



(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 I413105 B

(45)公告日：中華民國 102 (2013) 年 10 月 21 日

(21)申請案號：099146948

(22)申請日：中華民國 99 (2010) 年 12 月 30 日

(51)Int. Cl. : G10L13/00 (2006.01) G10L15/00 (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：李振宇 LI, JEN YU (TW)；涂家章 TU, JIA JANG (TW)；郭志忠 KUO, CHIH CHUNG (TW)

(74)代理人：洪堯順

(56)參考文獻：

US 5271088

US 2004/0030556A1

WO 2005/101905A1

審查人員：涂淑惠

申請專利範圍項數：14 項 圖式數：8 共 0 頁

(54)名稱

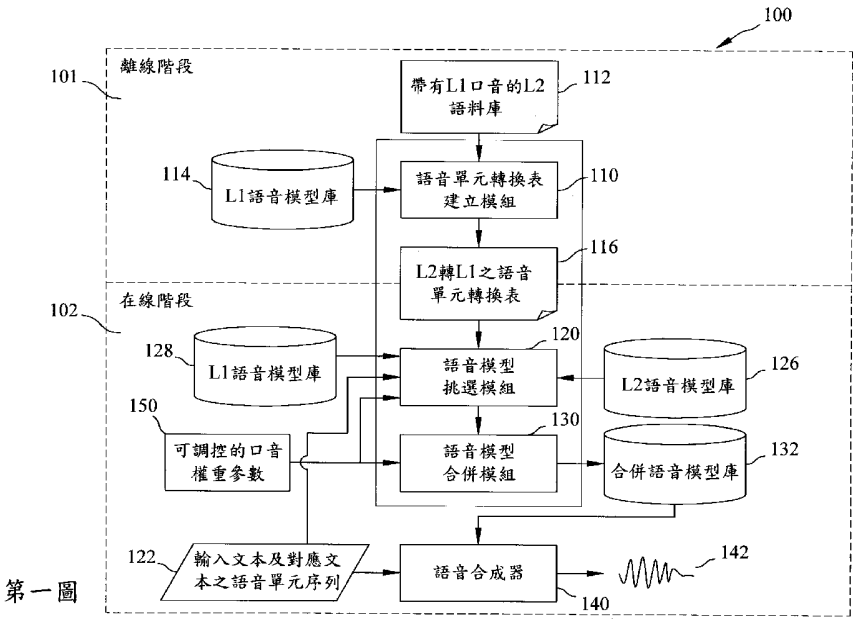
多語言之文字轉語音合成系統與方法

MULTI-LINGUAL TEXT-TO-SPEECH SYNTHESIS SYSTEM AND METHOD

(57)摘要

一種多語言之文字轉語音合成系統將欲合成之文本，藉由一語音模型挑選模組及一語音模型合併模組處理，利用一離線階段得到的一語音單元轉換表，於一在線階段時，此語音模型挑選模組，利用設定的至少一可調控的口音權重參數，選擇要採用的一轉換組合，找出一第二語音模型及一第一語音模型，此語音模型合併模組依照設定的至少一可調控的口音權重參數，將找出的兩語音模型合併成一合併語音模型，處理該轉換組合中所有的轉換後，產生合併語音模型庫，然後利用一語音合成器以及此合併語音模型庫將文本合成帶有第一語言口音之第二語言語音。

In one embodiment of a multi-lingual text-to-speech synthesis system, text to be synthesized is processed by a speech model selection module and a speech model combination module. In an on-line phase, the speech model selection module finds a second-language speech model and a first-language speech model, by using at least a controllable accent weight parameter to choose which transformation composition to adopt from a phonetic unit transformation table obtained in an off-line phase. The speech model combination module combines the two speech models into a new speech model, according to the at least a controllable accent weight parameter. A second-language speech with the first-language accent is further synthesized from the text by a speech synthesizer with the new speech model.



第一圖

- 100 . . . 多語言之文字轉語音合成系統
- 101 . . . 離線階段
- 102 . . . 在線階段
- L1 . . . 第一語言
- L2 . . . 第二語言
- 110 . . . 語音單元轉換表建立模組
- 112 . . . 帶有 L1 口音的 L2 語料庫
- 114 . . . L1 語音模型庫
- 116 . . . L2 轉 L1 之語音單元轉換表
- 120 . . . 語音模型挑選模組
- 150 . . . 可調控的口音權重參數
- 122 . . . 輸入文本及對應文本之 L2 語音單元序列
- 126 . . . L2 語音模型庫
- 128 . . . L1 語音模型庫
- 130 . . . 語音模型合併模組
- 132 . . . 合併語音模型庫
- 140 . . . 語音合成器
- 142 . . . 帶有 L1 口音之 L2 語音

六、發明說明：

【發明所屬之技術領域】

發明所屬之技術領域係關於一種多語言(multi-lingual)之文字轉語音(Text-To-Speech, TTS)合成(synthesis)系統與方法。

【先前技術】

在文章或句子中出現多種語言的交錯使用是很常見的，例如中文與英文夾雜使用。當人們需要將這些文字以語音合成技術轉為聲音時，依據使用的情境來決定如何處理非母語的文字是最佳的。例如有的情境以標準的英文讀出英文單字就已經是最好的，有的情境則略帶母語腔調的方式反而較為自然，例如小說電子書中出現的中英夾雜文句，寫給朋友的電子郵件等。目前多語言之文字轉語音合成系統普遍以多套語言的合成器進行切換，所以合成的語音在不同語言區塊交錯時，常會出現由不同語者發音，或是語句韻律中斷而不順暢等情形。

多語言語音合成的習知文獻有很多。相關的文獻例如美國專利號 US6,141,642 揭示的處理多種語言的文字轉語音裝置與方法(TTS Apparatus and Method for Processing Multiple Languages)，此技術直接以多套語言的合成器來進行切換。

有些專利文獻揭示的技術是直接將非母語音標完全

對應成母語音標，沒有將不同語言的語音模型之間的差異納入考慮。有些專利文獻揭示的技術則合併不同語言的語音模型中相似的部分，保留相異的部分，而沒有考慮口音權重的問題。有些論文如關於基於 HMM 之混合語言 (Mixed-language)，如中文-英文，的語音合成所揭示的技術也是沒有將口音权重納入考慮。

有一篇論文 "Foreign Accents in Synthetic Speech: Development and Evaluation" 是以不同的音標對應的方式來處理口音問題。另兩篇論文 "Polyglot speech prosody control" 及 "Prosody modification on mixed-language speech synthesis" 則處理韻律方面的問題，也沒有處理語音模型的部分。而論文 "New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer" 是以語者模型調適的方式來建立非母語 (non-native language) 的語音模型，但沒有揭示可控制口音的輕重。

【發明內容】

本揭露實施例可提供一種多語言之文字轉語音合成系統與方法。

在一實施例中，所揭露者是關於一種多語言之文字轉語音合成系統。此系統包含一語音模型挑選模組 (speech model selection module)、一語音模型合併模組

(pronunciation)與韻律(prosody)可以在兩種極端範圍中作調整。換句話說，在完全維持其原標準發音至完全改成以母語方式發音之間作調整。以解決目前合成多語言文字時，韻律或發音不自然的問題，並且可依照喜好的程度進行最佳的調整。

第一圖是一種多語言之文字轉語音合成系統的一個範例示意圖，與所揭露的某些實施範例一致。第一圖的範例中，多語言之文字轉語音合成系統 100 包含一語音模型挑選模組 120、一語音模型合併模組 130 及一語音合成器 140。於一在線(on-line)階段 102 時，語音模型挑選模組 120 對輸入文本及對應文本之 L2 語音單元序列 122，在 L2 語音模型庫 126 中找出一第二語言語音單元序列中各語音單元所對應的第二語音模型，再查詢一 L2 轉 L1 之語音單元轉換表 116，並依照設定的一可調控的口音權重參數 150，決定要採用的一轉換組合，選擇一相對應的第一語言語音單元序列，並在 L1 語音模型庫 128 中找出一第一語言語音單元序列中各語音單元所對應的第一語音模型。

語音模型合併模組 130，依照設定的可調控的口音權重參數 150，在 L2 語音模型庫 126 中找出的各語音單元所對應的模型(即第二語音模型)，及 L1 語音模型庫 128 中找出的各語音單元所對應的模型(即第一語音模型)，依據採用一轉換組合，合併成一合併語音模型，

第二圖是一範例示意圖，說明語音單元轉換表建立模組 110 如何產生語音單元轉換表，與所揭露的某些實施範例一致。在離線階段時，如第二圖的範例所示，建構 L2 轉 L1 之語音單元轉換表的流程可包含如下：(1)準備帶有 L1 口音的 L2 語料庫 112，此 L2 語料庫 112 包含有多個音檔 202 以及與音檔相對應的多個語音單元序列 204。(2)從 L2 語料庫 112 中挑選出一個音檔以及與此音檔的內容相對應的一 L2 語音單元序列，將此音檔以 L1 語音模型庫 114 來進行自由音節(free syllable)式語音辨識 212，產生音節辨識結果 214；關於音調(pitch)方面也可採取類似的方式以自由聲調辨識(free tone recognition)的結果作對應，也就是說，也可再包括進行一自由聲調式辨識來產生辨識結果 214，此時結果為具聲調的音節(tonal syllable)。(3)將 L1 語音模型庫 114 產生的音節辨識結果 214，透過音節轉成語音單元 216 處理，轉成一 L1 語音單元序列，(4)將步驟(2)的 L2 語音單元序列及步驟(3)轉成的 L1 語音單元序列利用動態編程(Dynamic Programming, DP) 218 來進行語音單元校準(alignment)，完成動態編程後，即可得到一筆轉換組合。也就是說，利用該動態編程來找出該欲合成之輸入文本之對應語音單元與轉換類型。

重複上述步驟(2)、(3)、(4)便可得到眾多的轉換組合，統計所得到的眾多轉換組合就可完成 L2 轉 L1 之語音單元轉換表 116。此語音單元轉換表可包含三種類型

的轉換，分別為代換(substitution)、插入(insertion)及刪除(deletion)，其中代換是一對一的轉換，插入是一對多的轉換，刪除是多對一的轉換。

舉例說明，假設從帶有 L1(中文)口音的 L2(英文)語料庫 112 中一個音檔為 SARS，其 L2 語音單元序列為 s/a:/r/s(國際音標表示法，語音單元為音素)。而此音檔由 L1 語音模型庫 114 進行自由音節式語音辨識 212 後，產生其音節辨識結果 214，經音節轉成語音單元 216 處理後，L1(中文)語音單元序列例如為“ㄙ/ㄚ/ㄖ/ㄙ(注音表示法，“ㄙ”為空韻母，語音單元為聲母/韻母)”。將 L2 語音單元序列“s/a:/r/s”及 L1 語音單元序列“ㄙ/ㄚ/ㄖ/ㄙ”利用動態編程 218 進行語音單元校準後，例如找到 $s \rightarrow \text{ㄙ}$ ， $a: \rightarrow \text{ㄚ}$ 的代換、 $s \rightarrow \text{ㄙ-ㄙ}$ 的插入ㄙ及 $a:-r \rightarrow \text{ㄚ}$ 的刪除 r 等轉換，此即為得到一筆轉換組合。

利用動態編程 218 進行語音單元校準的方法舉例說明如下。例如使用五個狀態(5-state)的隱馬可夫模型(HMM)來描述一個語音模型，每個狀態的特徵參數假設為梅爾倒頻譜(mel-cepstrum)，維度(dimension)假設為 25 維，特徵參數各維度的數值分布為高斯分布(Gaussian distribution)，以高斯密度函數 $g(\mu, \Sigma)$ 來表示，其中 μ 為平均值向量(維度為 25×1)， Σ 為共變異矩陣(維度為 25×25)，屬於 L1 的表示為 $g_1(\mu_1, \Sigma_1)$ ，屬於 L2 的表示為 $g_2(\mu_2, \Sigma_2)$ 。在動態編程過程中，可利用一種統計學

上計算兩離散機率分布之間的距離的巴特查裡亞距離 (Bhattacharyya distance) 來計算兩語音模型之間的本地距離，作為動態編程中的本地距離。巴特查裡亞距離 b 如公式(1)所示，

$$b = \frac{1}{8}(\mu_2 - \mu_1)^T \left[\frac{\Sigma_1 + \Sigma_2}{2} \right]^{-1} (\mu_2 - \mu_1) + \frac{1}{2} \ln \frac{|\Sigma_1 + \Sigma_2|/2}{|\Sigma_1|^{1/2} |\Sigma_2|^{1/2}} \quad (1)$$

依照此公式可以計算出 L1 語音模型的第 i 狀態 ($1 \leq i \leq 5$) 和 L2 語音模型第 i 狀態的距離，如上述使用五個狀態的 HMM，則把五個狀態的巴特查裡亞距離加總後，即可得到本地距離。以上述 SARS 的例子，第五圖進一步說明動態編程 218 的細節，其中 X 軸為 L1 語音單元序列，Y 軸為 L2 語音單元序列。

第五圖中，利用動態編程可以找出由起點(0,0)走到終點(5,5)的最短路徑，也就找到了 L1 語音單元序列與 L2 語音單元序列的轉換組合。找最短路徑的方法就是要找有最小累加距離的路徑。累加距離 $D(i,j)$ 的意義為，由起點(0,0)走到(i,j)這個點的所累積的總距離， i 是 X 軸座標， j 是 Y 軸座標。累加距離 $D(i,j)$ 的算法如下列公式所示：

$$D(i, j) = b(i, j) + \min \begin{cases} \omega_1 \cdot D(i-2, j-1) \\ \omega_2 \cdot D(i-1, j-1) \\ \omega_3 \cdot D(i-1, j-2) \end{cases}$$

其中， $b(i,j)$ 為點 (i,j) 的兩個語音模型之本地距離，在出發點的 $D(0,0)=b(0,0)$ 。本揭露實施例中以巴特查裡亞距離來當成本地距離，而 ω_1 、 ω_2 及 ω_3 分別為插入、刪除、及代換的權重，可以利用修改權重來調整插入、刪除、及代換發生時，對於累加距離影響的多寡， ω 越大影響越大。

第五圖中，線條 511-513 說明點 (i,j) 只能由這 3 條路徑過來，其他的路徑都不可以走，也就是限制由某點只能有 3 條路徑可以移動到下一點，其意義是只容許代換(路徑 512)、刪除 1 個語音單元(路徑 511)、插入 1 個語音單元(路徑 513)，共三種可容許的轉換類型。因為有了這項限制，在動態編程過程中，就有四條虛線範圍成為全域限制(global constraint)，因為超過虛線範圍的路徑都無法由起點走到終點，因此只要計算四條虛線範圍內所有的點，就可以找到一條最短路徑。首先，在此全域限制的範圍內，先計算各點的本地距離，接著再計算由 $(0,0)$ 走到 $(5,5)$ 各種可能路徑的累加距離，再找出最小值即可。在此例中假設找到的最短路徑是由箭頭實線所連接的路徑。

接著說明語音單元轉換表，L2 轉 L1 之語音單元轉換表的例子如第三圖所示。假設承上述帶有 L1(中文)口音的 L2(英文)語料庫 112 內總共有 10 筆內容為 SARS 的音檔，重複上述語音辨識、音節轉成語音單元、動態

編程步驟後，有 8 筆轉換組合如同前述結果($s \rightarrow \text{ㄙ}$ 、 $a:r \rightarrow \text{ㄚ}$ 、 $s \rightarrow \text{ㄙ-ㄖ}$)，而有 2 筆轉換組合為 $s \rightarrow \text{ㄙ}$ 、 $a:r \rightarrow \text{ㄚ}$ 、 $r \rightarrow \text{ㄖ}$ 、 $s \rightarrow \text{ㄙ-ㄖ}$ ，則統計所有的轉換組合後就可完成 L2 轉 L1 之語音單元轉換表的例子 300。在第三圖中，L2(英文)轉 L1(中文)之語音單元轉換表的例子 300 有兩種轉換組合，出現機率分別為 0.8 及 0.2。

接下來進一步說明在線階段 102 時，語音模型挑選模組、語音模型合併模組、及語音合成器的運作。語音模型挑選模組，依照設定的可調控的口音權重參數 150，可從語音單元轉換表中挑選所使用的轉換組合，以控制 L2 受到 L1 影響的程度。例如當設定的口音權重參數之值越小時，代表口音比較輕，就選擇出現機率越高的轉換組合，代表這種口音比較容易出現，易為大眾認知。反之，口音權重參數之值越大時，選擇出現機率越低的轉換組合，代表這種口音較少見、奇怪，也就代表口音比較重。例如第四圖所示，假設以 0.5 當作分界，當設定口音權重值 $w=0.4$ 時 ($w<0.5$)，選擇 L2 轉 L1 之語音單元轉換表的例子 300 中出現機率 0.8 的轉換組合；當設定口音權重值 $w=0.6$ ($w>0.5$)，選擇出現機率 0.2 的轉換組合。

參考第六圖的運作範例，語音模型挑選模組 120 利用 L2 轉 L1 之語音單元轉換表 116 以及設定的可調控的口音權重參數 150，根據至少含有 L2 的輸入文本及對

應文本之 L2 語音單元序列 122，進行模型挑選(model selection)，在 L2 語音模型庫 126 中找出各語音單元的模型，再查詢 L2 轉 L1 之語音單元轉換表 116，並依照設定的一可調控的口音權重參數 150，決定要採用的轉換組合，選擇一相對應的第一語言語音單元序列，並在 L1 語音模型庫 128 中找出各語音單元的模型。假設各語音模型如前述以五個狀態(5-state)的隱馬可夫模型(HMM)，例如為第一語音模型 614，其第 i 狀態($1 \leq i \leq 5$)之梅爾倒頻譜各維度的數值分布為 $g_1(\mu_1, \Sigma_1)$ ，及第二語音模型 616，其第 i 狀態之梅爾倒頻譜各維度的數值分布為 $g_2(\mu_2, \Sigma_2)$ 。語音模型合併模組 130 例如可使用下列公式(2)進行模型合併，將第一語音模型 614 及第二語音模型 616 合併為合併語音模型 622，此合併語音模型其第 i 狀態之梅爾倒頻譜各維度的數值分布表示為 $g_{new}(\mu_{new}, \Sigma_{new})$ 。

$$\begin{aligned}\mu_{new} &= w * \mu_1 + (1-w) * \mu_2 \\ \Sigma_{new} &= w * (\Sigma_1 + (\mu_1 - \mu_{new})^2) + (1-w) * (\Sigma_2 + (\mu_2 - \mu_{new})^2)\end{aligned}\quad (2)$$

； w 為設定的一可調控的口音權重參數 150，合理的數值範圍為 $0 \leq w \leq 1$ 。

如上述使用五個狀態的 HMM，則把五個狀態的 $g_{new}(\mu_{new}, \Sigma_{new})$ 都個別計算出來後，即可得到合併語音模型 622。例如 s->ㄥ的代換轉換，將 L1 語音單元 HMM (ㄥ)與 L2 語音單元 HMM (s)，以公式(2)計算得到新的

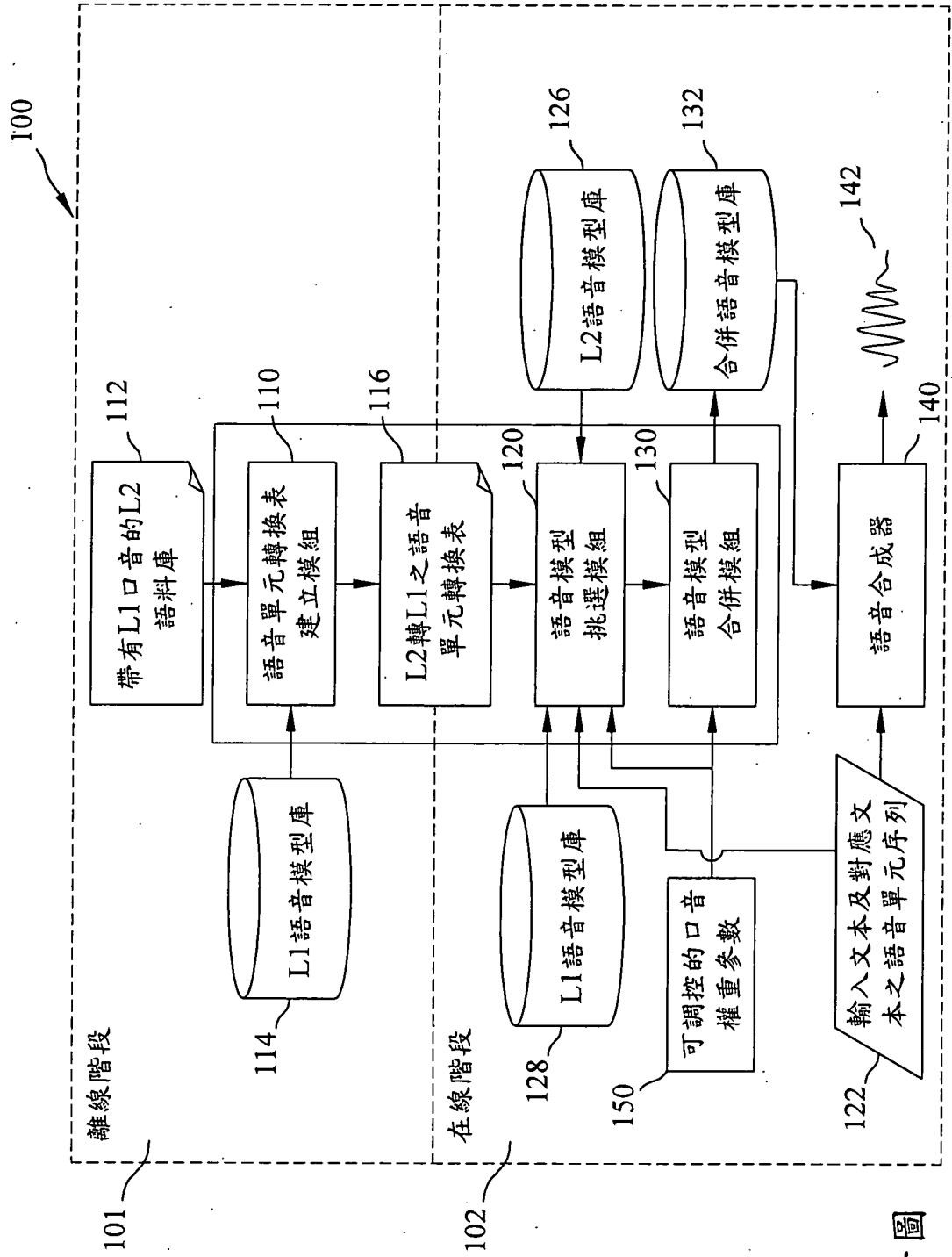
HMM (帶有 $\perp$ 之口音的 s)。而例如 $a:-r \rightarrow \gamma$ 的刪除轉換，則分別以 $a:- \rightarrow \gamma$ 和 $r \rightarrow$ 靜音(silence)方式來完成。同理， $s \rightarrow \perp$ - η 的插入轉換分別以 $s \rightarrow \perp$ 和 靜音 $\rightarrow \eta$ 的方式來完成。也就是說，當轉換是代換的類型時，可使用與 L2 對應之 L1 模型；當轉換是插入或刪除的類型時，使用靜音模型(silence model)當作對應模型。將各語音單元合併語音模型 622 全部計算完畢後，可得到一合併語音模型庫 132。此合併語音模型庫 132 再提供給語音合成器 140，合成為帶有 L1 口音的一 L2 語音 142。

上述例子說明 HMM 之聲學參數合併方式，在韻律參數方面，即音長(duration)和音調(pitch)，同樣也可利用公式(2)來得到新的韻律參數。對於音長參數的合併，可依照 L1 與 L2 之語音模型，找出各語音單元 HMM 之音長參數後，再利用公式(2)按照口音權重參數計算出合併後的新的音長參數(插入/刪除轉換所對應之靜音模型音長為 0)。對於音調參數的合併，代換轉換同樣也可利用公式(2)按照口音權重參數計算出合併後的新的音調參數，刪除轉換直接採用原語音單元之音調參數不變，例如 $a:-r \rightarrow \gamma$ 的刪除轉換，原 r 之音調參數不變。插入轉換則以插入之語音單元音調模型與最接近的有聲(voiced)語音單元之音調參數，利用公式(2)進行合併，例如 $s \rightarrow \perp$ - η 的插入轉換，以 η 之音調參數與有聲語音單元 $a:-$ 之音調參數進行合併(因為 s 為無聲語音單元，無聲調數值可供合併)。

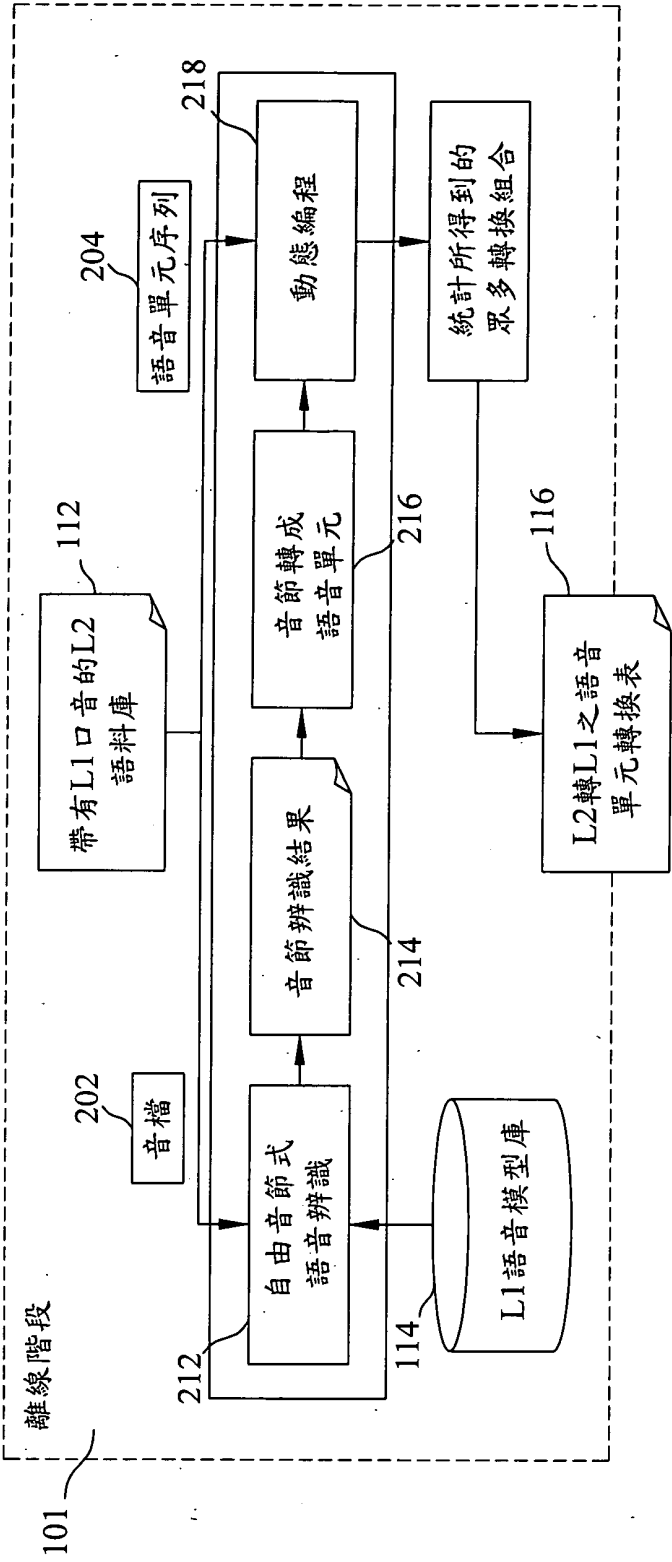
也就是說，語音模型合併模組 130 將找出的第二語言語音單元序列中各第二語言語音單元所對應的語音模型和第一語言語音模型庫中找出第一語言語音單元序列中各第一語言語音單元所對應的語音模型，依照轉換組合的對應關係，根據設定的口音權重參數合併成一合併語音模型，以及將各語音單元之合併語音模型之多個狀態之梅爾倒頻譜各維度的數值分布全部計算完畢後，得到此合併語音模型庫。

承上述，第七圖是一範例流程圖，說明一種多語言之文字轉語音合成方法的運作，與所揭露的某些實施範例一致。此多語言之文字轉語音合成方法係執行於一電腦系統上。此電腦系統備有一記憶體裝置，用來儲存多種語言語音模型，至少包括前述所找出的第一與第二語言語音模型。第七圖的範例中，首先，準備帶有第一語言口音的一第二語言語料庫及一第一語言語音模型庫，來建構一第二語言轉第一語言之語音單元轉換表，如步驟 710 所示。然後，對欲合成之輸入文本，及對應輸入文本之一第二語言語音單元序列，在一第二語言語音模型庫中找出第二語言語音單元序列中各語音單元所對應的第二語音模型後，再查詢此語音單元轉換表，並依照設定的一可調控的口音權重參數，決定要採用的一轉換組合，決定出一相對應的第一語言語音單元序列，並在第一語言語音模型庫中找出一第一語言語音單

八、圖式



第一圖

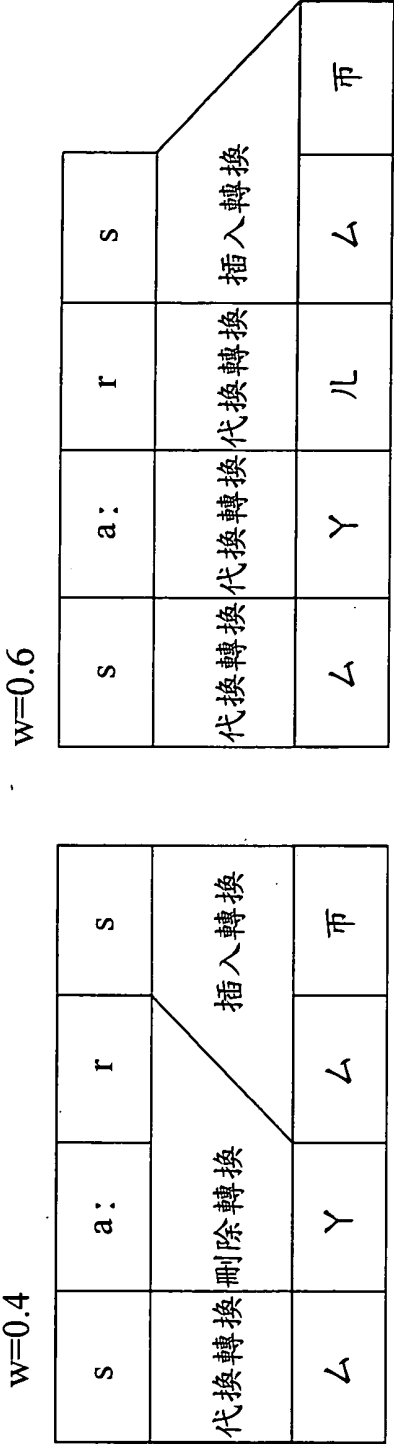


第二圖

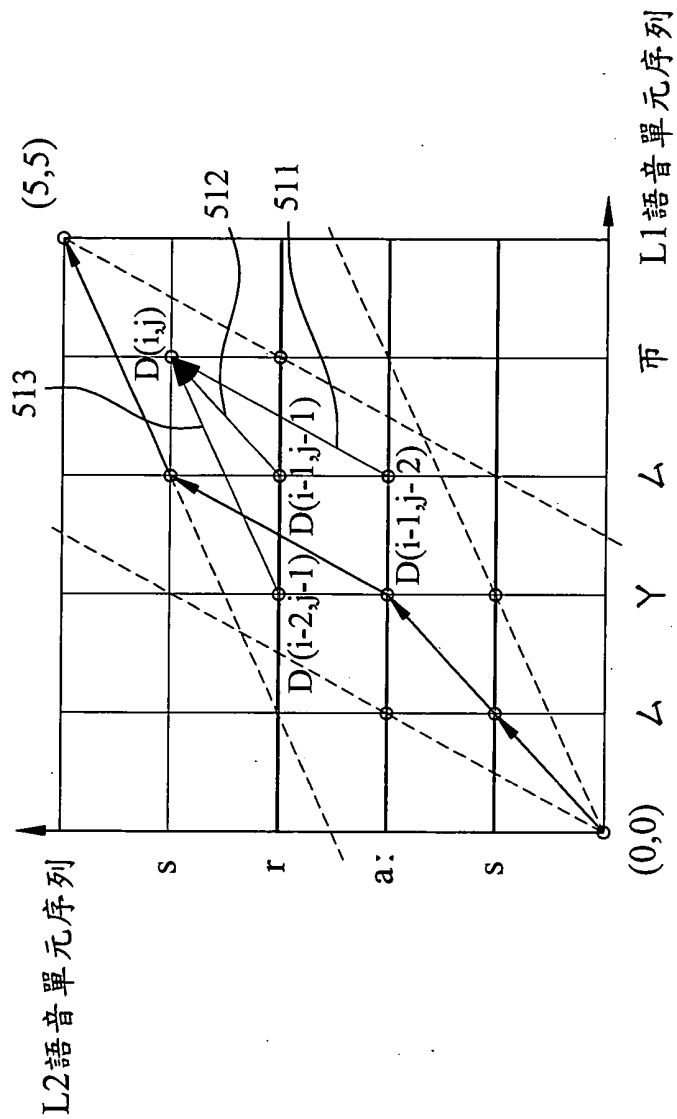
300

L2轉L1之語音單元轉換表						
輸入文本	英文語音單元序列	中文語音單元序列	出現機率	轉換組合		
SARS	sa:rs	ㄙㄩ.ㄌㄚ	0.8	s->ㄌ	a:r->ㄩ	s->ㄌㄚ
		ㄌㄩ.ㄌㄚ.ㄌㄚ	0.2	s->ㄌ	a:->ㄩ	r->ㄌ s->ㄌㄚ

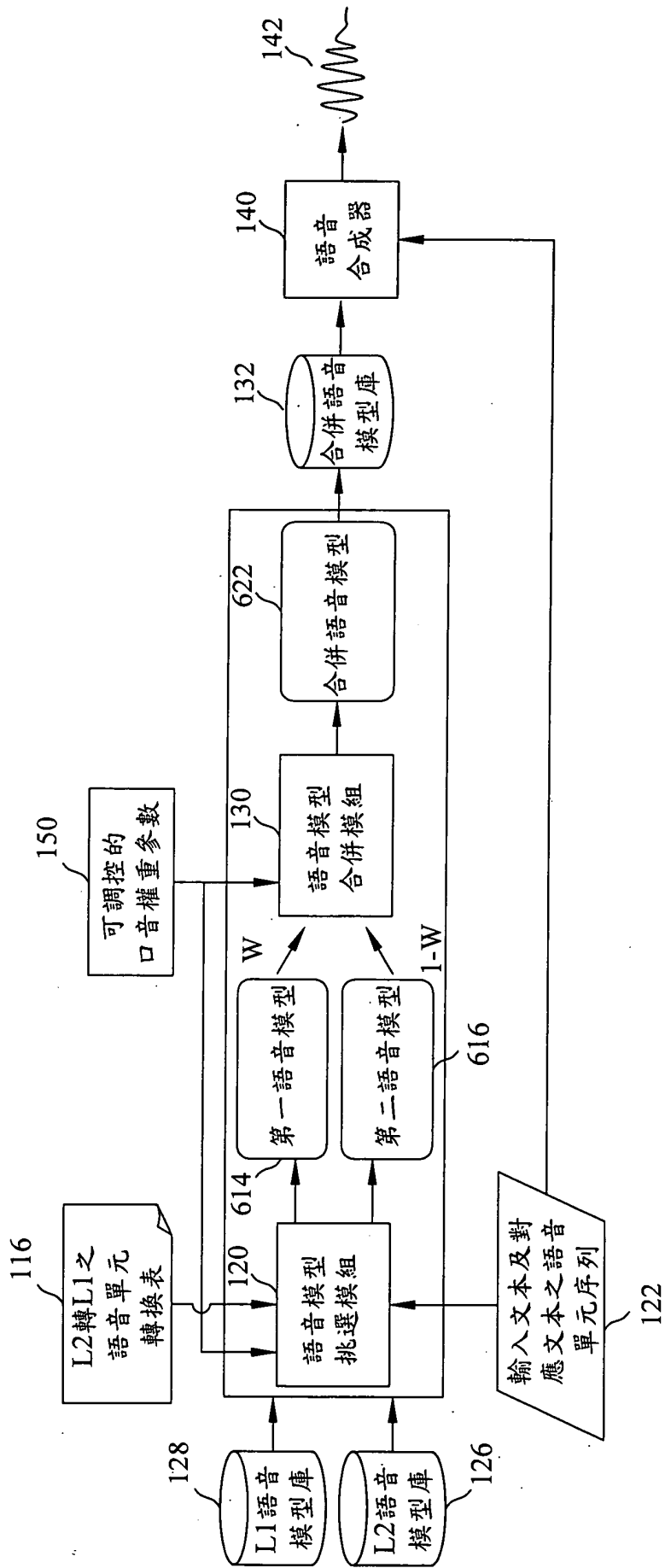
第三圖



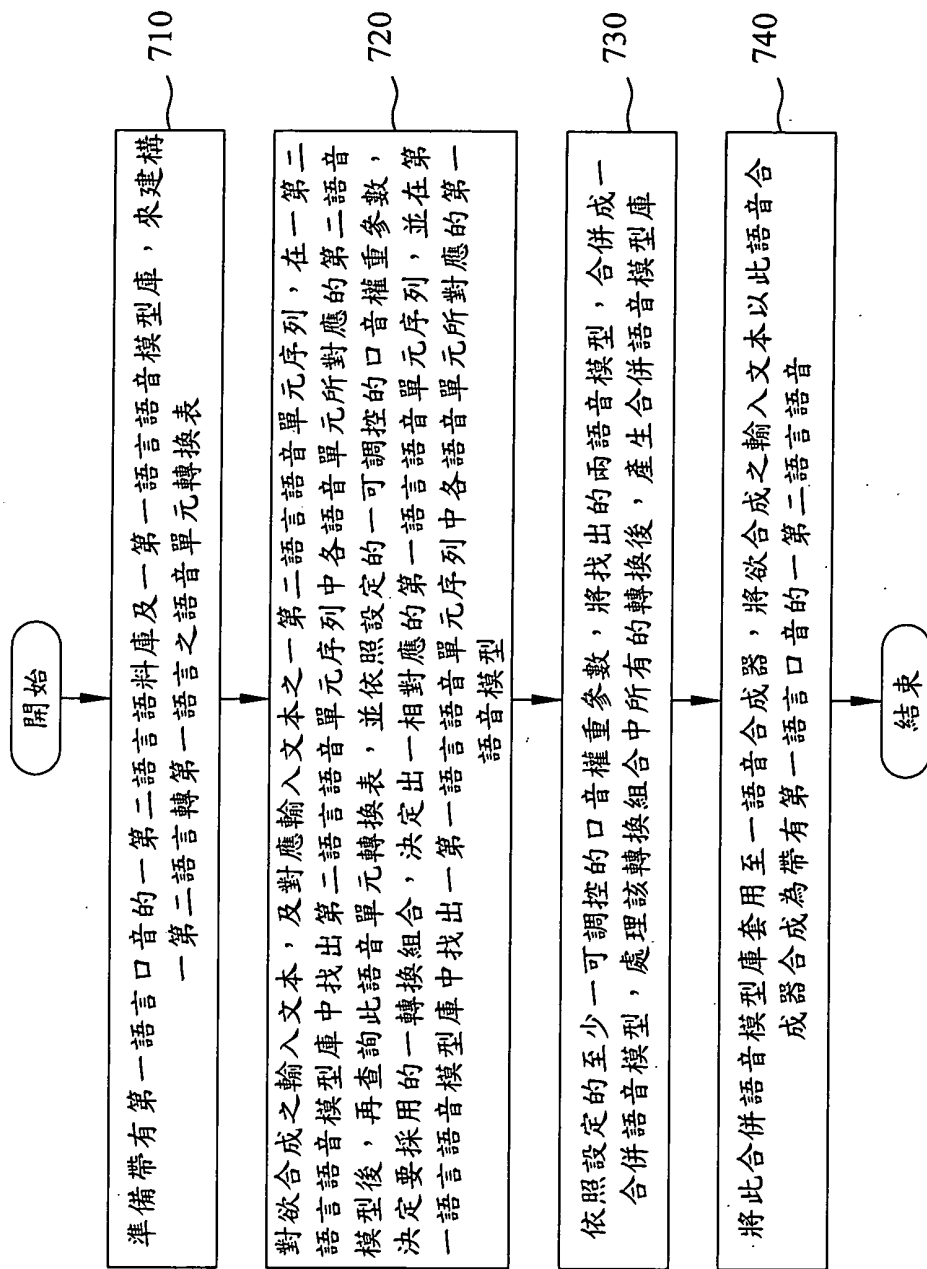
第四圖



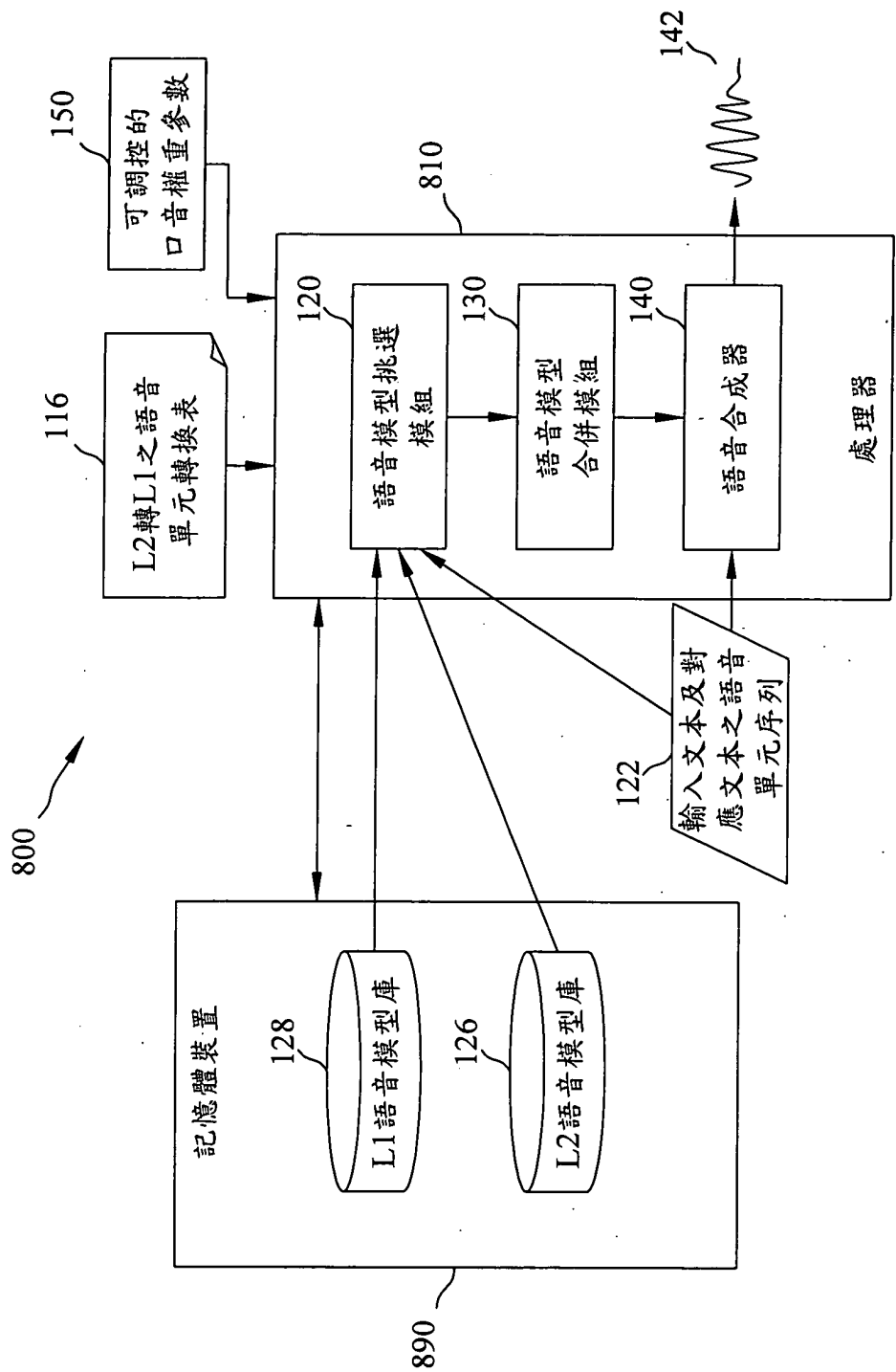
第五圖



第六圖



第七圖



第八圖

(speech model combination module) 及一語音合成器 (speech synthesizer)。此語音模型挑選模組對欲合成之含有第二語言之文本(text)及對應輸入文本之第二語言語音單元序列(phonetic unit sequence)，在一第二語言語音模型庫中找出該第二語言語音單元序列中各語音單元所對應的第二語音模型，再查詢一第二語言轉第一語言之語音單元轉換表，並利用設定的至少一可調控的口音權重參數，決定要採用一轉換組合，選擇出一相對應的第一語言語音單元序列，並在此第一語言語音模型庫中找出該第一語言語音單元序列中各語音單元所對應的第一語音模型。此語音模型合併模組將找出的第二與第一語音模型，依照設定的至少一可調控的口音權重參數，合併成一合併語音模型，處理該轉換組合中所有的轉換後，產生合併語音模型庫。此合併語音模型庫再套用至此語音合成器，以將輸入之文本合成為帶有第一語言口音之第二語言語音(L1-accent L2 speech)。

在另一實施例中，所揭露者是關於一種多語言之文字轉語音合成系統，此多語言之文字轉語音合成系統係執行於一電腦系統中，此電腦系統備有一記憶體裝置，用來儲存多種語言語音模型庫，至少包括一第一與一第二語言語音模型庫。此多語言之文字轉語音合成系統可包含一處理器，此處理器備有一語音模型挑選模組、一語音模型合併模組、及一語音合成器。其中，於一離線階段時，建立此語音單元轉換表，以提供給此處理器使

用。此語音模型挑選模組對欲合成之文本及對應輸入文本之一第二語言語音單元序列，在一第二語言語音模型庫中找出該第二語言語音單元序列中各語音單元所對應的第二語音模型，再查詢一第二語言轉第一語言之語音單元轉換表，並依照設定的至少一可調控的口音權重參數，決定要採用的一轉換組合，選擇出一相對應的第一語言語音單元序列，並在一第一語言語音模型庫中找出該第一語言語音單元序列中各語音單元所對應的第一語音模型。此語音模型合併模組將找出的第二與第一語音模型，依照設定的至少一可調控的口音權重參數，合併成一合併語音模型，處理該轉換組合中所有的轉換後，產生合併語音模型庫。此合併語音模型庫再套用至此語音合成器，以將輸入之文本合成為帶有第一語言口音之第二語言語音。

在又一實施例中，所揭露者是關於一種多語言之文字轉語音合成方法。此方法係執行於一電腦系統中，此電腦系統備有一記憶體裝置，用來儲存多種語言語音模型庫，至少包括一第一與一第二語言語音模型庫。此方法包含：對欲合成之輸入文本，利用一輸入之第二語言語音單元序列，在一第二語言語音模型庫中找出一第二語言語音單元序列中各語音單元所對應的第二語音模型後，再查詢一第二語言轉第一語言之語音單元轉換表，並依照設定的至少一可調控的口音權重參數，決定要採用的一轉換組合，選擇一相對應的第一語言語音單元序

列，並在一第一語言語音模型庫中找出一第一語言語音單元序列中各語音單元所對應的第一語音模型；依照設定的至少一可調控的口音權重參數，將找出的第二與第一語音模型，合併成一合併語音模型，處理該轉換組合中所有的轉換後，產生合併語音模型庫；以及將此合併語音模型庫套用至一語音合成器，並將欲合成之輸入文本以此語音合成器合成為帶有第一語言口音的一第二語言語音。

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】

本揭露實施例欲提供一種音韻模型統合之多語言文字轉語音合成技術，並且建立一種調整機制來調整非母語語句所帶的母語口音之權重，讓合成的語音在跨不同語言區塊時，能因應使用的情境來決定如何處理非母語的文字。讓合成的語音在跨不同語言區塊時韻律更加自然，發音腔調也更符合多數人所習慣的方式。換言之，本揭露實施例將非母語，即第二語言(second language, L2)，的文字轉換成帶有母語口音，即第一語言(first language1, L1)口音，之 L2 語音。

本揭露實施例是可用參數調整語音單元序列的對應以及語音模型的合併，來使非母語文字的發音

處理該轉換組合中所有的轉換後，產生合併語音模型庫 132。此合併語音模型庫 132 再套用至語音合成器 140，合成為 L1 語音及帶有 L1 口音的一 L2 語音 142。

多語言之文字轉語音合成系統 100 可再包括一語音單元轉換表建立模組 110，於一離線(off-line)階段 101 時，語音單元轉換表建立模組 110 根據帶有 L1 口音的一 L2 語料庫 112 及一 L1 語音模型庫 114，產生 L2 轉 L1 之語音單元轉換表 116。

在上述中，L1 語音模型庫 114 是供語音單元轉換表建立模組 110 所使用，而 L1 語音模型庫 128 則供語音模型合併模組 130 所使用，兩語音模型庫 114 及 128 可以採用相同的特徵參數，也可以採用不同的特徵參數，但 L2 語音模型庫 126 採用的參數與 L1 語音模型庫 128 是採用相同的特徵參數。

欲合成之輸入文本 122 可以是同時包含 L1 以及 L2 的文本，例如中英夾雜的句子：他今天感覺很 high、Cindy 昨天 mail 給我、這件衣服是 M 號的。此時 L1 為中文語言，L2 為英語，而合成語音在 L1 的部分維持正常發音不變，L2 的部分則合成帶有 L1 口音的 L2 語音。輸入文本 122 也可以是只包含 L2 的文本，例如合成帶有台語口音的中文語言，此時 L1 為台語，L2 為中文語言。也就是說，欲合成之輸入文本 122 至少含有 L2 的文本。

元序列中各語音單元所對應的第一語音模型，如步驟 720 所示。依照設定的至少一可調控的口音權重參數，將找出的兩語音模型，合併成一合併語音模型，處理該轉換組合中所有的轉換後，產生合併語音模型庫，如步驟 730 所示。最後，將此合併語音模型庫套用至一語音合成器，將欲合成之輸入文本以此語音合成器合成為帶有第一語言口音的一第二語言語音，如步驟 740 所示。

上述多語言之文字轉語音合成方法的運作可簡化為步驟 720~步驟 740。而第二語言轉第一語言之語音單元轉換表可在一離線階段時建構，也可以有其他多種建構方式。本揭露之文字轉語音合成方法的實施範例可於線上階段時，再查詢一已建構好的第二語言轉第一語言之語音單元轉換表即可。

各步驟之實施細節，例如步驟 710 中建構一第二語言轉第一語言之語音單元轉換表、步驟 720 中依照設定的一可調控的口音權重參數，決定要採用的轉換組合及找出兩語音模型、步驟 730 中依照設定的至少一可調控的口音權重參數，將找出的兩語音模型，合併成合併語音模型等，如同上述所載，不再重述。

本揭露實施之多語言之文字轉語音合成系統也可執行於一電腦系統上，如第八圖的實施例所示。此電腦系統(未示於圖示)備有一記憶體裝置(memory

device)890，用來儲存多種語言語音模型，至少包括前述所採用的 L1 語音模型庫 128 與 L2 語音模型庫 126，多語言之文字轉語音合成系統 800 可包含一處理器 810。處理器 810 裡可備有語音模型挑選模組 120、語音模型合併模組 130 及語音合成器 140，來執行這些模組之上述功能。可於一離線階段時，先建立一語音單元轉換表及設定至少一可調控的口音權重參數，以提供給語音模型挑選模組 120、語音模型合併模組 130 使用。如何建立此語音單元轉換表，如同上述所載，不再重述。此語音單元轉換表可在離線階段時，由此電腦系統或其他電腦系統建立。

承上述，本揭露實施例可提供一種可調控式之多語言文字轉語音合成系統與方法，可用參數調整語音單元的對應以及語音模型的合併，可使得合成的語音在跨不同語言區塊時，使得第二語言詞彙的發音與韻律，可以在完全維持其原標準發音，到完全以第一語言方式發音的兩種極端範圍中作調整。可應用的情境例如有聲電子書、家用機器人、數位教學等，能使電子書裡多語言夾雜的對白呈現多角色語者、能使機器人增加娛樂效果、能使數位教學提供可程式化的語言教學等。

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】

第一圖是一種多語言之文字轉語音合成系統的一個範例示意圖，與所揭露的某些實施範例一致。

第二圖是一範例示意圖，說明語音單元轉換表建立模組如何產生語音單元轉換表，與所揭露的某些實施範例一致。

第三圖是 L2 轉 L1 之語音單元轉換表的一個例子，與所揭露的某些實施範例一致。

第四圖是一範例示意圖，說明依設定的權重值來選擇 L2 轉 L1 之語音單元轉換表中的轉換組合，與所揭露的某些實施範例一致。

第五圖說明動態編程的細節，與所揭露的某些實施範例一致。

第六圖是一範例示意圖，說明在線階段時，各模組的運作，與所揭露的某些實施範例一致。

第七圖是一範例流程圖，說明一種多語言之文字轉語音合成方法的運作，與所揭露的某些實施範例一致。

第八圖是多語言之文字轉語音合成系統執行於一電腦系統中的一範例示意圖，與所揭露的某些實施範例一致。

【主要元件符號說明】

100 多語言之文字轉語音合成系統

101 離線階段

102 在線階段

L1 第一語言

L2 第二語言

- 110 語音單元轉換表建立模組 112 帶有 L1 口音的 L2 語料庫
- 114 L1 語音模型庫 116 L2 轉 L1 之語音單元轉換表
- 120 語音模型挑選模組
- 122 輸入文本及對應文本之 L2 語音單元序列
- 126 L2 語音模型庫 128 L1 語音模型庫
- 130 語音模型合併模組 132 合併語音模型庫
- 140 語音合成器 142 帶有 L1 口音之 L2 語音
- 150 可調控的口音權重參數
-
- 202 音檔 204 語音單元序列
- 212 自由音節式語音辨識 214 音節辨識結果
- 216 音節轉成語音單元 218 動態編程
-
- 300 L2 轉 L1 之語音單元轉換表的例子
-
- 511-513 3 條路徑
-
- 614 第一語言模型 616 第二語言模型
- 622 合併語音模型
-
- 710 準備帶有第一語言口音的一第二語言語料庫及一第一語言語音模型庫，來建構一第二語言轉第一語言之語音單元轉換表
- 720 對欲合成之輸入文本，及對應輸入文本之一第二語言語音單元序列，在一第二語言語音模型庫中找出第二語言語音單元序列中各語音單元所對應的第二語音模型後，再查詢此語音單元轉換表，並依照設定的一可調控的口音權重參數，決定要採用的一轉換組合，決定出一相對應的第一語言語音單元

序列，並在第一語言語音模型庫中找出一第一語言語音單元
序列中各語音單元所對應的第一語音模型

730 依照設定的至少一可調控的口音權重參數，將找出的兩語音
模型，合併成一合併語音模型，處理該轉換組合中所有的轉
換後，產生合併語音模型庫

740 將此合併語音模型庫套用至一語音合成器，將欲合成之輸入
文本以此語音合成器合成為帶有第一語言口音的一第二語
言語音

800 多語言之文字轉語音合成系統

810 處理器

890 記憶體裝置

發明專利說明書

100年1月13日修正
補充頁

(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：99146948

※申請日：99.12.30

※IPC 分類：G10L13/00 (2006.01)

G10L15/00 (2013.01)

一、發明名稱：(中文/英文)

多語言之文字轉語音合成系統與方法/

MULTI-LINGUAL TEXT-TO-SPEECH SYNTHESIS SYSTEM AND METHOD

二、中文發明摘要：

一種多語言之文字轉語音合成系統將欲合成之文本，藉由一語音模型挑選模組及一語音模型合併模組處理，利用一離線階段得到的一語音單元轉換表，於一在線階段時，此語音模型挑選模組，利用設定的至少一可調控的口音權重參數，選擇要採用的一轉換組合，找出一第二語音模型及一第一語音模型，此語音模型合併模組依照設定的至少一可調控的口音權重參數，將找出的兩語音模型合併成一合併語音模型，處理該轉換組合中所有的轉換後，產生合併語音模型庫，然後利用一語音合成器以及此合併語音模型庫將文本合成帶有第一語言口音之第二語言語音。

三、英文發明摘要：

In one embodiment of a multi-lingual text-to-speech synthesis system, text to be synthesized is processed by a speech model selection module and a speech model combination module. In an on-line phase, the speech model selection module finds a second-language speech model and a first-language speech model, by using at least a controllable accent weight parameter to choose which transformation composition to adopt from a phonetic unit transformation table obtained in an off-line phase. The speech model combination module combines the two speech models into a new speech model, according to the at least a controllable accent weight parameter. A second-language speech with the first-language

accent is further synthesized from the text by a speech synthesizer with the new speech model.

四、指定代表圖：

(一)本案指定代表圖為：第（一）圖。

(二)本代表圖之元件符號簡單說明：

100 多語言之文字轉語音合成系統

101 離線階段

102 在線階段

L1 第一語言

L2 第二語言

110 語音單元轉換表建立模組 112 帶有 L1 口音的 L2 語料庫

114 L1 語音模型庫

116 L2 轉 L1 之語音單元轉換表

120 語音模型挑選模組

150 可調控的口音權重參數

122 輸入文本及對應文本之 L2 語音單元序列

126 L2 語音模型庫

128 L1 語音模型庫

130 語音模型合併模組

132 合併語音模型庫

140 語音合成器

142 帶有 L1 口音之 L2 語音

五、本案若有化學式時，請揭示最能顯示發明特徵的化學式：

七、申請專利範圍：

1. 一種多語言之文字轉語音合成系統，利用一第一語言語

音模型庫與一第二語言語音模型庫將文字轉為語音，該系統包含：

一語音模型挑選模組，對欲合成之一含有一第二語言之輸入文本及對應該輸入文本之該第二語言的部分之一第二語言語音單元序列，在該第二語言語音模型庫中找出該第二語言語音單元序列中各語音單元所對應的一第二語音模型後，再查詢一第二語言轉第一語言之語音單元轉換表，並利用設定的至少一可調控的口音權重參數，決定要採用一轉換組合，選擇出一相對應的第一語言語音單元序列，並在該第一語言語音模型庫中找出該第一語言語音單元序列中各語音單元所對應的一第一語音模型；

一語音模型合併模組，將找出的該第二語音模型與該第一語音模型，依照設定的至少一可調控的口音權重參數，合併成一合併語音模型，處理該轉換組合中所有的轉換後，產生一合併語音模型庫；以及

一語音合成器，該合併語音模型庫被套用至該語音合成器，並且該語音合成器將該欲合成之輸入文本合成為帶有一第一語言口音的一第二語言語音。

2. 如申請專利範圍第1項所述之系統，其中一語音單元轉

換表建立模組於一離線階段時，透過一語音單元轉換表建立模組，根據帶有該第一語言口音的一第二語言語料庫及該第一語言語音模型庫，產生該第二語言轉第一語

言之語音單元轉換表。

3. 如申請專利範圍第1項所述之系統，其中該語音模型合併模組將找出的該第二語音模型與該第一語音模型以一權重方式計算，合併成該合併語音模型庫。
4. 如申請專利範圍第1項所述之系統，其中該第二語音模型與該第一語音模型至少包含一聲學參數。
5. 如申請專利範圍第4項所述之系統，其中該第二語音模型與該第一語音模型還包括一音長參數及一音調參數。
6. 一種多語言之文字轉語音合成系統，係執行於一電腦系統中，該電腦系統備有一記憶體裝置，至少儲存一第一語言語音模型庫與一第二語言語音模型庫，該文字轉語音合成系統包含：

一處理器，該處理器備有一語音模型挑選模組、一語音模型合併模組、及一語音合成器，該語音模型挑選模組對欲合成之一含有第二語言之輸入文本及對應該輸入文本第二語言的部分之一第二語言語音單元序列，在一第二語言語音模型庫中找出該第二語言語音單元序列中各語音單元所對應的一第二語音模型，再查詢一第二語言轉第一語言之語音單元轉換表，並利用設定的至少一可調控的口音權重參數，決定要採用一轉換組合，選擇出一相對應的第一語言語音單元序列，並在該第一語言語音模型庫中找出該第一語言語音單元序列中各語音單元所對應的一第一語音模型，該語音模型合併模組將找出的該第二語音模型與該第一語音模型，依照至少一可調控的口音權重參數，合併成一合併語音模型庫，

該合併語音模型庫再套用至該語音合成器，以合成為帶有一第一語言口音之第二語言語音。

7. 一種多語言之文字轉語音合成方法，係執行於一電腦系統中，該電腦系統備有一記憶體裝置，至少儲存一第一語言語音模型庫與一第二語言語音模型庫，該方法包含：

對欲合成之含有第二語言之輸入文本，利用對應該輸入文本第二語言的部分之一第二語言語音單元序列，在該第二語言語音模型庫中找出該第二語言語音單元序列中各語音單元所對應的一第二語音模型後，再查詢一第二語言轉第一語言之語音單元轉換表，並依照設定的至少一可調控的口音權重參數，決定要採用的一轉換組合，選擇一相對應的第一語言語音單元序列，並且在該第一語言語音模型庫中找出該第一語言語音單元序列中各語音單元所對應的一第一語音模型；

依照設定的至少一可調控的口音權重參數，將該找出的該第二語音模型與該第一語音模型，合併成一合併語音模型，處理該轉換組合中所有的轉換後，產生一合併語音模型庫；以及

將該合併語音模型庫套用至一語音合成器，並將欲合成之輸入文本以該語音合成器合成為帶有一第一語言口音的一第二語言語音。

8. 如申請專利範圍第 7 項所述之方法，該方法還包括建構該語音單元轉換表，建構該語音單元轉換表還包括：
- 從一第二語言語料庫中挑選出多個音檔以及與音檔相

對應的多個第二語言語音單元序列;

對挑選出的該多個音檔的每一音檔，由該第一語言語音模型來進行一自由音節式語音辨識，產生一辨識結果並將該辨識結果轉成一第一語言語音單元序列，並且將與該音檔相對應的一第二語言語音單元序列及轉成的該第一語言語音單元序列利用一動態編程來進行語音單元校準，完成該動態編程後，得到一筆轉換組合;以及統計由上述所得到的多筆轉換組合，產生該語音單元轉換表。

9. 如申請專利範圍第 8 項所述之方法，其中該動態編程還包括利用一種統計學上計算兩離散機率分布之間的距離的巴特查裡亞距離來計算兩語音單元之間的本地距離。
10. 如申請專利範圍第 7 項所述之方法，其中該語音單元轉換表包含代換、插入、及刪除，共三種類型的轉換。
11. 如申請專利範圍第 10 項所述之方法，其中代換是一對一的轉換，插入是一對多的轉換，刪除是多對一的轉換。
12. 如申請專利範圍第 10 項所述之方法，該方法利用該動態編程，找出該欲合成之輸入文本之對應語音單元與轉換類型。
13. 如申請專利範圍第 7 項所述之方法，其中該合併語音模型還包括以一高斯密度函數表示為 $g_{new}(\mu_{new}, \Sigma_{new})$ ，並以下列的形式來表達：

$$\mu_{new} = w * \mu_1 + (1 - w) * \mu_2$$

$$\Sigma_{new} = w * (\Sigma_1 + (\mu_1 - \mu_{new})^2) + (1 - w) * (\Sigma_2 + (\mu_2 - \mu_{new})^2)$$

其中，該找出的第一語音模型以高斯密度函數表示為 $g_1(\mu_1, \Sigma_1)$ ，該找出的第二語音模型以高斯密度函數表示為 $g_2(\mu_2, \Sigma_2)$ ， μ 為平均值向量， Σ 為共變異矩陣， $0 \leq w \leq 1$ 。

14. 如申請專利範圍第 8 項所述之方法，其中產生該辨識結果還包括進行一自由聲調式辨識。

accent is further synthesized from the text by a speech synthesizer with the new speech model.

四、指定代表圖：

(一)本案指定代表圖為：第（一）圖。

(二)本代表圖之元件符號簡單說明：

100 多語言之文字轉語音合成系統

101 離線階段

102 在線階段

L1 第一語言

L2 第二語言

110 語音單元轉換表建立模組 112 帶有 L1 口音的 L2 語料庫

114 L1 語音模型庫

116 L2 轉 L1 之語音單元轉換表

120 語音模型挑選模組

150 可調控的口音權重參數

122 輸入文本及對應文本之 L2 語音單元序列

126 L2 語音模型庫

128 L1 語音模型庫

130 語音模型合併模組

132 合併語音模型庫

140 語音合成器

142 帶有 L1 口音之 L2 語音

五、本案若有化學式時，請揭示最能顯示發明特徵的化學式：