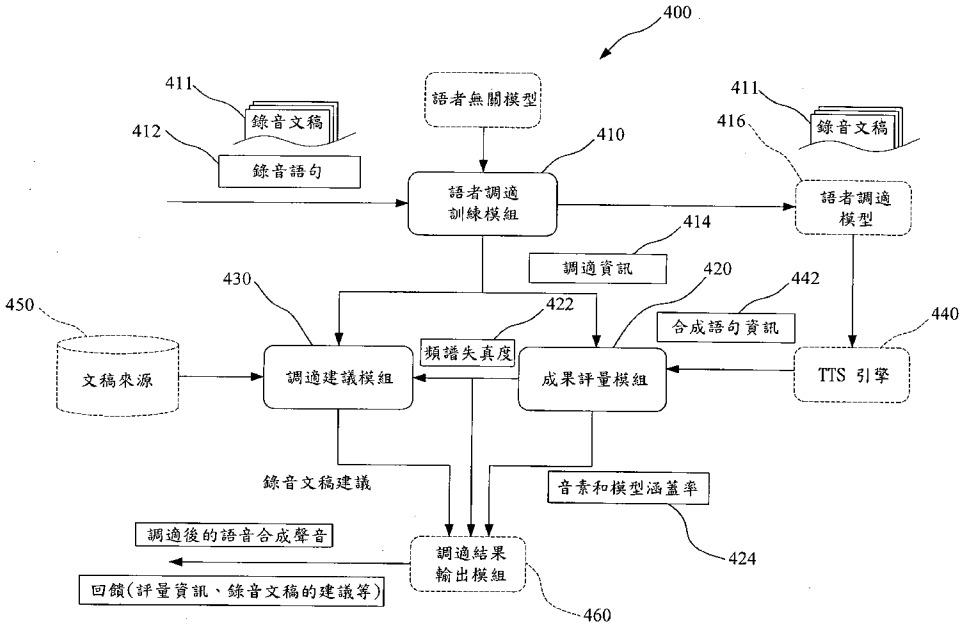
引導式語者調適語音合成的系統與方法及電腦程式產品

GUIDED SPEAKER ADAPTIVE SPEECH SYNTHESIS SYSTEM AND METHOD AND COMPUTER PROGRAM PRODUCT

(57)摘要

根據一種引導式語者調適語音合成系統的一實施例，一語者調適訓練模組，根據輸入之錄音文稿與對應的錄音語句，輸出調適資訊與語者調適模型。一文字轉語音合成引擎，接收該錄音文稿、該語者調適模型，輸出合成語句資訊。一成果評量模組，接收該調適資訊、該合成語句資訊，估算出評量資訊。一調適建議模組，根據該調適資訊以及該評量資訊內容，從文稿來源中選取出後續要錄製的錄音文稿，做為下一次調適的建議。

According to an exemplary embodiment of a guided speaker adaptive speech synthesis system, a speaker adaptive training module generates adaptation information and an adapted voice model based on the input recording text and recorded speech. A text to speech engine loads the adapted voice model and then turns the recording text into synthetic speech information. A performance assessment module receives the adaptation information and synthetic speech information to produce assessment information. An adaptation recommendation module picks up suitable recording texts from a text storage medium for next speaker adaption process by referring to the adaptation information and assessment information.



第四圖

400・・・語音合成系統

410・・・語者調適訓練模組

420・・・成果評量模組

430・・・調適建議模組

440・・・TTS 引擎

412・・・錄音語句

414・・・調適資訊

416・・・語者調適模型

442・・・合成語句資訊

424・・・音素與模型涵蓋率

422・・・頻譜失真度

450・・・文稿來源

460・・・調適結果輸出模組

TTS・・・文字轉語音

411・・・錄音文稿

I471854

公告本

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：101138942

※申請日：101.10.19　※IPC 分類：G10L 13/02 (2006.01)

一、發明名稱：（中文/英文）

引導式語者調適語音合成的系統與方法及電腦程式產品/
GUIDED SPEAKER ADAPTIVE SPEECH SYNTHESIS SYSTEM AND METHOD AND COMPUTER PROGRAM PRODUCT

二、中文發明摘要：

根據一種引導式語者調適語音合成系統的一實施例，一語者調適訓練模組，根據輸入之錄音文稿與對應的錄音語句，輸出調適資訊與語者調適模型。一文字轉語音合成引擎，接收該錄音文稿、該語者調適模型，輸出合成語句資訊。一成果評量模組，接收該調適資訊、該合成語句資訊，估算出評量資訊。一調適建議模組，根據該調適資訊以及該評量資訊內容，從文稿來源中選取出後續要錄製的錄音文稿，做為下一次調適的建議。

三、英文發明摘要：

According to an exemplary embodiment of a guided speaker adaptive speech synthesis system, a speaker adaptive training module generates adaptation information and an adapted voice model based on the input recording text and recorded speech. A text to speech engine loads the adapted

voice model and then turns the recording text into synthetic speech information. A performance assessment module receives the adaptation information and synthetic speech information to produce assessment information. An adaptation recommendation module picks up suitable recording texts from a text storage medium for next speaker adaption process by referring to the adaptation information and assessment information.

四、指定代表圖：

(一)本案指定代表圖為：第（四）圖。

(二)本代表圖之元件符號簡單說明：

400 語音合成系統	410 語者調適訓練模組
420 成果評量模組	430 調適建議模組
440 TTS 引擎	412 錄音語句
414 調適資訊	416 語者調適模型
442 合成語句資訊	424 音素與模型涵蓋率
422 頻譜失真度	450 文稿來源
460 調適結果輸出模組	TTS 文字轉語音
411 錄音文稿	

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】

本揭露係關於一種引導式語者調適(guided speaker adaptation)語音合成(speech synthesis)的系統與方法及電腦程式產品。

【先前技術】

建立語者相關(speaker dependent)語音合成系統，不論是採用語料庫(corpus based)或是統計模型為主(statistical model based)等，通常需要在專業的錄音環境下，錄製大量、穩定且說話特性一致的聲音樣本，例如收錄大於 2.5 個小時，且聲音樣本控制在穩定一致的狀態的聲音樣本。基於隱藏式馬可夫模型(Hidden Markov Model，HMM)語音合成系統搭配語者調適技術可提供快速且穩定的個人化語音合成系統的建立方案。此技術藉由一預先建立好的初始語音模型，新的語者只要輸入少於約 10 分鐘的語料就可將一平均語音模型調適成具有個人音色特質的語音模型。

基於 HMM 架構的語音合成系統，如第一圖所示，一開始輸入一串文字，經過文本分析(Text Analysis)110 可轉成文字轉語音(Text-To-Speech，TTS)系統可讀取的全標籤(full label)格式的字串 112，例如 sil-P14+P41/A:4^0/B:0+4/C:1=14/D:1@6。接著進行三種模型決策樹比對後，取得各個模型檔所對應的模型編號。此

三種模型決策樹為頻譜模型決策樹 122、音長(duration)模型決策樹 124、以及音高(pitch)模型決策樹 126。每一模型決策樹決定出約有數百到數千個 HMM 模型，也就是說，頻譜模型決策樹決定出約有數百到數千個 HMM 頻譜模型、音高模型決策樹決定出約有數百到數千個 HMM 音高模型。例如，前述全標籤格式的字串 sil-P14+P41/A:4^0/B:0+4/C:1=14/D:1@6 轉成音素與模型資訊如下：

音素:P14;

狀態 1 至 5 的頻譜模型編號:123、89、22、232、12;

狀態 1 至 5 的韻律模型編號:33、64、82、321、19。

之後，參考這些音素與模型資訊來進行合成 130。

語音合成技術不勝枚舉。一般的語者調適策略是語句越多越好，針對每個人說話特性不同並沒有設計最合適的調適內容。在現有的技術或文獻中，有些語者調適的演算法從少量的語料去調適全部的語音模型，並設計模型之間彼此共享調適資料的行為。理論上，每一語音模型代表了不同的聲音特性，所以過度共享不同特性的資料來進行語者調適，也會模糊化模型原本的特性而影響到合成的品質。

有的語音合成技術的語者調適策略是先區分語者相關特徵參數、以及語者無關特徵參數，再調整語者相關特徵後，整合之前的語者特徵無關參數後再進行合成。有的

語者調適策略是利用類似語音轉換技術來調適原始音高與共振峰。有的語者調適語音合成進行語者調適的演算法後，並無再探討相關的調適成果以及調適語句推薦的部分。有的語音合成技術在設計語料庫時，並無涉以涵蓋率與聲音失真度為準則的語句挑選方式。

有的語音合成技術如第二圖所示，在語者調適階段210中結合高層描述訊息，例如是上下文相關韻律訊息，共同來調適目標語者的頻譜、基頻與時長模型。此技術著重在加入高層描述訊息來進行語者調適，對於語者調適後的模型沒有進行任何評量或預測的動作。有的語音合成技術如第三圖所示，比較語者調適模型所合成的語音參數與真實語音的聽感誤差，並且採用基於生成參數聽感誤差最小化的準則回頭調整原始語者到目標語者的模型轉移矩陣。此技術是著重在改變語者調適演算法的估計法則，對於語者調適後的模型沒有進行任何評量或預測的動作。

上述或現有的語音合成技術中，有的僅由文字層面分析使用者應該輸入的資料，沒有考慮實際調適之後的結果。有的預設的文稿無法在事前就知道每一使用者(客戶端)最需要調適的地方在何處。文字層面的分析通常基於目標語言的音素類別而定，而非針對初始語音模型的架構而定。語音模型的分類常會使用到大量的語言學知識，僅基於音素的語音合成是無法窺探整個語音模型的全貌。所以該預設文稿無法讓語音模型間得到平均的語音資料來

進行估算，容易出現前述模型特性模糊化的現象。

因此，如何設計一種對於語者調適後的模型進行評量或預測、考量涵蓋率與聲音失真度為準則來挑選語句、以及可推薦調適語句的語音合成技術，來提供好的聲音品質與相似度，是一個重要的議題。

【發明內容】

本揭露實施例可提供一種引導式語者調適語音合成系統與方法及電腦程式產品。

所揭露的一實施例是關於一種引導式語者調適語音合成系統。此系統包含一語者調適訓練模組(speaker adaptive training module)、一文字轉語音引擎(text to speech engine)、一成果評量模組(performance assessment module)、以及一調適建議模組(adaptation recommandation module)。此語者調適訓練模組根據輸入之錄音文稿(recording text)以及對應的錄音語句(recorded speech)，輸出調適資訊以及語者調適模型。此文字轉語音合成引擎，接收此錄音文稿、此語者調適模型，輸出合成語句資訊。此成果評量模組，將參考調適資訊、此合成語句資訊，估算出評量資訊。此調適建議模組根據此錄音語句、此調適結果、以及此評量資訊，從文稿來源中選取出後續要錄製的錄音文稿，做為下一次調適的建議。

所揭露的另一實施例是關於一種引導式語者調適語音合成方法。此方法包含:輸入錄音文稿以及錄音語句，輸出一語者調適模型以及調適資訊;載入語者調適模型以及給定錄音文稿，輸出一合成語句資訊;輸入此調適資訊、此合成語句資訊，估算出評量資訊;以及根據此錄音語句、此調適資訊、以及此評量資訊，從文稿來源中選取出後續要錄製的錄音文稿，做為下一次調適的建議。

所揭露的又一實施例是關於一種引導式語者調適語音合成的電腦程式產品。此電腦程式產品包含備有多筆可讀取程式碼的一儲存媒體，並且藉由一硬體處理器讀取此多筆可讀取程式碼來執行:輸入錄音文稿以及錄音語句，輸出一語者調適模型以及調適資訊;載入語者調適模型以及給定錄音文稿，輸出一合成語句資訊;輸入此調適資訊、此合成語句資訊，估算出評量資訊;以及根據此錄音語句、此調適資訊、以及此評量資訊，從文稿來源中選取出後續要錄製的錄音文稿，做為下一次調適的建議。

茲配合下列圖示、實施例之詳細說明及申請專利範圍，將上述及本發明之其他優點詳述於後。

【實施方式】

本揭露實施例之引導式語者調適語音合成技術是藉由輸入的錄音語句以及文稿內容等資料做出下一次調適語句的推薦，由此引導使用者針對前一次調適過程中的不

足之處再次輸入語料進行補強。其中資料的評量可分為涵蓋率以及頻譜失真度的評量。在本揭露實施例中，涵蓋率以及頻譜失真度的估算結果可搭配一演算法，例如貪婪式演算法等的設計，再從一文稿來源中挑選出最適合的調適語句並且將該評量結果回饋給使用者或客戶端、或一處理文稿與語音輸入的模組等。其中涵蓋率可根據輸入文稿轉換為可讀取的全標籤(full label)格式的字串後，分析對應到音素以及語者無關模型內容的涵蓋比例。頻譜失真度藉由比對錄音語句與調適後的合成語句兩者的頻譜參數，經過時間校正後所量測出的頻譜失真度而定。

語者調適基本上是利用調適語料來調整所有的語音模型，這些語音模型例如是採用基於 HMM 架構於進行合成時所參考的多個 HMM 頻譜模型、多個 HMM 音長模型、以及多個 HMM 音高模型。在本揭露實施例中，語者調適過程中被調適的語音模型例如是，但不限定於，採用基於 HMM 架構於進行合成時所參考的 HMM 頻譜模型、HMM 音長模型、HMM 音高模型。舉前述基於 HMM 模型為例來說明語者調適及訓練。理論上，當進行調適的錄音語料所轉成之可讀取的全標籤格式的字串所對應到的模型編號足夠廣泛，也就是說能包含原本 TTS 系統中的大部分模型分佈，那麼獲得的調適成果可以更好。基於此基本的理論點，本揭露實施例設計一種可利用演算法，例如貪婪演算法(greedy algorithm)，進行最大化的模型涵蓋率的挑選方法，來選取出後續要錄製的錄音文稿，以更有效

率地進行語者調適。

既有的語者調適是根據輸入的錄音語句，進行語者無關(Speech Independent，SI)語音合成模型的調適訓練，產生語者調適的(Speech Adaptive，SA)語音合成模型，並且由一 TTS 引擎直接根據此 SA 語音合成模型來進行語音合成。與既有的語音合成技術不同的是，本揭露實施例之語音合成系統在進行既有的語者調適訓練後，還加入了一成果評量模組與一調適建議模組，使得語者調適過程中可以根據目前調適成果做不同後續文稿建議，以及提供目前調適語句的評量資訊供使用者(客戶端)參考。此成果評量模組可以估算出調適語句的音素涵蓋率、模型涵蓋率、以及頻譜失真度。此調適建議模組可以根據語者調適訓練後的調適結果、以及成果評量模組估算出的目前調適語句的評量資訊，從文稿來源中選取出後續要錄製的文稿，做為下一次調適的推薦。依此，經由不斷地調適與提供文稿建議的方式進行有效率的語者調適，使得此語音合成的系統可以提供好的聲音品質與相似度。

承上述，第四圖是根據本揭露一實施例，說明一種引導式語者調適語音合成系統。參考第四圖，語音合成系統 400 包含一語者調適訓練模組 410、一文字轉語音(TTS)引擎 440、一成果評量模組 420、以及一調適建議模組 430。語者調適訓練模組 410 根據錄音文稿 411 以及錄音語句 412 調適出一語者調適模型 416。語者調適訓練模組 410

根據錄音文稿 411 內容進行分析後，可收集到錄音文稿 411 所對應的音素與模型資訊。語者調適訓練模組 410 調適後的一調適資訊 414 至少包括輸入的錄音語句 412、分析錄音語句 412 所產生的切音資訊、錄音文稿 411 所對應的音素與多種模型資訊。此多種模型資訊例如可採用頻譜模型資訊與韻律模型資訊。此韻律模型即前述的音高模型，因為頻譜決定了音色，而音高決定了韻律的大致趨勢。

一文字轉語音(TTS)引擎 440 根據錄音文稿 411 以及語者調適模型 416，輸出合成語音資訊 442。此合成語音資訊 442 至少包括合成語句以及合成語句的切音資訊。

成果評量模組 420 結合調適資訊 414 以及合成語句資訊 442，估算出目前調適語句的評量資訊，此評量資訊包含如音素與模型涵蓋率 424、以及一或多個語音差異評估參數(例如頻譜失真度 422 等)。音素與模型涵蓋率 424 包括如音素涵蓋率、頻譜模型涵蓋率、韻律型涵蓋率等。一旦有了音素和模型的統計資訊之後，套用音素涵蓋率公式以及模型涵蓋率公式即可求得音素與模型涵蓋率。此一或多個語音差異評估參數(如頻譜失真度及/或韻律失真度等)的估算可利用語者調適訓練模組 410 所輸入的錄音語句、錄音語句的切音資訊、以及 TTS 引擎 440 提供的合成語句和合成語句的切音資訊，並透過多個執行程序來求得。如何估算出音素與模型涵蓋率與語音差異評估參數的細節與範例說明將再描述。

調適建議模組430根據語者調適訓練模組410所輸出的調適資訊414、以及成果評量模組420估算出的目前錄音語句的評量資訊，例如頻譜失真度，從一文稿來源(例如文稿資料庫)450中選取出後續要錄製的錄音文稿，做為下一次調適的建議。調適建議模組430選取錄音文稿的策略例如是，能夠讓音素/模型的涵蓋率最大化。語音合成系統400可輸出成果評量模組420估算出的目前調適語句的評量資訊，如音素與模型涵蓋率、頻譜失真度等，以及調適建議模組430做出的下一次調適語句的建議，如錄音文稿的建議，至一調適結果輸出模組460。調適結果輸出模組460可將這些資訊，如評量資訊、錄音文稿的建議等，回饋給使用者或客戶端、或一處理文字與語音輸入的模組等。依此，經由不斷地調適與提供文稿建議的方式進行有效率的語者調適，使得語音合成系統400也可經由調適結果輸出模組460輸出調適後的語音合成聲音。

第五圖是根據本揭露一實施例，說明語者調適訓練模組從一輸入文稿收集到每一筆全標籤資訊所對應的音素與模型資訊的範例。在第五圖的例子中，語者調適訓練模組將輸入文稿轉成多筆全標籤資訊516，將此多筆全標籤資訊516進行比對後，收集到每一筆全標籤資訊所對應的音素資訊、狀態(state)1至5的頻譜模型編號、以及狀態1至5的韻律模型編號。當模型的種類收集越多(表示涵蓋率越高)時，則代表平均語音模型可能獲得更好的調適結

果。

從第五圖的例子中可窺知，當輸入一筆全標籤資訊到一語音合成系統後，經過如決策樹比對之後可獲得它的頻譜模型編號與韻律模型編號。從全標籤資訊本身也可看出它的音素資訊，以 sil-P14+P41/A:4^0/B:0+4/C:1=14/D:1@6 為例，它的音素即 P14(注音為ㄒ)，而左音素則為 sil(代表靜音(silence))，右音素則為 P41(注音為ㄧ)。因此收集調適語料的音素與模型資訊是相當直覺的，此資訊收集過程是執行於調適訓練模組之中。有了音素與模型的統計資訊之後，就可以套用音素涵蓋率公式以及模型涵蓋率公式來估算出音素與模型涵蓋率。

第六圖是根據本揭露一實施例，估算音素涵蓋率與模型涵蓋率的公式範例。在第六圖的涵蓋率計算公式 610 中，估算音素涵蓋率的公式中，分母的值(此例為 50)代表 TTS 引擎有 50 種不同的音素;估算模型涵蓋率的公式中，假設頻譜或韻律模型皆有 5 個不同的狀態。當模型為頻譜模型時，模型涵蓋率的公式中，$StateCoverRate_s$ 中的分母(即變數 $ModelCount_s$)代表狀態 s 的頻譜模型種類數，分子(即變數 $Num\_UniqueNodel_s$)代表狀態目前收集到的頻譜模型種類數，依此模型涵蓋率的公式估算出頻譜模型涵蓋率。類似地，當模型為韻律模型時，從模型涵蓋率的公式中，可估算出韻律模型涵蓋率。

成果評量模組420估算出的語音差異評估參數包含頻譜失真度時，相較於涵蓋率的估算是比較複雜的。如第七圖所示，在本揭露的實施例中，頻譜失真度的估算可利用調適訓練模組 410 所輸出錄音語句、錄音語句的切音資訊、以及 TTS 引擎 440 所提供的合成語句、合成語句的切音資訊，再執行特徵擷取(feature extraction)710、時間校正(time alignment)720、以及頻譜失真計算(spectral distortion calculation)730 來求得。

特徵擷取是先求取語音的特徵參數，例如可採用梅爾倒頻譜(Mel-Cepstral)參數，或是線性預測編碼(Linear Prediction Coding，LPC)、或是線頻譜(Line Specturm Frequency，LSF)、或是感知線性預測(Perceptual Linear Prediction，PLP)等方法作為參考語音特徵，接著再進行錄音語句與合成語句的時間校正比對。錄音語句及合成語句的切音資訊雖然是已知的，但是錄音語句與合成語句之間，每一字的發音長度並不一致，因此進行頻譜失真度計算之前，需先進行時間校正。時間校正的做法可採用動態時間扭曲(Dynamic Time Warping，DTW)。最後利用如梅爾倒頻譜失真(Mel-Cepstral Distortion，MCD)作為頻譜失真度指標計算的基礎。MCD 的計算公式如下:

$$MCD_{frame} = \frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{N}(mcp_i^{(syn)} - mcp_i^{(tar)})^2}$$ ，其中

mcp 是梅爾倒頻譜參數，syn 是來自調適語句(adapted speech)的合成音框(synthesized frame)，tar 是來自實際語句

(real speech)的目標音框(target frame)，N 是 mcp 維度(dimension)。每一語音單位(例如音素)的頻譜失真度(Distortion)可估算如下:

$Distortion = \frac{\sum_{f=1}^{K} MCD_f}{K}$，其中 K 是音框的個數。

當 MCD 值越高時，表示合成結果相似度越低。因此，系統目前的調適結果可採用此指標來表示。

調適建議模組 430 結合來自語者調適訓練模組 410 的調適資訊 414、以及成果評量模組 420 估算出的評量資訊如頻譜失真度，從一文稿來源中選取出後續錄音文稿的建議。如第八圖所示，在本揭露的實施例中，調適建議模組 430 還利用基於音素與模型涵蓋率最大化(Phone/Model based coverage maximization)演算法 820，例如貪婪演算法(greedy algorithm)，來挑選最適合的錄音文稿，並且在執行此演算法的過程中，先參考權重重估算(weight re-estimation)810 的結果;最後輸出後續錄音文稿的建議。

承上述之引導式語者調適語音合成系統及各模組的描述，第九圖是根據本揭露的一實施例，說明一種引導式語者調適語音合成方法。如第九圖所示，此語音合成方法 900 先輸入錄音文稿以及對應的錄音語句進行語者調適訓練，輸出語者調適模型以及調適資訊(步驟 910)。接著將語者調適模型以及錄音文稿提供給一 TTS 引擎，輸出合成

語音資訊(步驟920)。此語音合成方法900再根據此調適資訊、以及此合成語音資訊，估算出目前錄音語句的評量資訊(步驟930)。最後再根據此調適資訊、以及此評量資訊，從一文稿來源中選取出後續要錄製的錄音文稿，做為下一次調適的建議(步驟940)。

承上述，此引導式語者調適語音合成方法可包含:輸入錄音文稿以及錄音語句，輸出一語者調適模型以及調適資訊;載入語者調適模型以及給定錄音文稿,輸出一合成語句資訊;輸入此調適資訊、此合成語句資訊，估算出評量資訊;以及根據此調適資訊、以及此評量資訊，從文稿來源中選取出後續要錄製的錄音文稿，做為下一次調適的建議。

此調適資訊至少包括錄音語句以及錄音語句的切音資訊以及錄音語句對應的該音素與模型資訊。此合成語音資訊至少包括合成語句以及合成語句的切音資訊。此評量資訊至少包括音素與模型涵蓋率、以及一或多個語音差異評估參數(如頻譜失真度)。

在語音合成方法900中，如何從一輸入文稿的錄音語句收集到所對應的音素與模型資訊、如何估算音素涵蓋率與模型涵蓋率、如何估算頻譜失真度、以及選取錄音文稿的策略等相關內容皆已描述於前述本揭露實施例中，此處不再重述。如之前所述，本揭露的實施例是先進行一權重

重估算後，再利用基於音素與模型涵蓋率最大化的演算法來挑選錄音文稿。第十圖與第十一圖是根據本揭露的實施例，分別說明基於音素與模型涵蓋率最大化的演算法的流程。

參考第十圖之演算法的流程，首先，此基於音素涵蓋率最大化演算法根據一當次的評量資訊，進行權重重估算(步驟 1005)。進行權重重估算後可得到一音素之新的權重 Weight(PhoneID)、以及此音素的一更新的影響力 Influence(PhoneID)，其中 PhoneID 是音素的識別碼(identifier)。此權重重估算的細節將於第十二圖中描述。然後，初始化一文稿來源中每一候選語句的分數為 0(步驟 1010);此演算法根據一分數函數(score function)的定義，計算文稿來源中每一句子的分數，並且將分數正規化(步驟 1012);例如可根據此句子中音素的個數來進行此正規化(例如將總分數除以音素的個數)。定義一音素的分數函數的範例如下:

$$Score = Weigtht(PhoneID) \times 10^{Influence(PhoneID)}$$

在上述的分數函數中，一音素的分數是依此音素的權重和影響力來決定。音素的權重 Weight(PhoneID)的系統初始值是取此音素出現次數的倒數當作此音素的權重(weight)，所以在儲存媒體例如資料庫中出現越多次者，其權重越低。音素的影響力 Influence(PhoneID)初始值假設定為 20，表示每一音素最多出現 20 次，之後其分數影響

力可視為不計;當音素被挑選過 1 次之後，此音素的Influence(PhoneID)將被減 1，對其分數的貢獻將變成10^{19}，以此類推，當此音素被挑選過 j 次之後，對其分數的貢獻將變成 10^{20-j}。也就是說，一音素的Influence(PhoneID)與此音素被挑選過的次數有關，被挑選過的次數越多者，其影響力越低。

音素種類越多元的候選語句獲得的分數則越高，最後從中挑選分數最高者從該文稿來源移出到調適建議的句子集合中(步驟 1014)，並且該挑選到的句子其所包含的音素之影響力將被降低(步驟 1016)，以利提高其他音素下次被挑選的機會。當被挑選出的句子的個數未超過一預定值時(步驟 1018)，則進行步驟 1012，而重新計算該文稿來源中的所有剩下的候選語句的分數，重覆上述過程，直到挑選出的句子的個數超過一預定值為止。

也就是說，此基於音素涵蓋率最大化演算法定義一音素的分數函數，對於一文稿來源中每一個候選語句進行分數估算，音素種類越多元的候選語句獲得的分數則越高，最後從中挑選分數最高者從該文稿來源移出到調適建議的句子集合中，並且該挑選到的句子其所包含的音素之影響力將被降低，以利提高其他音素下次被挑選的機會。接著重新計算該文稿來源中的所有候選語句的分數，重覆上述過程，直到挑選出的句子的個數超過一預定值為止。

參考第十一圖之演算法的流程，首先，此基於模型涵蓋率最大化演算法根據一當次的評量資訊，進行權重重估算(步驟 1105)。進行權重重估算後可得到兩模型之新的 MCP 權重和 LF0 權重以及此兩模型的兩更新影響力，即 $Influence(M_s^L)$ 與 $Influence(P_s^L)$，其中 M_s^L 表示當狀態為 S 且文稿標籤資訊為 L 時所對應到的頻譜(MCP)模型，同理 P_s^L 表示當狀態為 S 且文稿標籤資訊為 L 時所對應到的韻律(LF0)模型。此文稿標籤資訊定義為輸入的錄音文稿，經由語者調適訓練模組的文稿分析後所得的全標籤資訊，如圖五中的 516。此權重重估算的細節將於第十二圖中描述。然後，初始化一文稿來源中每一候選語句的分數為 0(步驟 1110); 此演算法根據一分數函數(score function)的定義，計算文稿來源中每一句子的分數，並且將分數正規化(步驟 1112); 例如可根據此句子中的 L(文稿標籤)個數來進行此正規化(例如將總分數除以音素的個數)。定義一模型的分數函數的範例如下:

$$\text{Score} = \sum_{s=1}^{5} \left(MCPScore(M_s^L) + LF0Score(P_s^L) \right)$$

$$MCPScore(M_s^L) = Weight(M_s^L) \times 10^{Influence(M_s^L)}$$

$$LF0Score(P_s^L) = Weight(P_s^L) \times 10^{Influence(P_s^L)}$$

在上述的分數函數中，分數是依此一頻譜模型分數與一韻律模型分數來決定，並且一頻譜或韻律模型的分數是依此模型的權重和影響力來決定。在上述的模型分數函數中，頻譜模型的權重 $Weight(M_s^L)$ 以及韻律模型的權重 $Weight(P_s^L)$

的系統初始值分別是取其出現次數的倒數分別當作 MCP 模型的權重與 LF0 模型的權重，所以模型在儲存媒體例如資料庫中出現越多次者，其模型權重越低。$Influence(M_s^L)$ 與 $Influence(P_s^L)$ 的值一開始例如皆為 5，每出現一次，其值減 1。也就是說，$Influence(M_s^L)$ 及 $Influence(P_s^L)$ 的值與其模型被挑選過的次數有關，被挑選過的次數越多者，其影響力越低。

MCP 模型與 LF0 模型種類越多元的候選語句獲得的分數則越高，最後從中挑選分數最高者從該文稿來源移出到調適建議的句子集合中(步驟 1114)，並且該挑選到的句子其所包含的模型之影響力將被降低(步驟 1116)，以利提高其他模型下次被挑選的機會。當被挑選出的句子的個數未超過一預定值時(步驟 1118)，則進行步驟 1112，而重新計算該文稿來源中的所有剩下的候選語句的分數，重覆上述過程，直到挑選出的句子的個數超過一預定值為止。

也就是說，此基於模型涵蓋率最大化演算法定義一模型的分數函數，對於一文稿來源中每一個候選語句進行分數估算，模型種類越多元的候選語句獲得的分數則越高，最後從中挑選分數最高者從該文稿來源移出到調適建議的句子集合中，並且該挑選到的句子其所包含的模型之影響力將被降低，以利提高其他模型下次被挑選的機會。接著重新計算該文稿來源中的所有候選語句的分數，重覆上述過程，直到挑選出的句子的個數超過一預定值為止。

承上述第十圖與第十一圖的流程，在基於音素涵蓋率最大化或是基於模型涵蓋率最大化的演算中，權重重估算扮演了關鍵性角色。它根據頻譜失真度來決定新的音素權重、及模型權重，例如新的$Weight(PhoneID)$、及$Weight(M_s^L)$、$Weight(P_s^L)$，並且是利用一種音色相似度的方法來動態調整權重的高低。此權重重估算是利用音色相似度的方法來動態調整權重的高低，使得後續挑選文稿的參考不只是考量到涵蓋率(只根據文本參考)，也能兼顧合成結果的回饋。而音色相似度通常是以頻譜失真度來估算，假如一語音單位(例如音素或音節或字)的頻譜失真度過高，表示它調適的結果不夠好，後續的文稿應該要加強此單位的挑選，因此它的權重應該要調升;反之，當一語音單位的頻譜失真度很低，表示它調適的結果已經夠好，後續應調降它的權重，讓其他語音單位被挑選的機會增加。依此，在本揭露實施例中，權重調整原則為，當一語音單位的頻譜失真度高於一高門檻值(例如，原始語句的平均失真度+原始語句的標準差)時，調升此語音單位的權重;當一語音單位的頻譜失真度低於一低門檻值(例如，原始語句的平均失真度－原始語句的標準差)時，調降此語音單位的權重。

第十二圖是根據本揭露一實施例，說明一種權重重估算的調整方式。在第十二圖之權重重估算的調整方式的公式 1200 中，Di 表示某一語音單位(例如以音素為單位)的第 i 個失真度(distortion)，D_{mean} 表示調適語料的平均失真度，D_{std} 表示調適語料的標準差失真度。N 表示參與此次

權重調整的單位個數(例如 P14 這個音素共有 5 個參與計算)，同一種單位所估算的各個因子 $Factor_i$ 不盡相同，因此求取這些 $Factor_i$ 的平均(即平均因子 F)作為代表。最後，新權重是根據平均因子 F 來進行調整，調整公式的範例為，新權重=權重x(1+F)，其中平均因子 F 的值可能為正值或負值。

第十三圖是合成語句和原始語句的頻譜失真度分布的一個範例圖，其中橫軸代表不同的音素，縱軸代表其頻譜失真度(縱軸的單位為 dB)，計算頻譜失真度的語音單位為音素。因為音素 5 至音素 8 的頻譜失真度皆高於(D_{mean}+D_{std})，因此根據本揭露實施例之權重調整原則，可依第十二圖的調整方式來調升音素 5、音素 6、音素 7、以及音素 8 的權重;而音素 11、音素 13、音素 20、以及音素 37 的頻譜失真度皆低於(D_{mean}−D_{std})，因此根據本揭露實施例之權重調整原則，可依第十二圖的調整方式來調降音素 11、音素 13、音素 20、以及音素 37 的權重。

上述本揭露實施例之引導式語者調適語音合成的方法可藉由一電腦程式產品來實現。此電腦程式產品可藉由至少一硬體處理器讀取內嵌於一儲存媒體的程式碼來執行此方法。依此，根據本揭露又一實施例，此電腦程式產品可包含備有多筆可讀取程式碼的一儲存媒體，並且藉由至少一硬體處理器讀取此多筆可讀取程式碼來執行:輸入錄音文稿以及錄音語句，輸出一語者調適模型以及調適資

訊;載入語者調適模型以及給定錄音文稿,輸出一合成語句資訊;輸入此調適資訊、此合成語句資訊,估算出評量資訊;以及根據此調適資訊、以及此評量資訊,從文稿來源中選取出後續要錄製的錄音文稿,做為下一次調適的建議。

綜上所述,本揭露實施例提供一種引導式語者調適語音合成系統與方法。其技術先輸入錄音文稿和錄音語句,輸出為調適資訊以及語者調適模型;一 TTS 引擎讀取此語者調適模型以及此錄音文稿,輸出合成語句資訊;接著結合此調適資訊以及此合成語句資訊,估算出評量資訊;再根據此調適資訊、以及此評量資訊,來選取出後續要錄製的錄音文稿,做為下一次調適的建議。此技術考量音素與模型涵蓋率,以聲音失真度為準則來挑選語句,以及做出下一次調適語句的推薦,由此引導使用者/客戶端針對前一次調適過程中的不足之處補強輸入語料,以提供好的聲音品質與相似度。

以上所述者僅為本揭露實施例,當不能依此限定本揭露實施之範圍。即大凡本發明申請專利範圍所作之均等變化與修飾,皆應仍屬本發明專利涵蓋之範圍。

【圖式簡單說明】

第一圖是基於 HMM 架構的語音合成技術的一範例示意圖。

第二圖是一種結合高層描述信息和模型自適應的語者轉換技術的一範例示意圖。

第三圖是一種基於生成參數聽感誤差最小化的模型自適應技術的一範例示意圖。

第四圖是根據本揭露一實施例，說明一種引導式語者調適語音合成系統。

第五圖是根據本揭露一實施例，說明語者調適訓練模組從一輸入文稿的範例，收集到每一筆全標籤資訊所對應的音素與模型資訊。

第六圖是根據本揭露一實施例，估算音素涵蓋率與模型涵蓋率的公式範例。

第七圖是根據本揭露一實施例，說明成果評量模組估算頻譜失真度的運作。

第八圖是根據本揭露一實施例，說明調適建議模組的運作。

第九圖是根據本揭露的一實施例，說明一種引導式語者調適語音合成方法。

第十圖是根據本揭露的一實施例，說明基於音素涵蓋率最大演算法的流程。

第十一圖是根據本揭露的實施例，說明基於模型涵蓋率最大演算法的流程。

第十二圖是根據本揭露一實施例，說明一種權重重估算的

調整方式。

第十三圖是一個句子的範例代表圖，其頻譜失真度計算的單位為音素。

【主要元件符號說明】

110 文本分析　　112 全標籤格式的字串

122 頻譜模型決策樹　　124 音長模型決策樹

126 音高模型決策樹　　130 合成

210 語者調適階段　　411 錄音文稿

400 語音合成系統　　410 語者調適訓練模組

420 成果評量模組　　430 調適建議模組

440 TTS 引擎　　412 錄音語句

414 調適資訊　　416 語者調適模型

442 合成語句資訊　　424 音素與模型涵蓋率

422 頻譜失真度　　450 文稿來源

460 調適結果輸出模組　　TTS 文字轉語音

516 多筆全標籤資訊

610 涵蓋率計算公式

710 特徵擷取　　720 時間調整

730 頻譜失真計算

810 權重重估算

820 基於音素與模型涵蓋率最大化演算法

910 輸入錄音文稿以及對應的錄音語句進行語者調適訓練，輸出語者調適模型以及調適資訊

920 將語者調適模型以及錄音文稿提供給一 TTS 引擎，輸出合成語音資訊

930 根據此調適資訊、以及此合成語音資訊，估算出目前錄音語句的評量資訊

940 根據此調適資訊、以及此評量資訊，從一文稿來源中選取出後續要錄製的錄音文稿，做為下一次調適的建議

1005 根據一當次的評量資訊，進行權重重估算

1010 初始化一文稿來源中每一候選語句的分數為 0

1012 根據一分數函數的定義，計算文稿來源中每一句子的分數，並且將分數正規化

1014 從中挑選分數最高者從該文稿來源移出到調適建議的句子集合中

1016 該挑選到的句子其所包含的音素之影響力將被降低

1018 當被挑選出的句子的個數未超過一預定值時

1105 根據一當次的錄音語料資訊，進行權重重估算

1110 初始化一文稿來源中每一候選語句的分數為 0

1112 根據一分數函數的定義，計算文稿來源中每一句子的分數，並且將分數正規化

1114 從中挑選分數最高者從該文稿來源移出到調適建議的句子集合中

1116 該挑選到的句子其所包含的模型之影響力將被降低

1118 被挑選出的句子的個數未超過一預定值時

1200 權重重估算的調整方式的公式

Di 某一語音單位(例如音素)的第 *i* 個失真度

D_{mean} 調適語料的平均失真度

D_{std} 調適語料的標準差失真度

N 參與此次權重調整的單位個數

NewWeight 新權重　　*Weight* 新權重

$Factor_i$ 各個因子　　*F* 平均因子

七、申請專利範圍：

1. 一種引導式語者調適語音合成系統，包含：

一語者調適訓練模組，根據輸入之錄音文稿與對應的錄音語句，輸出至少包含頻譜模型資訊與韻律模型資訊的調適資訊與語者調適模型；

一文字轉語音合成引擎，接收該錄音文稿與該語者調適模型，輸出合成語句資訊；

一成果評量模組，接收該調適資訊、該合成語句資訊，估算出評量資訊；以及

一調適建議模組，根據該調適資訊與該評量資訊內容，從文稿來源中選取出後續要錄製的錄音文稿，以做為下一次調適的建議。

2. 如申請專利範圍第 1 項所述之系統，其中該調適訓練模組所輸出的該調適資訊至少包括：

該錄音文稿、該錄音語句、該錄音文稿對應的音素與模型資訊、以及該錄音語句對應的切音資訊。

3. 如申請專利範圍第 2 項所述之系統，其中該模型資訊至少包括該頻譜模型資訊、與該韻律模型資訊。

4. 如申請專利範圍第 1 項所述之系統，該文字轉語音合成引擎所輸出的該合成語句資訊至少包括：該錄音文稿的合成語句，以及該合成語句的切音資訊。

5. 如申請專利範圍第 1 項所述之系統，其中該評量資訊至少包括該錄音語句的音素與模型涵蓋率。

6. 如申請專利範圍第 5 項所述之系統，其中該音素與模型涵蓋率包括音素涵蓋率、頻譜模型涵蓋率、以及韻律模

型涵蓋率。

7. 如申請專利範圍第 1 項所述之系統，其中該評量資訊至少包括一或多個語音差異評估參數。

8. 如申請專利範圍第 7 項所述之系統，其中該一或多個語音差異評估參數至少包括該錄音語句和該合成語句的頻譜失真度。

9. 如申請專利範圍第 1 項所述之系統，其中該調適建議模組選取錄音文稿的策略是能夠讓該音素與模型的涵蓋率最大化。

10. 如申請專利範圍第 1 項所述之系統，其中該系統是採用基於隱藏式馬可夫模型或者隱藏式半馬可夫模型架構的語音合成系統。

11. 如申請專利範圍第 1 項所述之系統，其中該系統經由不斷地調適與提供文稿建議的方式來進行語者調適。

12. 如申請專利範圍第 1 項所述之系統，其中該系統輸出該合成語句、該成果評量模組估算出的該目前錄音語句的評量資訊、以及該調適建議模組做出的下一次調適語句的建議。

13. 一種引導式語者調適語音合成方法，包含：

輸入錄音文稿與對應的錄音語句，輸出語者調適模型與至少包含頻譜模型資訊與韻律模型資訊的調適資訊；

載入該語者調適模型，輸入該錄音文稿，以合成出合成語音資訊；

結合該調適資訊與該合成語音資訊，估算出評量資

訊；以及

根據該調適資訊與該評量資訊內容，從文稿來源中選取出後續要錄製的錄音文稿，做為下一次調適的建議。

14. 如申請專利範圍第 13 項所述之方法，其中該評量資訊包括該目前錄音語句的音素涵蓋率、頻譜模型涵蓋率、韻律模型涵蓋率、以及一或多個語音差異評估參數。

15. 如申請專利範圍第 13 項所述之方法，其中該一或多個語音差異評估參數至少包括頻譜失真度。

16. 如申請專利範圍第 13 項所述之方法，其中該方法先進行一權重重估算後，再利用一基於音素涵蓋率最大化的演算法與一基於模型涵蓋率最大化演算法來選取出後續要錄製的該錄音文稿，該音素涵蓋率係套用音素涵蓋率公式而求得，該模型涵蓋率係套用模型涵蓋率公式而求得。

17. 如申請專利範圍第 16 項所述之方法，其中該權重重估算是根據頻譜失真度來決定新的音素權重、及模型權重，並且是利用一種音色相似度的方法來動態調整權重的高低。

18. 如申請專利範圍第 17 項所述之方法，其中該調整權重的原則為，當一語音單位的頻譜失真度高於一高門檻值，調升該語音單位的權重；反之當一語音單位的頻譜失真度低於一低門檻值時，調降該語音單位的權重。

19. 如申請專利範圍第 18 項所述之方法，其中該語音單位是字、音節、或音素的其中一種或多種組合。

20. 如申請專利範圍第16項所述之方法，其中該基於音素涵蓋率最大化演算法定義一音素的分數函數，對於一文稿來源中每一個候選語句進行分數估算，音素種類越多元的候選語句獲得的分數則越高，最後從中挑選分數最高者從該文稿來源移出到調適建議的句子集合中，並且該挑選到的句子其所包含的音素之影響力將被降低，以利提高其他音素下次被挑選的機會，接著重新計算該文稿來源中的所有候選語句的分數，重覆上述過程，直到挑選出的句子的個數超過一預定值為止。

21. 如申請專利範圍第20項所述之方法，其中根據該音素的分數函數定義，一音素的分數是依該音素的權重和影響力來決定。

22. 如申請專利範圍第16項所述之方法，其中該基於模型涵蓋率最大化演算法定義一模型的分數函數，對於一文稿來源中每一個候選語句進行分數估算，模型種類越多元的候選語句獲得的分數則越高，最後從中挑選分數最高者從該文稿來源移出到調適建議的句子集合中，並且該挑選到的句子其所包含的模型之影響力將被降低，以利提高其他模型下次被挑選的機會，接著從新計算該文稿來源中的所有候選語句的分數，重覆上述過程，直到挑選出的句子的個數超過一預定值為止。

23. 如申請專利範圍第22項所述之方法，其中根據該模型的分數函數定義，一模型的分數是依該一頻譜模型分

數與一韻律模型分數來決定，並且一頻譜或韻律模型的分數是依該頻譜或韻律模型的權重和影響力來決定。

24. 一種引導式語者調適語音合成的電腦程式產品，包含備有多筆可讀取程式碼的一儲存媒體，並且藉由至少一硬體處理器讀取該多筆可讀取程式碼來執行：

輸入錄音文稿與對應的錄音語句，輸出語者調適模型與至少包含頻譜模型資訊與韻律模型資訊的調適資訊；

載入該語者調適模型，輸入該錄音文稿，以合成出合成語音資訊；

結合該調適資訊、與合成語音資訊，估算出評量資訊；以及

根據該調適資訊與該評量資訊內容，從文稿來源中選取出後續要錄製的錄音文稿，做為下一次調適的建議。

25. 如申請專利範圍第 24 項所述之電腦程式產品，其中該評量資訊包括該目前錄音語句的音素涵蓋率、頻譜模型涵蓋率、韻律模型涵蓋率、以及一或多個語音差異評估參數。

26. 如申請專利範圍第 24 項所述之電腦程式產品，其中該一或多個語音差異評估參數至少包括頻譜失真度。

27. 如申請專利範圍第 24 項所述之電腦程式產品，其中該方法先進行一權重重估算後，再利用一基於音素涵蓋率最大化的演算法與一基於模型涵蓋率最大化的演算法來選取出後續要錄製的該錄音文稿。

28. 如申請專利範圍第27項所述之電腦程式產品，其中該權重重估算是根據頻譜失真度來決定新的音素權重、及模型權重，並且是利用一種音色相似度的方法來動態調整權重的高低。

29. 如申請專利範圍第28項所述之電腦程式產品，其中該調整權重的原則為，當一語音單位的頻譜失真度高於一高門檻值，調升該語音單位的權重；反之當一語音單位的頻譜失真度低於一低門檻值時，調降該語音單位的權重。

30. 如申請專利範圍第29項所述之電腦程式產品，其中該語音單位是字、音節、或音素其中一種或多種組合。

31. 如申請專利範圍第27項所述之電腦程式產品，其中該基於音素涵蓋率最大化演算法定義一音素的分數函數，對於一文稿來源中每一個候選語句進行分數估算，音素種類越多元的候選語句獲得的分數則越高，最後從中挑選分數最高者從該文稿來源移出到調適建議的句子集合中，並且該挑選到的句子其所包含的音素之影響力將被降低，以利提高其他音素下次被挑選的機會，接著重新計算該文稿來源中的所有候選語句的分數，重覆上述過程，直到挑選出的句子的個數超過一預定值為止。

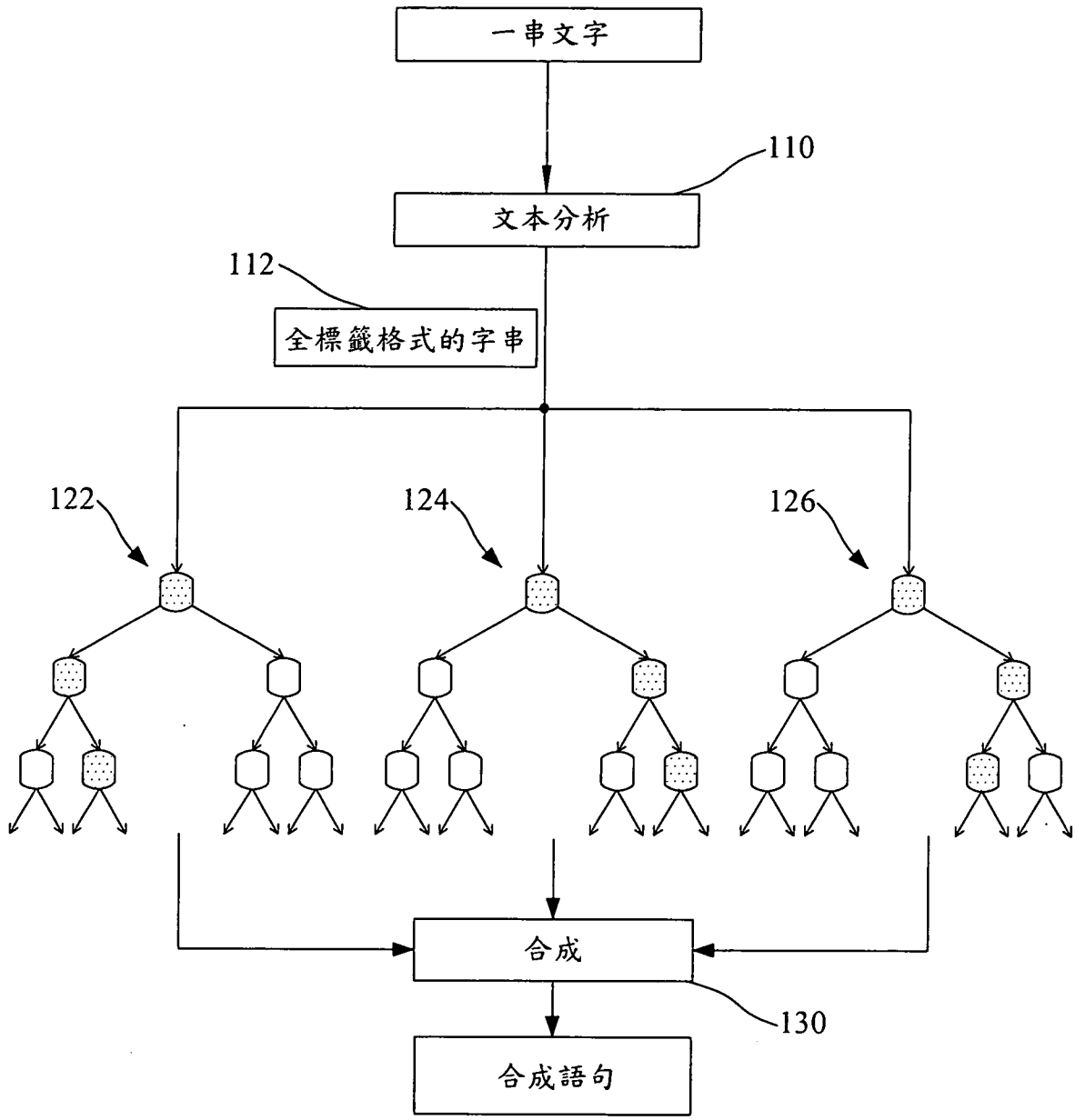
32. 如申請專利範圍第31項所述之電腦程式產品，其中根據該音素的分數函數定義，一音素的分數是依該音素的權重和影響力來決定。

33. 如申請專利範圍第27項所述之電腦程式產品，其中該

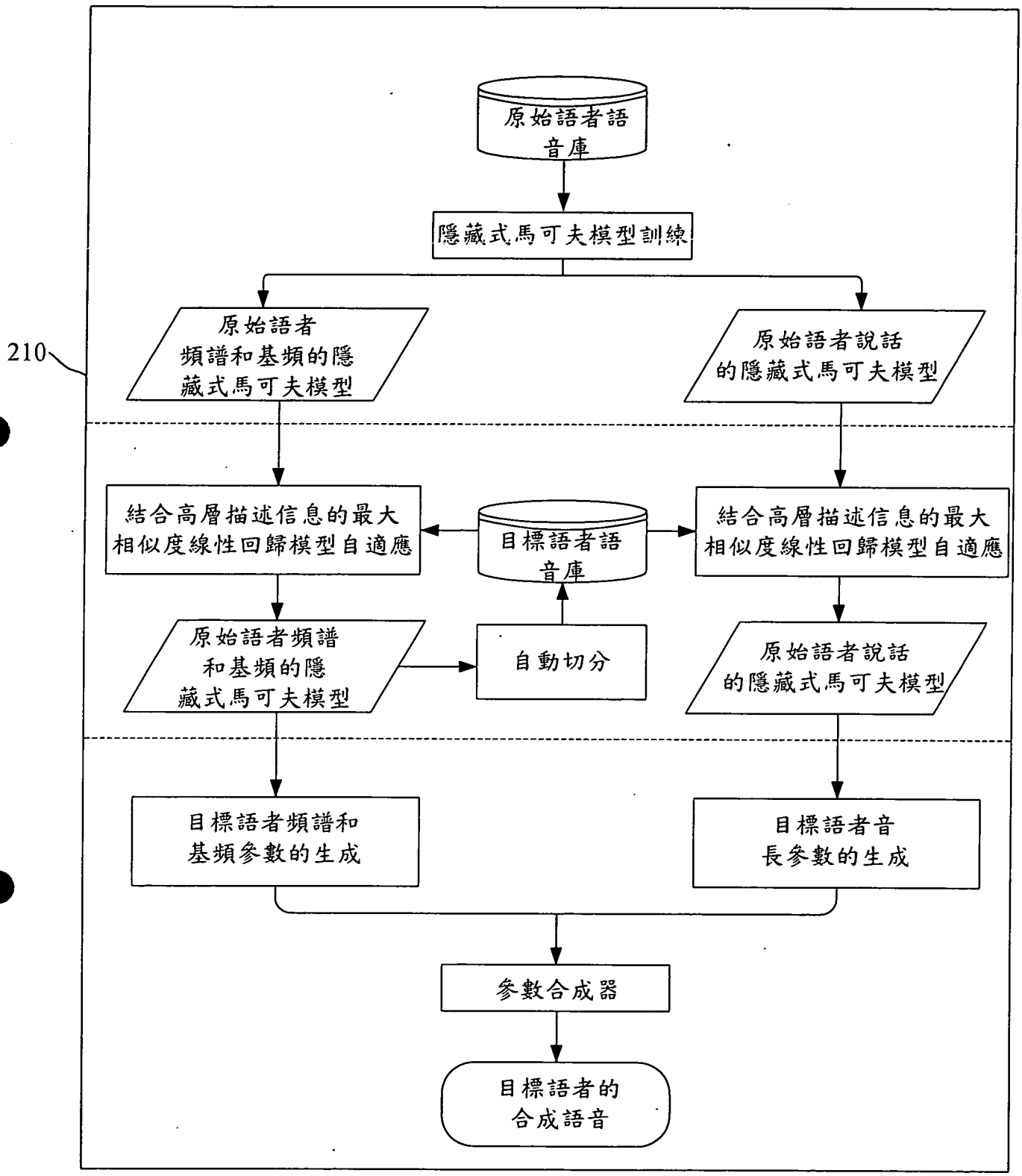
基於模型涵蓋率最大化演算法定義一模型的分數函數，對於一文稿來源中每一個候選語句進行分數估算，模型種類越多元的候選語句獲得的分數則越高，最後從中挑選分數最高者從該文稿來源移出到調適建議的句子集合中，並且該挑選到的句子其所包含的模型之影響力將被降低，以利提高其他模型下次被挑選的機會，接著從新計算該文稿來源中的所有候選語句的分數，重覆上述過程，直到挑選出的句子的個數超過一預定值為止。

34. 如申請專利範圍第 33 項所述之電腦程式產品，其中根據該模型的分數函數定義，一模型的分數是依該一頻譜模型分數與一韻律模型分數來決定，並且一頻譜或韻律模型的分數是依該頻譜或韻律模型的權重和影響力來決定。

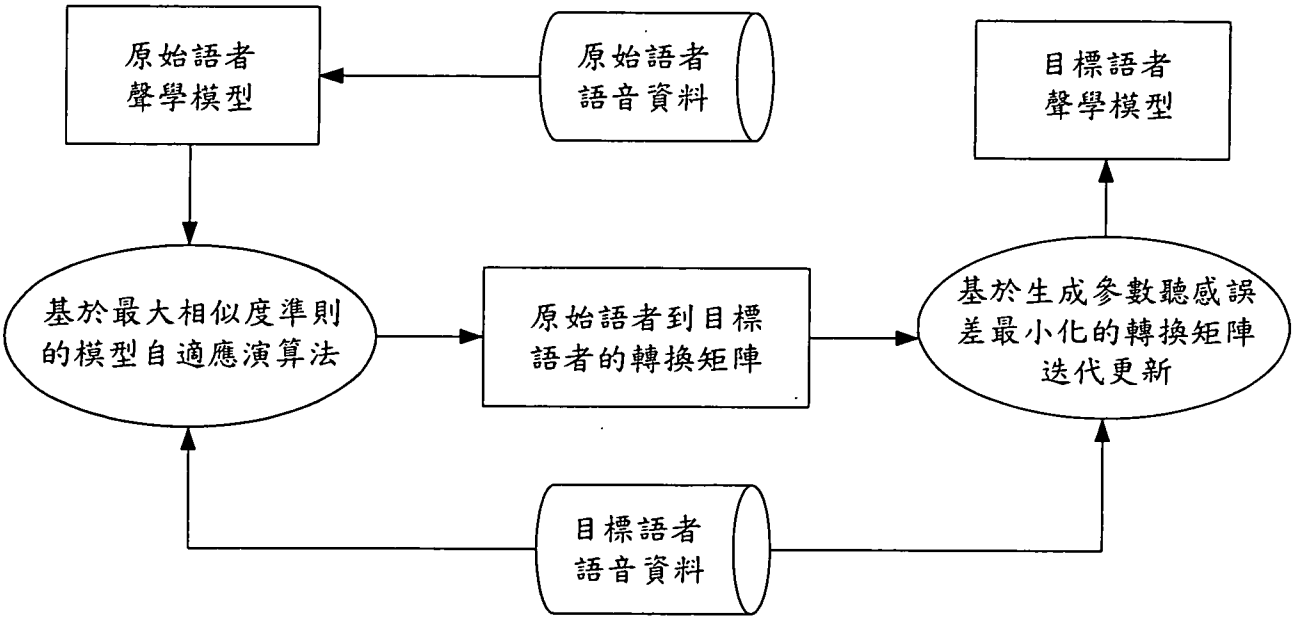
八、圖式



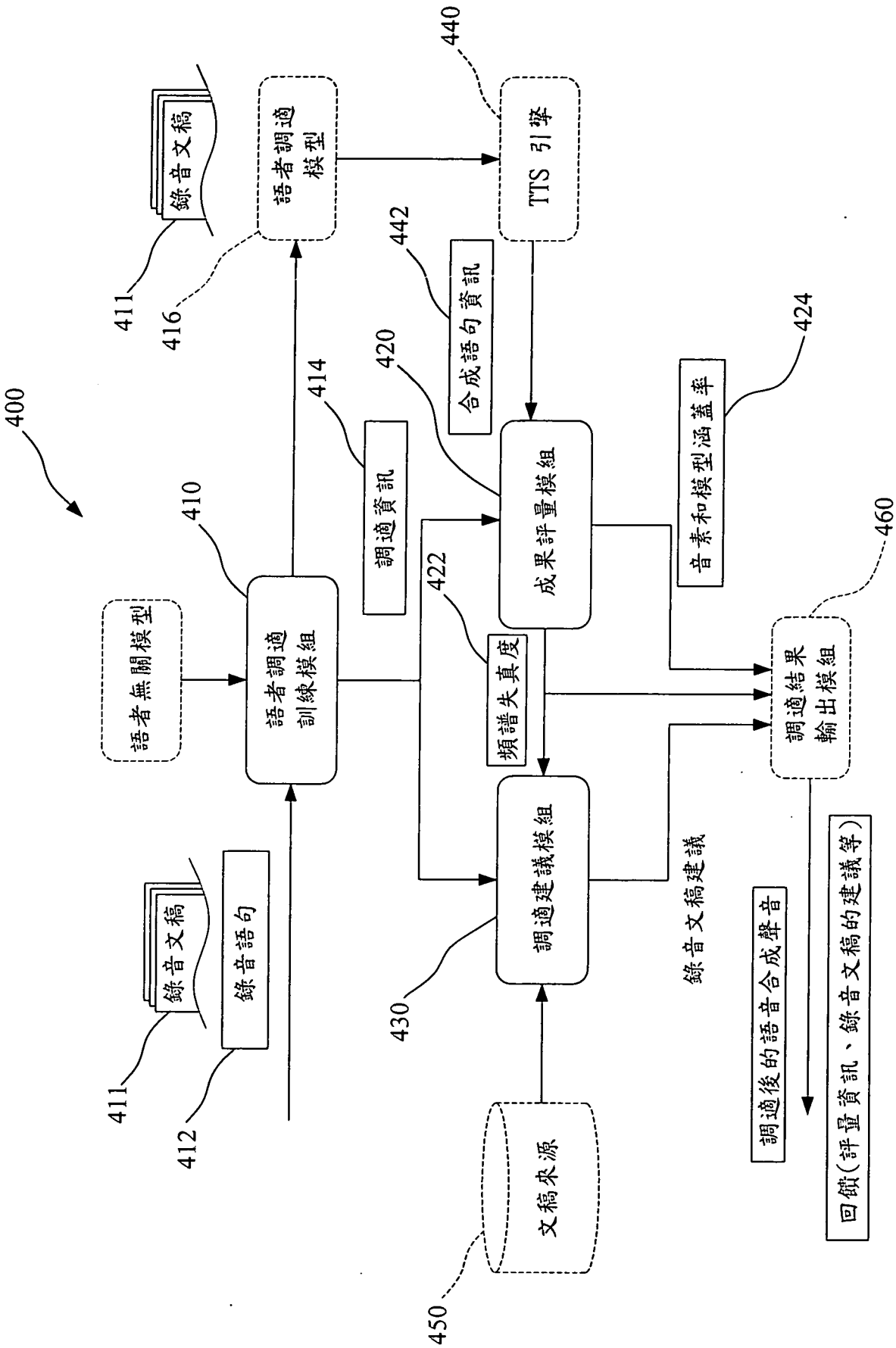
第一圖


210
原始語者語音庫
隱藏式馬可夫模型訓練
原始語者頻譜和基頻的隱藏式馬可夫模型
原始語者說話的隱藏式馬可夫模型
結合高層描述信息的最大相似度線性回歸模型自適應
目標語者語音庫
結合高層描述信息的最大相似度線性回歸模型自適應
原始語者頻譜和基頻的隱藏式馬可夫模型
自動切分
原始語者說話的隱藏式馬可夫模型
目標語者頻譜和基頻參數的生成
目標語者音長參數的生成
參數合成器
目標語者的合成語音

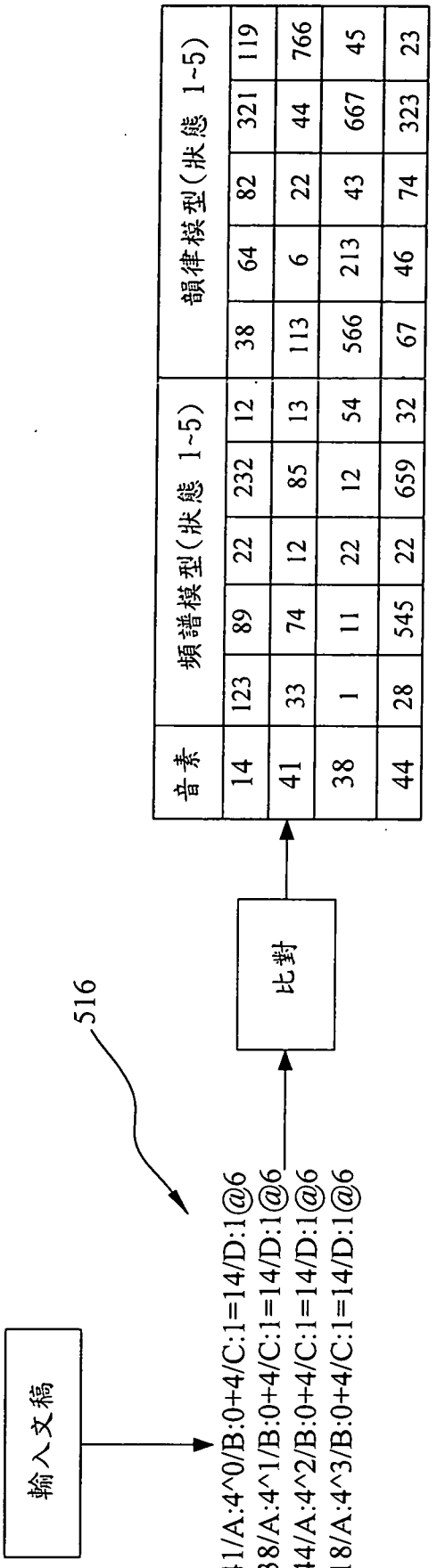
第二圖


原始語者
聲學模型
原始語者
語音資料
目標語者
聲學模型
基於最大相似度準則
的模型自適應演算法
原始語者到目標
語者的轉換矩陣
基於生成參數聽感誤
差最小化的轉換矩陣
迭代更新
目標語者
語音資料

第三圖

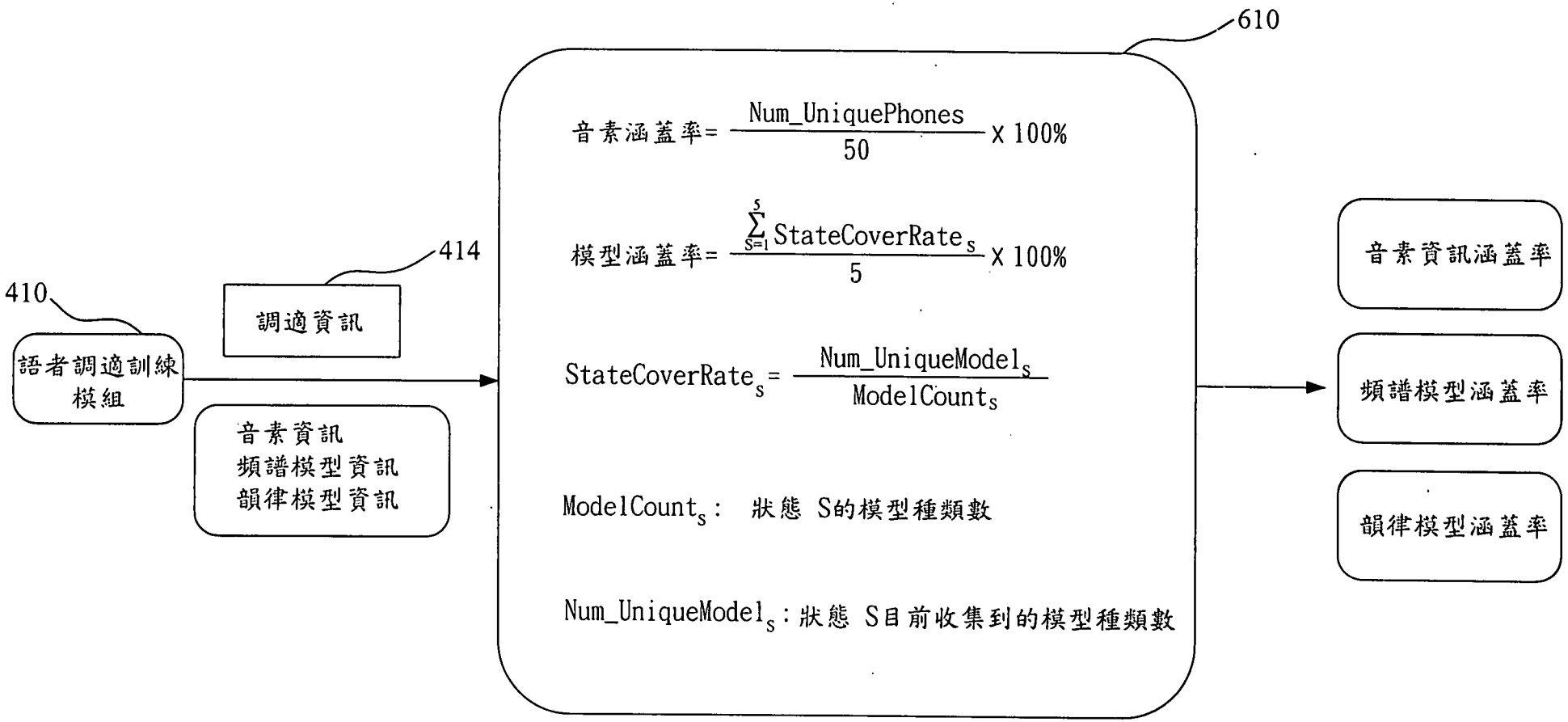

400
語者無關模型
錄音文稿
411
錄音語句
412
語者調適訓練模組
410
語者調適模型
414
錄音文稿
411
416
調適資訊
414
合成語句資訊
442
TTS 引擎
440
成果評量模組
420
音素和模型涵蓋率
424
頻譜失真度
422
調適結果輸出模組
460
調適建議模組
430
錄音文稿建議
文稿來源
450
調適後的語音合成聲音
回饋(評量資訊、錄音文稿的建議等)

第四圖

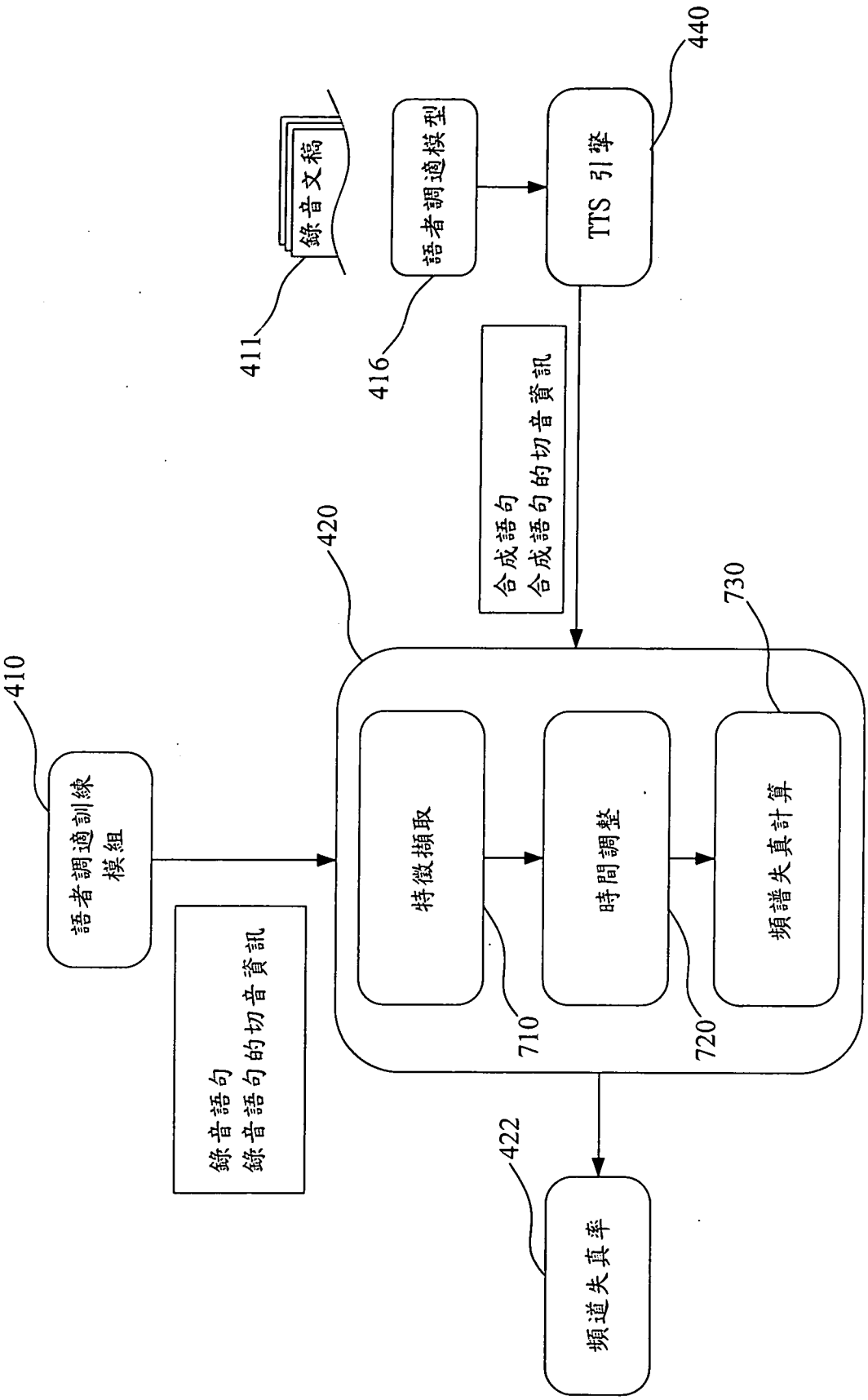


音素	頻譜模型(狀態 1~5)					韻律模型(狀態 1~5)				
14	123	89	22	232	12	38	64	82	321	119
41	33	74	12	85	13	113	6	22	44	766
38	1	11	22	12	54	566	213	43	667	45
44	28	545	22	659	32	67	46	74	323	23

第五圖

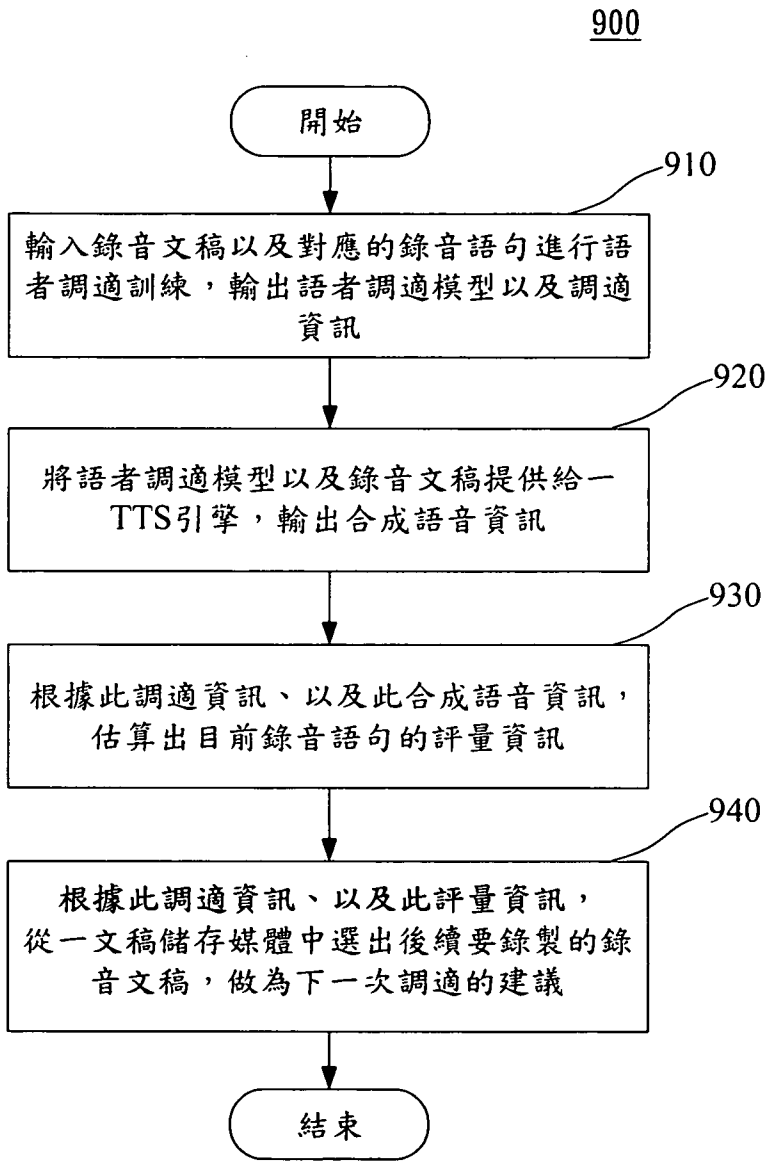

610
音素涵蓋率= Num_UniquePhones / 50 X 100%
模型涵蓋率= (∑ StateCoverRate_s) / 5 X 100%
StateCoverRate_s = Num_UniqueModel_s / ModelCount_s
ModelCount_s： 狀態 S的模型種類數
Num_UniqueModel_s：狀態 S目前收集到的模型種類數
414
調適資訊
410
語者調適訓練模組
音素資訊
頻譜模型資訊
韻律模型資訊
音素資訊涵蓋率
頻譜模型涵蓋率
韻律模型涵蓋率

第六圖

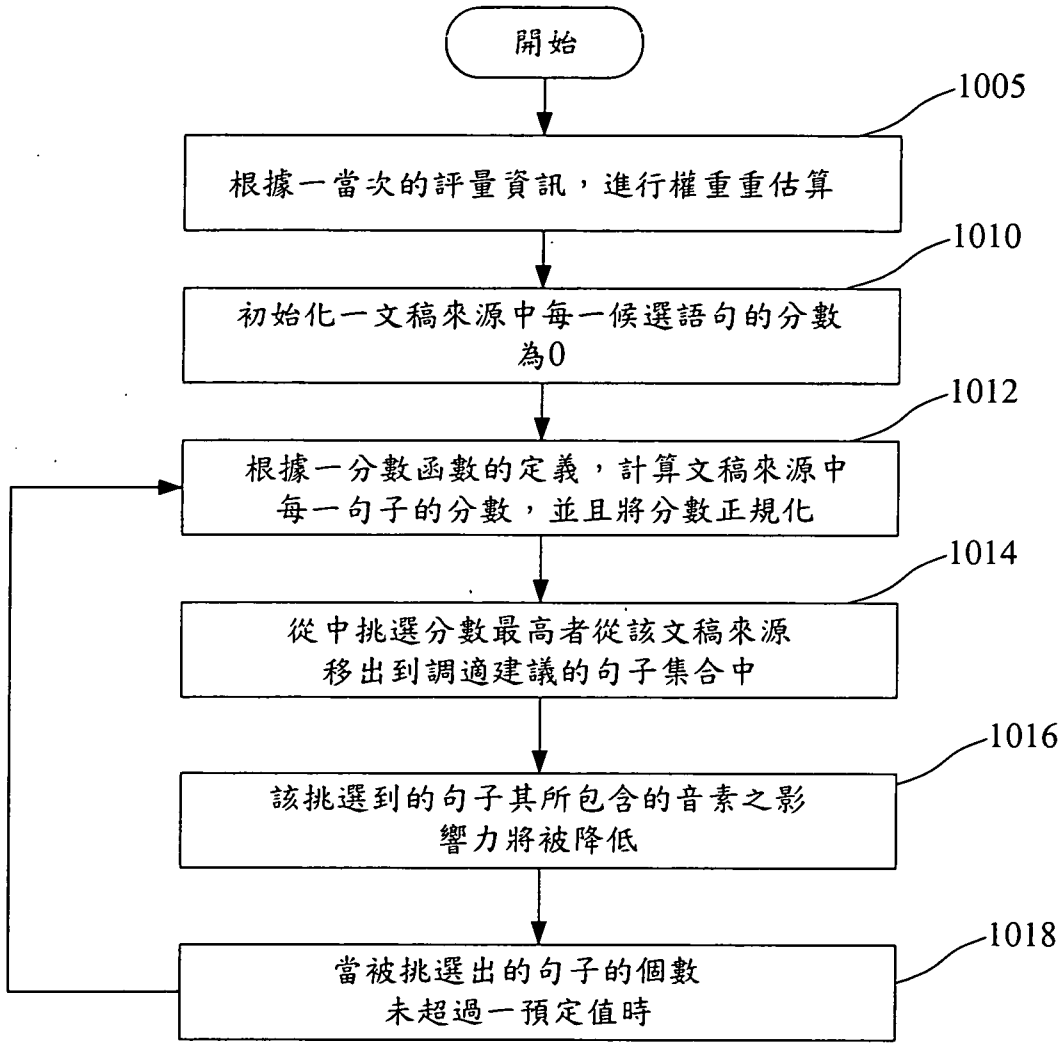

錄音文稿
411
語者調適模型
416
TTS 引擎
440
合成語句
合成語句的切音資訊
420
730
410
語者調適訓練
模組
錄音語句
錄音語句的切音資訊
特徵擷取
時間調整
頻譜失真計算
710
720
422
頻道失真率

第七圖

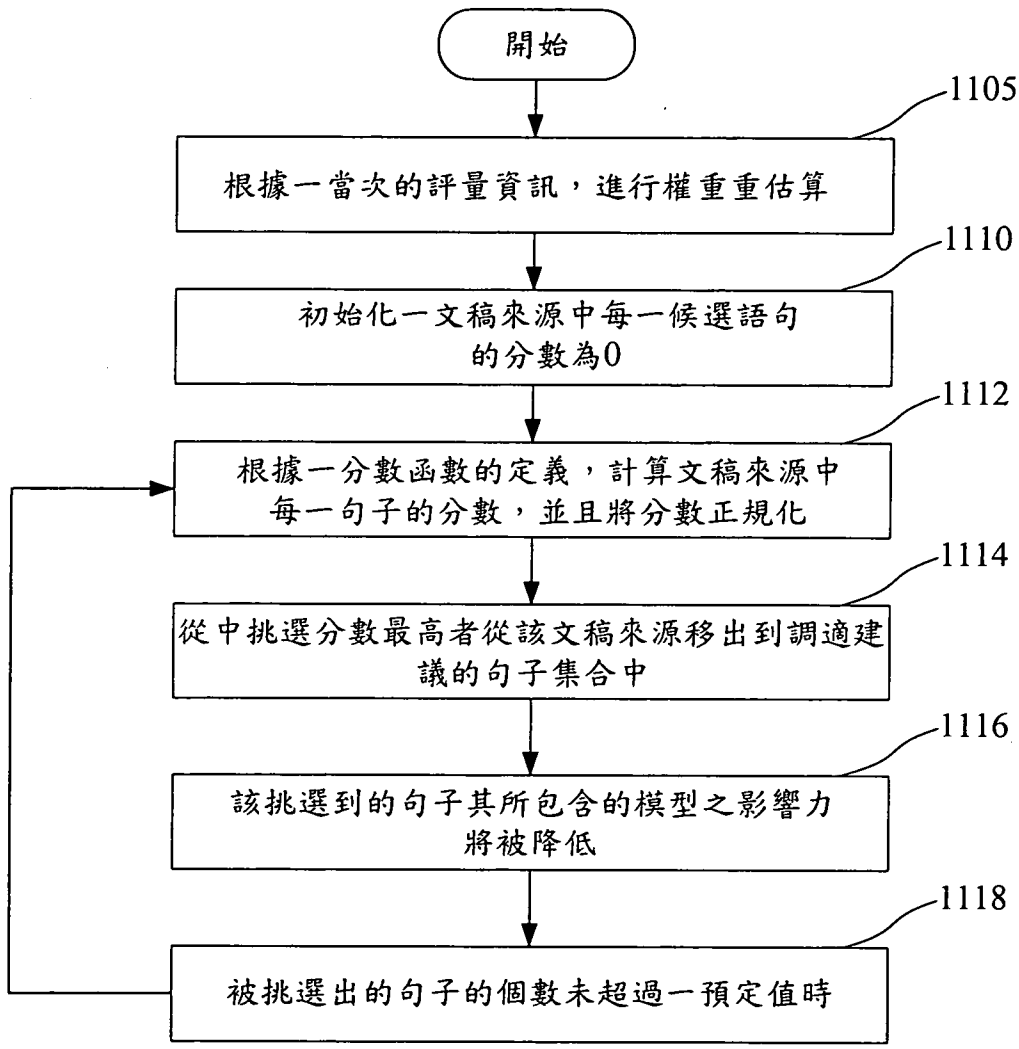
第八圖



第九圖


開始
1005
根據一當次的評量資訊，進行權重重估算
1010
初始化一文稿來源中每一候選語句的分數為0
1012
根據一分數函數的定義，計算文稿來源中每一句子的分數，並且將分數正規化
1014
從中挑選分數最高者從該文稿來源移出到調適建議的句子集合中
1016
該挑選到的句子其所包含的音素之影響力將被降低
1018
當被挑選出的句子的個數未超過一預定值時

第十圖


開始
1105
根據一當次的評量資訊，進行權重重估算
1110
初始化一文稿來源中每一候選語句的分數為0
1112
根據一分數函數的定義，計算文稿來源中每一句子的分數，並且將分數正規化
1114
從中挑選分數最高者從該文稿來源移出到調適建議的句子集合中
1116
該挑選到的句子其所包含的模型之影響力將被降低
1118
被挑選出的句子的個數未超過一預定值時

第十一圖

1200

新權重 = 權重X(1+F)

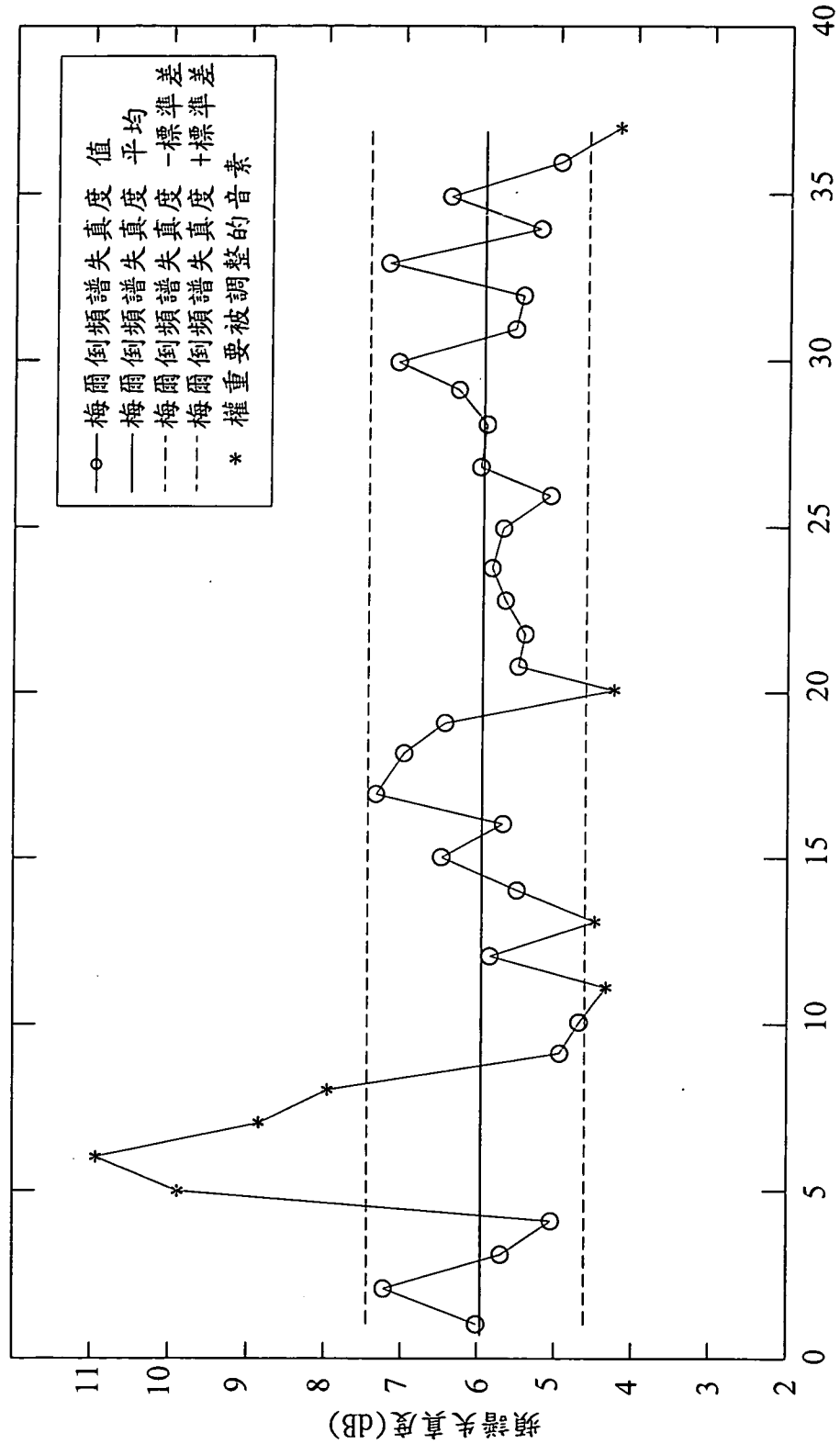
$$F = \frac{\sum_{i=1}^{N} Factor_i}{N}$$

$$Factor_i = \frac{D_i - (D_{mean} + D_{std})}{D_i}, \quad 當 \quad D_i \geq D_{mean} + D_{std})$$

$$Factor_i = -\frac{(D_{mean} - D_{std}) + D_i}{D_{mean} - D_{std}}, \quad 當 \quad 0 \leq D_i \leq D_{mean}) - D_{std}$$

當 $D_{mean} - D_{std} < D_i < D_{mean} < D + D_{std}$，則不計算 $Factor_i$

第十二圖


梅爾倒頻譜失真度 值
梅爾倒頻譜失真度 平均
梅爾倒頻譜失真度 -標準差
梅爾倒頻譜失真度 +標準差
* 權重要被調整的音素
頻譜失真度(dB)
音素
0
5
10
15
20
25
30
35
40
2
3
4
5
6
7
8
9
10
11

第十三圖