200926139

發明專利說明書

（本說明書格式、順序及粗體字，請勿任意更動，※記號部分請勿填寫）

※ 申請案號：97139410

※ 申請日期：97 年 10 月 14 日

※IPC 分類：
G10L 13/02
G10L 15/26

一、發明名稱：（中文/英文）

使用聲音資料之字素至音素轉換

GRAPHEME-TO-PHONEME CONVERSION USING ACOUSTIC DATA

二、申 請 人：（共 1 人）

姓名或名稱：（中文/英文）

美商．微軟公司

Microsoft Corporation

代表人：（中文/英文）

艾苹那諾爾 D 巴特萊

EPPENAUER, D. BARTLEY

住居所或營業所地址：（中文/英文）

美國華盛頓州列德蒙微軟路 1 號

One Microsoft Way, Building 8, Redmond, WA 98052-6399, U.S.A.

國 籍：（中文/英文）

美國/USA

三、發明人：（共 3 人）

姓 名：（中文/英文）

1.李笑/LI, XIAO

2.古那沃達那亞希拉 JR/GUNAWARDANA, ASELA J. R.

3.亞塞羅亞力詹德/ACERO, ALEJANDRO

國　籍：(中文/英文)

1.中國/CHINA

2.美國/USA

3.美國/USA

四、聲明事項：

☐ 主張專利法第二十二條第二項☐第一款或☐第二款規定之事實，其事實發生日期為：　年　月　日。

☑ 申請前已向下列國家（地區）申請專利：

【格式請依：受理國家（地區）、申請日、申請案號　順序註記】

☑ 有主張專利法第二十七條第一項國際優先權：

美國；2007 年 12 月 7 日；11/952,267

☐ 無主張專利法第二十七條第一項國際優先權：

☐ 主張專利法第二十九條第一項國內優先權：

【格式請依：申請日、申請案號　順序註記】

☐ 主張專利法第三十條生物材料：

☐ 須寄存生物材料者：

國內生物材料　【格式請依：寄存機構、日期、號碼　順序註記】

國外生物材料　【格式請依：寄存國家、機構、日期、號碼　順序註記】

☐ 不須寄存生物材料者：

所屬技術領域中具有通常知識者易於獲得時，不須寄存。

國　籍：(中文/英文)

1.中國/CHINA

2.美國/USA

3.美國/USA

四、聲明事項：

☐ 主張專利法第二十二條第二項☐第一款或☐第二款規定之事實，其事實發生日期為：　年　月　日。

☑ 申請前已向下列國家（地區）申請專利：

【格式請依：受理國家（地區）、申請日、申請案號　順序註記】

☑ 有主張專利法第二十七條第一項國際優先權：

美國；2007 年 12 月 7 日；11/952,267

☐ 無主張專利法第二十七條第一項國際優先權：

☐ 主張專利法第二十九條第一項國內優先權：

【格式請依：申請日、申請案號　順序註記】

☐ 主張專利法第三十條生物材料：

☐ 須寄存生物材料者：

國內生物材料　【格式請依：寄存機構、日期、號碼　順序註記】

國外生物材料　【格式請依：寄存國家、機構、日期、號碼　順序註記】

☐ 不須寄存生物材料者：

所屬技術領域中具有通常知識者易於獲得時，不須寄存。

九、發明說明：

【發明所屬之技術領域】

本發明係與使用聲音資料之字素(grapheme)至音素(phoneme)轉換有關。

【先前技術】

字素至音素（G2P）轉換係指從字詞拼字（或者字素序列）自動建立發音（或者音素序列）。字素至音素為大型語音撥號系統中一種廣為使用的元件。

許多字素至音素系統係依據統計模型，其係使用人工編撰（hand-authored）之發音字典加以訓練。然而，該發音字典中具有的字素-音素關係通常係由語言學家所編寫，往往無法反應實際上人們如何發音字詞，或者無法涵蓋足夠變化。此使得此一字素至音素模型在語音相關任務上不夠理想。

舉例來說，考慮識別名稱之情形。其中一挑戰在於地域性差別；即某些存在於名稱中的字素-音素關係可能不存在於一發音字典中。即使某些名稱以及其發音可能被加入至該字典中，但大規模時如此是不實際的，且可能有大量的特殊名稱，再者稀有名稱常具有不尋常的發音。

另一挑戰為說話者的差異。來自不同地理區域以及種族族群的人們可能以不同方式發音相同名稱。一人工編撰之發音字典無法合理地抓住這些差異。

【發明內容】

提供此發明內容以利用一簡化形式介紹代表性概念之

一選擇，該些概念將於下文之實施方式中加以進一步描述。此發明內容並無意識別該申請專利主題之關鍵特徵或本質特徵，且無意以任何方式被用於限制該申請專利主題之範圍。

簡言之，本文中描述之主題的各種態樣係指一種技術，聲音資料、音素序列、字素序列及音素序列與字素序列間的一排列藉該技術提供一字音素（graphoneme）模型，其被用於再訓練語音辨識中可使用的一字素至音素模型。一般而言，該再訓練包括使用聲音資料最佳化該字音素模型。在一態樣中，最佳化該字音素模型包含使用該聲音資料執行該字音素模型之參數的最大或然率訓練或辨別訓練。在一態樣中，再訓練包括結合一發音字典以及聲音資訊，例如藉由內插（interpolate）字音素模型參數或者取得在該發音字典中與資料結合的字素-音素對。

在一示範實施中，對於再訓練中使用之聲音資料自動（例如不具有監督）收集字素標記。該聲音資料被接收為語音輸入，例如一語音撥號系統。該聲音資料被記錄及辨識為一潛在字素標記。若該說話者確認該潛在字素標記為正確的，則該聲音資料被持續與該字素標記相關聯。可進行多次與該說話者的互動以接收額外聲音資料並取得該確認。不符合一確信臨界值的語音輸入可被過濾掉而不被用於該再訓練模型中。

可從以下實施方式並結合圖示而明瞭其他優點。

【實施方式】

本發明中描述之技術的各種態樣一般係指調整（leverage）聲音資料，以適應於語音辨識之一字素至音素模型。在一態樣中，此可被用於直接地以聲音資料所建立之發音而增加或修改一既有發音字典。舉例來說，說話的名稱識別可使用從一大型語音撥號系統所取得之資料。

如下文所述，透過適應資料之適應性係於一字音素層級加以執行，其考慮一字素序列、一音素序列以及該字素序列與音素序列間的一排列及分組。將了解一產生的字素至音素轉換不僅增進該適應資料中具有之字詞（包括名稱）的發音，且亦推及未見過的字詞。因此在適應字音素模型參數中提供聲音及字音素的結合模型，伴隨前述兩示範訓練方法，分別名為最大或然率訓練以及辨別訓練。

將了解本文中提出之各種範例主要被描述為與識別說出之名稱及/或英文字詞及名稱有關。然而，可輕易了解該技術可適用於任何類型的語音辨識及/或其他語言。再者，雖然可輕易了解使用聲音資訊而增進字素至音素轉換可大幅增進語音辨識，但該相同技術亦可被用於增進語音合成的品質。

因此，本發明並不限於本文中描述之任何特定實施例、態樣、概念、結構、功能或範例。反之，本文中描述之任何實施例、態樣、概念、結構、功能或範例為非限制性的，且本發明可以各種方式加以使用，其一般在語音辨識及/或語音合成中提供利益及優勢。

就訓練而言，字素與音素之間的或然率關係係使用一

字音素 n-gram 模型（亦稱為一結合的多 gram 模型）加以建立。更特言之，為了從一發音字典建立可藉以訓練一字音素 n-gram 模型的字素序列，以下表格可被使用作為設置一字音素序列之概念的一範例（對於字詞「letter」）：

字素序列	l	e	t	t	e	r
音素序列	l	eh	t	ϵ	ax	r
字音素序列	l:l	e:eh	t:t	t:ϵ	e:ax	r:r

如本文中使用者，一任意變數 g 表示一字素序列，而 φ 表示一音素序列。該變數 s 表示 φ 與 g 的一排列以及一分組，如此範例中所定義者。

考慮該英文語言字詞字母，g 等於(l, e, t, t, e, r)而 φ 等於(l, eh, t, ax, r)。排列 g 及 φ 的一種可能方式被顯示於前文中，其中 ϵ 表示一空音素。在此一排列下，可藉由關聯字素與其音素配對物，如最後一行（字音素序列）中所示。

隨後，鄰近字音素單元可被群聚以形成較大單元。在前述範例中，結合 1:1 與 e:eh、e:ax 與 r:r，產生：

l&e:l&eh　t&t:t&ϵ　e&r:ax&r　　(1)

(1)之形式被定義為一字音素序列，其被(g,φ,s)所完全決定。

如(1)中的此字音素序列可從平行字素及音素序列之

一發音字典所建立，亦即 s 可由一組 (g,φ) 所推斷。至此，一步驟自動排列 g 及 φ 以透過一期望值最大化（EM）方法形成基元字音素，其排列係使用字音素單 gram 統計所推斷。另一步驟使用依據相互資訊之一演算法並藉由允許一字素單元具有最大的 k 字素及 l 音素而整合字素為較大單元。該結果為字音素序列的一語料庫。

若具有字音素序列之語料庫，則可訓練具有撤退（backoff）的一標準 n-gram 模型。注意依據訓練資料之數量，可使用一切斷臨界值以調整模型複雜度，例如一 n-gram 若其具有大於此臨界值的計數則將被排除於該模型之外。隨後，字素至音素轉換可藉由應用一最佳-第一搜尋演算法（或者其他適合的搜尋演算法）而達成。

現考量在適應於此一模型中使用聲音—即如本文中所述之最佳化字素至音素轉換而增進語音（包括名稱）辨識。關於此一最佳化，聲音資料在學習存在於真實世界應用程式中的字素-音素關係中為非常有用的。另一隨機變數 x 被用於表示聲音，其可與前述之其他變數（假設 x, g, φ 與 s）共同建構成為以下方程式(2)：

$$\log p_\theta(x,g,\phi,s)=\log p(x|\phi)+\log p_\theta(g,\phi,s) \qquad (2)$$

該參數化係依循 x 於特定 φ 與 g 及 s 無關的假設。其中，該結合可能性係由一聲音模型分數 $p(x|\phi)$ 及一字素模型分數 $p_\theta(g,\phi,s)$ 所表示，其中 θ 表示將被適應化的 n-gram 模型

參數。注意在此範例中使用一固定的聲音模型，故 $p(x|\phi)$ 並未被參數化。再者，可使用一縮放參數，其功用類似於語音辨識中的一語言模型縮放參數；方程式(2)變為：

$$\approx \log p(x|\phi) + a \log p_\theta(g,\phi,s) \qquad (3)$$

該縮放參數 a 的一合理數值 0.25 已被使用。注意為了簡單性，a 在以下範例中被省略，然而實際上此一參數 a 被用於方程式(2)中。

如前文中可見者，方程式(2)（因而以及方程式(3)）提供一種機制，其結合地建構聲音、一字素序列、一音素序列、以及該音素序列以及字素序列之間的排列。該結合或然率可依據方程式(2)加以參數化。注意方程式(2)中的第一項可為任何聲音模型，其為固定的。該第二項為一 n-gram 型字音素模型，其將被最佳化。注意如本文中所述，「最佳化」一詞以及其變化詞（例如最佳化中）與「最大化」一詞以及其變化詞（例如最大化中）不必然意指理想的或完美的，反之可意指移向此狀態（例如達到某收斂臨界值、更接近一真實最大值等等而較先前更佳）。

一般而言，x 及 g 兩者（該聲音資料以及其字素標記）為可見的，然而 φ 及 s（該音素序列以及排列）為隱藏的。亦可具有一適應資料（x_i, g_i）之集合。注意如本文中所述（例如參照第 4 圖之示範流程圖），聲音資料 x 的字素標記 g 可於一未監督之方式中加以取得。

現參考第 1 圖，該圖顯示一般概念圖，其包括使用標

記化聲音資料作為該適應資料 104 而再訓練一辨識器 102 的元件。在第 1 圖之範例中，呼叫記錄開採（call log mining）106 從該聲音資料記錄 104 中析取一組波形，即字素序列對（x_i, g_i）。下文參照第 4 圖之示範流程圖描述如何可在一未監督之方式中從實際使用者說話收集此波形(即字素序列對 108)的一範例。

如下文所述，該辨識器 102 一開始使用一或然率式字音素對音素模型 110，模型 110 係使用一發音字典加以訓練以辨識特定波形、字素序列對 108 的音素序列，例如提供字素序列、音素序列對（g_i, ϕ_i）作為辨識輸出 112。另如下文所述，一再訓練機制 114 隨後產生一字素至音素模型 116，其被用於後續辨識中。如前文所見，基本上提供一反饋迴路，因而當更多標記的聲音資料（波形、字素對）變為可取得時，該字素至音素模型 116 變得更佳。

若具有該組標記的適應資料 108，一字音素 n-gram 模型之適應的一潛在方式為重新估計模型參數(模型參數係最大化該結合可能性記錄 $p(x,g)$，其導致最大可能性估計（MLE））。替代地，可使用一辨別訓練（DT）方式直接最大化該條件可能性記錄 $p(g|x)$。這兩種方式將分別參照第 2 及 3 圖於下文加以描述。

現考量使用聲音資料之字音素模型之最大可能性訓練情形，若具有字素序列以及聲音取樣對，則訓練將方程式(4)最大化的 n-gram 式字音素模型參數：

$$\sum_{i=1}^{m}\log p_\theta(x_i,g_i)=\sum_{i=1}^{m}\log\sum_{\phi_i,s_i}p_\theta(x_i,g_i\phi_i,s_i) \qquad (4)$$

換言之，給定一組（x_i,g_i）對，該最大可能性估計的目標為將方程式(4)最大化。在一範例中，可使用第2圖之步驟中所說明之演算法達成該訓練。

更特言之，一標準 EM 演算法可被用於應付隱藏變數 $\{\phi_i,s_i\}_{i=1}^{m}$。替代地，該 Viterbi 演算法可被應用，其被描述於本文中。

一種此特殊最佳化程序被顯示於第2圖中，開始於第2圖之步驟202，其表示從前述在一發音字典上訓練的一基線字音素模型(θ_0)開始。

步驟204表示在給定觀察的（x_i,g_i）下以及目前模型θ估計而找出最有可能的 ϕ_i 和 s_i：

$$\begin{aligned}\hat{\phi}_i,\hat{s}_i &= \underset{\phi_i,s_i}{\operatorname{argmax}}\ \log p_\theta(\phi_i,s_i|x_i,g_i)\\ &= \underset{\phi_i,s_i}{\operatorname{argmax}}\ \log p(x_i|\phi_i)+\log p_\theta(g_i,\phi_i,s_i)\end{aligned} \qquad (5)$$

步驟206藉由以下式子重新估計該模型：

$$\hat{\theta}=\underset{\theta}{\operatorname{argmax}}\sum_{i=1}^{m}\log p_\theta(g_i,\hat{\phi}_i,\hat{s}_i) \qquad (6)$$

藉由重複步驟204及206來反覆執行步驟208直到收斂為止，其於訓練機制中為典型的。

注意為了降低運算，在前述演算法的每個反覆中，該目前的字音素模型被用於對一特定字素序列建立n種最佳

的音素序列假設，之後方程式(5)被用於重新評價（重新評分）該n種最佳的假設。換言之，就運算方便性而言，方程式(5)中的「argmax」運算對於每個i僅用於產生最高之 $\log p_\theta(g_i; \phi_i; s_i)$ 分數之最前面的n種音素序列。再者注意該n種最佳的列表可由一語音解碼器所建立，但此可能導致『語言學上的不正確』發音，其非訓練一字素至音素模型所希望的。

另一個考慮的議題在於當 g_i 並非 x_i 的正確標記時，其可特別發生在使用未監督資料集合時（如以下參照第4圖之描述）。此惱人的範例將『污染』該字音素模型。因此提供一態樣來藉由設置一確信臨界值而過濾掉惱人資料，亦即使用一聲音模型確信數值α，一取樣於以下情形被忽略：

$$\log p(x_i \mid \hat{\phi}_i) < \alpha \qquad (7)$$

此過濾之一基礎為當 g_i 並非 x_i 的正確標記時，則該n種最佳的 ϕ_i（因此以及 $\hat{\phi}$）之任一者將不大可能產生一高聲音模型分數。

前述方式產生一字音素模型，其於一適應集合方面被最佳化。依據適應資料之數量，此模型可能十分普及或者無法十分普及。一種更穩固的方法可利用源自訓練該基線字音素模型之發音字典的資訊。此與嘗試藉由利用既有、區域外資料而學習一新區域（通常具有有限的資料）的模

型相似。

在適應一字音素模型之內容中可使用各種策略，包括模型內插（interpolation）及/或資料結合。一般而言，模型內插係將使用聲音資料訓練之字音素模型參數線性地內插入使用一發音字典之字音素模型參數。換言之，模型內插從方程式(6)取得一模型 θ^{ML}（在收斂之後），並以該基線字音素模型 θ_0 線性地內插之。該內插權值被調整於一開發設置上。

關於資料結合，一般而言給定字素序列以及聲音取樣對，該對應的音素序列係由方程式(5)於訓練收斂之後所取得。以此方式取得的字素-音素對可被直接與一發音字典結合以供再訓練。換言之，資料結合對每個 i 從方程式(5)取得 $\hat{\phi}$（仍在收斂之後）。資料結合隨後結合 $\{(g_i,\hat{\phi}_i)\}_{i=1}^{m}$ 與該原始發音字典，並如前文所述再訓練一模型。在此方面，$\{(g_i,\hat{\phi}_i)\}_{i=1}^{m}$ 之功能類似於從聲音資料所建立的一『發音字典』。然而，不像一典型發音字典中每個 (g,φ) 數值為獨特的，$\{(g_i,\hat{\phi}_i)\}_{i=1}^{m}$ 可含有相同的輸入，對於多個 i 而言 $(g_i = g, \hat{\phi}_i = \phi)$。注意此冗餘可為有用的，由於其自然地定義不具有一發音字典的一先前分佈 $p(g,\varphi)$。注意此冗餘亦可藉由在資料結合後破壞相同輸入而加以移除。

關於一替代訓練機制，若具有給定字素及聲音取樣對，即使用聲音資料之字音素模型參數之辨別訓練（DT）的一示範實施訓練最大化方程式(8)的 n-gram 式字音素模型參數，其如下文所述依據方程式(9)使用近似。注意雖然

最大可能性估計目標找出最佳描述該資料的參數，且在該模型結構為正確的假設之下於統計上是一致的，而該訓練資料係從該真實分佈所建立，且具有非常大量的此種訓練資料，但此狀態實際上是鮮能被滿足的。因此，直接目標為更佳分類/辨識效能之辨別訓練通常產生較佳效能。

在字素-音素轉換之內容中，辨別訓練嘗試估計字音素模型參數之一方式係此模型所建立之發音最大地降低辨識錯誤者。以下方程式係指給定聲音下最大化一字素序列之條件可能性：

$$\sum_{i=1}^{m} \log p_\theta(g_i|x_i) = \sum_{i=1}^{m} \log \frac{p_\theta(x_i, g_i)}{\sum_{g_i'} p_\theta(x_i, g_i')} \qquad (8)$$

$p(x_i, g_i)$之運算涉及 φ_i, s_i 上的忽略。在此建立一近似為：

$$p_\theta(x_i, g_i) = \sum_{\phi_i, s_i} p_\theta(x_i, g_i) \approx p_\theta(x_i, g_i, \hat{\phi}_i, \hat{s}_i) \qquad (9)$$

其中 $\hat{\phi}_i, \hat{s}_i$ 被定義於方程式(5)中。因此方程式(8)變為：

$$\approx \sum_{i=1}^{m} \log \frac{p(x_i|\hat{\phi}_i) p_\theta(\hat{\phi}_i, \hat{s}_i, g_i)}{\sum_{g_i'} p(x_i|\hat{\phi}_i) p_\theta(\hat{\phi}_i', \hat{s}_i', g_i')} \qquad (10)$$

可應用隨機梯度下降以找出一區域最佳估計 θ^{DT}。

更特言之，一示範訓練程序係參照第 3 圖之流程圖而

加以描述，其開始於步驟 302，其表示以前述之一 ML 適應之字音素模型 θ^{ML} 開始。步驟 304 藉由使用一語音辨識器以及使用該 ML 適應之字音素而取得 x_i 之 n 種最佳的辨識結果 g'_i。於步驟 306，對於聲音/字素對 (x_i, g_i) 而言，$\hat{\phi}_i, \hat{s}_i$ 係由方程式(5)所取得；類似地對於每對 (x_i, g'_i) 而言，$\hat{\phi}'_i, \hat{s}'_i$ 係由方程式(5)所取得。

步驟 308 係將隨機梯度下降應用至關於 θ 之方程式(10)。早期停止被應用以避免過適。注意在具有撤退（backoff）之一 n-gram 模型中，若一 n-gram 不存在於此模型中，則其或然率係由撤退至一較低階分佈而加以運算。關於如何在辨別訓練中控制撤退具有數種選擇，一撤退權值為固定的而較低階 n-gram 參數於步驟 308 中被更新。

步驟 310 重複步驟 304、306 及 308 直到收斂為止。

就在 n 種最佳字素序列上（方程式(10)）執行最佳化之訓練的一致性而言，該辨別之訓練模型可藉由類似方式加以評估。對於該測試集合中的每個 x_i 而言，n 種最佳的 g'_i 係使用一語音辨識器以及使用該 ML 適應之字音素模型 θ^{ML} 所建立。以下使用辨識訓練所取得之模型重新評分 g'_i：

$$\begin{aligned} \hat{g}_i &= \underset{g'_i}{\operatorname{argmax}}\, p_{\theta^{DT}}(g'_i|x_i) \\ &= \underset{g'_i}{\operatorname{argmax}}\, p(x_i|\hat{\phi}'_i) p_{\theta^{DT}}(\hat{\phi}'_i, \hat{s}'_i, g'_i) \end{aligned} \qquad (11)$$

注意該相同近似被用作為方程式(10)中所使用者。依據 $\hat{g}_i$ 之辨識錯誤率可被測量為自重新評分所取得者。

現考量另一態樣，其描述聲音資料之字素標記的未監督取得。至此部分，各種類型的聲音資料可自動從對應至實際使用者呼叫之呼叫記錄加以取得。為確認辨識而與該使用互動（或者若無法確認時最終提供一拼字）提供該聲音資料之字素標記。

舉例來說，第 4 圖顯示此一與名稱識別有關之未監督取得機制的示範步驟。在第 4 圖中，當一使用者說出一個人的名稱（例如 John Doe）時，該聲音資料被接收並記錄（步驟 402）。注意於步驟 404 該辨識器判定一名稱，並使用文字至語音提示該使用者確認該辨識之名稱，例如『你剛是說 John Doe 嗎？』

若於步驟 406 該使用者對該提示確認『是』，則於步驟 404 中對應至該辨識之名稱的文字在步驟 408 中被使用作為步驟 402 之記錄聲音的字素標記。該處理隨後結束，即使傳送該呼叫至該請求之使用者等等的適當動作被當然執行。

相反地，若該使用者於步驟 406 確認『否』，在此範例中該系統/辨識器於步驟 410 請求該使用者重複該名稱；（一替代選項為在請求重複前提示該使用者第二最為可能的名稱）。步驟 412、414 及 416 基本上類似於步驟 402、404 及 416，亦即該使用者說出一名稱、該辨識器從該接收之語音辨識一名稱並且請求該呼叫者確認該重新辨識之名稱。注意與該第一集合起碼有些許不同之一第二集合之聲音資料於步驟 412 被接收（一般而言，除非該呼叫者掛

斷）；若該辨識器判定該相同名稱（其最可能是不正確的，除非該呼叫者不小心說出『否』），該辨識器可提示關於此聲音資料輸入所辨識的第二最可能名稱。

若該使用者於步驟 416 對此範例中的此第二嘗試確認『是』，則於步驟 414 識別之名稱在步驟 418 被使用作為每個該記錄之聲音的字素標記，如步驟 402 所記錄者（被標記為『第一』）以及步驟 412 所記錄者（被標記為『第二』）。注意一典型使用者將說出該相同名稱兩次且因而該第一及第二聲音資料集合係與該相同的字素標記相關聯。然而，若該使用者清楚說出不同名稱而非重複該相同名稱，則該聲音資料之第一集合可被方程式(7)等等所得到之一過濾機制加以忽略。

回到步驟 416，若該使用者對該第二嘗試確認『否』，則在此範例中該辨識器/系統於步驟 420 請求拼出該名稱。步驟 422 表示該使用者提供該拼字並對從該拼字所擷取之一名稱確認『是』；注意若該系統可從該輸入辨識一正確的拼字名稱(舉例來說與自動拼字檢查類似)，仍將使用一不正確拼字的名稱。假設一正確拼字之名稱被輸入（或者相符），該拼字之名稱於步驟 424 被使用作為每個該（第一及第二）記錄之聲音的字素標記。注意一不適當拼字之一名稱仍可維持為一正確的字素標記，因為其他的呼叫者可能類似地錯誤拼字此一名稱。

可輕易了解人工轉寫（manual transcription）在一

大型語音撥號系統中由於該文法中的大量名稱(且有時為易混淆的)而為一昂貴且易錯的任務。如前所述依據成功確認的一對話分析自動取得聲音資料之一子集和字素標記。在此對話階段之末尾，該系統記錄該呼叫被傳輸至該請求方的一事件。由於在傳輸前該使用者與該系統確認的正確性(『是』)，可合理地假設被傳輸之該被呼叫者的名稱為該對應波形的正確字素標記。如第4圖之範例中所呈現者，該系統於該使用者給予一正面確認之前可經歷多次互動，藉此從該確認之傳輸所取得之字素標記可對應該對話其中的多個波形。

然而可輕易了解此一假設在一未監督系統中可能導致該資料中的雜訊。舉例來說，一呼叫所傳輸的目標可能不是該一或多個對應波形的真實字素標記。事實上，一使用者可疏忽地或以其他方式就一不正確辨識的名稱對該系統確認『是』(通常是因為文字至語音所建立的易混淆發音)，因而該呼叫被傳輸至一錯誤者而該字素不符合該聲音資料。如前所述，方程式(7)被用於從該適應集合忽略(過濾掉)此雜訊資料，如第4圖中透過步驟426所表示者。

示範操作環境

第5圖說明可實施第1-4圖之範例於其上的一適用運算系統環境500。該運算系統環境500僅為一適用運算環境之一範例且無意對本發明之使用或功能範圍假設任何限制。該運算環境500亦不應被解釋為具有與該示範操作環

境500之任一者或結合有關之任何依賴或需求。

本發明可伴隨各種其他一般目的或特殊目的之運算系統環境或配置加以操作。可伴隨本發明家以使用之著名運算系統、環境及/或配置的範例包括但不限於個人電腦、伺服器電腦、手持或筆記型電腦裝置、平板裝置、多處理器系統、微處理器式系統、機上盒、可程式化消費型電子裝置、網路個人電腦、微電腦、大型主機電腦、包括前述系統或裝置之任一者等等的分散式運算環境。

可於電腦可執行指令之一般內容中描述本發明，例如由一電腦執行之程式模組。一般而言，程式模組包括常式、程式、物件、元件、資料結構等等，其執行特定任務或實施特定抽象資料類型。本發明亦可被實施於一分散式運算環境中，其中任務係由透過一通信網路連結之遠端處理裝置所執行。在一分散式運算環境中，程式模組可位於區域及/或遠端電腦儲存媒體中，其包括記憶體儲存裝置。

參照第5圖，用於實施本發明之各種態樣的一示範系統包括一一般目的運算裝置，其形式為一電腦510。該電腦510之元件可包括但不限於一處理單元520、一系統記憶體530以及一系統匯流排521，其連接包括該系統記憶體之各種系統元件至該處理單元520。該系統匯流排521可為各種匯流排結構之任一者，包括一記憶體匯流排或記憶體控制器、一週邊匯流排、以及使用各種匯流排結構之任一者的一區域匯流排。不受限地舉例來說，此結構包括工業標準結構（ISA）匯流排、微通道結構（MCA）匯流

排、增強型 ISA(EISA)匯流排、視訊工業標準協會(VESA)區域匯流排以及亦稱為 Mezzanine 匯流排之週邊元件互連(PCI)匯流排。

該電腦 510 典型地包括各種電腦可讀取媒體。電腦可讀取媒體可為任何可被該電腦 510 所存取之可用媒體，且其包括依電性及非依電性記憶體以及可移除及不可移除記憶體。不受限地舉例來說，電腦可讀取媒體可包含電腦儲存媒體及通信媒體。電腦儲存媒體包括以任何方法或技術所實施以供儲存例如電腦可讀取指令、資料結構、程式模組或其他資料的依電性及非依電性、可移除及不可移除媒體。電腦儲存媒體包括但不限於隨機存取記憶體(RAM)、唯讀記憶體(ROM)、電子可抹除可程式化唯讀記憶體(EEPROM)、快閃記憶體或其他記憶體技術、CD-ROM、數位多媒體光碟(DVD)或其他光碟儲存、磁匣、磁帶、磁碟儲存或其他雌性儲存裝置、或者任何其他可被用於儲存該所需資訊且可被電腦 510 所存取之媒體。通信媒體典型地包含位於一模組化資料信號例如一載波或其他傳輸機制中的電腦可讀取指令、資料結構、程式模組或的其他資料，並包括任何資訊傳遞媒體。『模組化資料信號』一詞係指其一或多個特性係以一方式加以設置或改變以加密該信號中的資訊。不受限地舉例來說，通信媒體包括有線媒體例如一有線網路或者直接連線、以及無線媒體例如聲波、無線電頻率(RF)、紅外線及其他無線媒體。前述任何結合亦應被包括於電腦可讀取媒體之範圍中。

該系統記憶體 530 包括依電性及/或非依電性記憶體形式之電腦儲存媒體，例如唯讀記憶體（ROM）531 或隨機存取記憶體（RAM）532。含有例如於啟動過程中協助電腦 510 中各元件傳輸資訊之基本常式的一基本輸入/輸出系統 533(BIOS)典型地被儲存於 ROM 531 中。RAM 532 典型地含有可立即被存取及/或目前正被處理單元 520 操作的資料及/或程式模組。不受限地舉例來說，第 6 圖說明作業系統 534、應用程式 535、其他程式模組 536 以及程式資料 537。

該電腦 510 亦可包括其他可移除/不可移除、依電性/非依電性電腦儲存媒體。僅舉例而言，第 5 圖說明讀寫不可移除、非依電性磁性媒體的一硬碟機 541、讀寫一可移除、非依電性磁碟 552 的一磁碟機 551、以及讀寫一可移除、非依電性光碟 556 如一 CD-ROM 或其他光學媒體的一光碟機 555。其他可被用於該示範作業環境之可移除/不可移除、依電性/非依電性電腦儲存媒體包括但不限於磁帶匣、快閃記憶卡、數位多媒體碟片、固態 RAM、固態 ROM 等等。該硬碟機 541 典型地透過一不可移除記憶體介面如介面 340 被連接至該系統匯流排 521，而磁碟機 551 及光碟機 555 典型地藉由一可移除記憶體介面如介面 550 被連接至該系統匯流排 521。

前文討論並於第 5 圖中說明之機器以及其相關的電腦儲存媒體提供電腦可讀取指令、資料結構、程式模組及用於該電腦 510 之其他資料的儲存。例如在第 5 圖中，硬碟

機 541 被說明為儲存作業系統 544、應用程式 545、其他程式模組 546 以及程式資料 547。注意這些元件可為相同或不同於作業系統 534、應用程式 535、其他程式模組 536 以及程式資料 537 者。作業系統 544、應用程式 545、其他程式模組 546 以及程式資料 547 於此處被賦予不同編號以說明至少其為不同複本。一使用者可透過輸入裝置例如一平板、或者電子數位轉換器 564、一麥克風 563、一鍵盤 562、以及一指標裝置 361 一般係指滑鼠、軌跡球或觸控板輸入指令及資訊至該電腦 310 中。未顯示於第 5 圖之其他輸入裝置可包括一搖桿、遊戲控制盤、衛星碟、掃瞄器等等。這些及其他輸入裝置通常透過連接至該系統匯流排之一使用者輸入介面 560 被連接至該處理單元 520，但亦可由其他介面及匯流排結構加以連接，例如一平行埠、遊戲埠或一通用序列匯流排（USB）。一螢幕 591 或其他類型之顯示裝置亦透過一介面例如一視訊介面 590 被連接至該系統匯流排 521。該螢幕 591 亦可被整合一觸控畫面面板等等。注意該螢幕及/或觸控畫面面板可被實體連接至該運算裝置 510 被裝入之一容置空間，例如一平板型個人電腦中。除此之外，如運算裝置 510 之電腦亦可包括其他週邊輸出裝置例如喇叭 595 以及印表機 596，其可透過一輸出週邊介面 594 等等加以連接。

該電腦 510 使用邏輯連接至一或多個遠端電腦例如一遠端電腦 580 而被操作於一網路化環境中。該遠端電腦 580 可為一個人電腦、一伺服器、一路由器、一網路個人電腦、

一點裝置或其他共用網路節點，且典型地包括許多或所有前文關於該電腦 510 所描述之元件，即使在第 5 圖中僅說明一記憶體儲存裝置 581。第 5 圖中描繪之邏輯連接包括一區域網路（LAN）571 以及一廣域網路（WAN）573，但亦可包括其他網路。此網路環境在辦公室、企業範圍電腦網路、內部網路以及網際網路中均為常見的。

當使用於一 LAN 網路環境中時，該電腦 510 透過一網路介面或配接器 570 被連接至該 LAN 571。當使用於一 WAN 網路環境中時，該電腦 510 典型地包括一數據機 572 或其他用於在該 WAN 573 上建立通信的方式，例如該網際網路。可為內接或外接式的數據機 572 在一網路環境中可透過該使用者輸入介面 560 或其他適合機制被連接至該系統匯流排 321。一無線網路元件 574 例如包含一介面以及天線可透過一適用裝置如一存取點或同儕電腦被連接至一 WAN 或 LAN。在一網路環境中，關於該電腦 510 所描繪之程式模組或者其部分可被儲存於遠端記憶體儲存裝置中。不受限地舉例來說，第 5 圖說明遠端應用程式 585 位於遠端電腦 581 上。將了解圖示之網路連接僅為例示性且可使用能於該電腦間建立一通信連結的其他方式。

一輔助子系統 599（例如用於輔助之內容顯示）可透過該使用者介面 560 加以連接以允許例如程式內容、系統狀態及事件通知等資料被提供給該使用者，即使該電腦系統之主要部分處於一低電源狀態中。該輔助子系統 599 可被連接至該數據機 572 及/或網路介面 570 以於該主處理單

元 520 處於一低電源狀態中時允許這些系統之間的通信。

結論

雖然本發明容許各種修改及替代結構，但其特定說明性實施例被顯示於圖示中且已於前文加以詳細描述。然而，應了解並無意將本發明限制於已揭露之特定形式，相反地，其議題涵蓋所有修改、替代結構以及位於本發明之精神及範圍中的均等物。

【圖式簡單說明】

本發明係由範例加以說明且不受限於附隨圖示，圖示中相似的編號係指相似的元件，其中：

第 1 圖為一方塊圖，其表示用於訓練及適應一語音識別模型的示範元件；

第 2 圖為一流程圖，其表示可被用於訓練一語音識別模型以使用最大或然率估計而調整參數的示範步驟；

第 3 圖為一流程圖，其表示可被用於訓練一語音識別模型以使用辨別訓練而調整參數的示範步驟；

第 4 圖為一流程圖，其表示可被用於取得聲音資料並指派字素標記至該未監督資料集合中的聲音資料的一示範步驟；及

第 5 圖顯示可納入本發明之各種態樣之一電腦環境的一說明性範例。

【主要元件符號說明】

102 辨識器

104 聲音資料記錄

106 呼叫記錄挖掘

108 適應資料

110 初始模型

112 辨識輸出

114 再訓練機制

116 字素至音素模型

500 運算系統環境

510 電腦

520 處理單元

521 系統匯流排

530 系統記憶體

531 唯讀記憶體（ROM）

532 隨機存取記憶體（RAM）

534, 544 作業系統

535, 545 應用程式

536, 546 程式模組

537, 547 程式資料

540, 550 記憶體介面

541 硬碟機

551 磁碟機

552 磁碟

555 光碟機

556 光碟

560 使用者輸入介面

561 指標裝置

562 鍵盤

563 麥克風

564 電子數位轉換器

570 配接器

571 區域網路（LAN）

572 數據機

573 廣域網路（WAN）

580 遠端電腦

581 記憶體儲存裝置

585 遠端應用程式

590 視訊介面

591 螢幕

594 輸出週邊介面

595 喇叭

596 印表機

五、中文發明摘要：

本發明描述使用聲音資料以增進語音辨識之字素至音素轉換，例如在一語音撥號系統中更正確地辨識說出之名稱。本發明描述聲音及字音素之一結合模型（聲音資料、音素序列、字素序列以及音素序列與字素序列之間的一排列），其於使用聲音資料來適應字音素模型參數中係由最大可能性訓練及辨別訓練加以再訓練。本發明亦描述接收之聲音資料之字素標記的未監督集合，藉以自動取得可被用於再訓練之一實質數量的實際取樣。不符合一確信臨界值的語音輸入可被過濾掉而不會被該再訓練模型使用。

六、英文發明摘要：

Described is the use of acoustic data to improve grapheme-to-phoneme conversion for speech recognition, such as to more accurately recognize spoken names in a voice-dialing system. A joint model of acoustics and graphonemes (acoustic data, phonemes sequences, grapheme sequences and an alignment between phoneme sequences and grapheme sequences) is described, as is retraining by maximum likelihood training and discriminative training in adapting graphoneme model parameters using acoustic data. Also described is the unsupervised collection of grapheme labels for received acoustic data, thereby automatically obtaining a substantial number of actual samples that may be used in retraining. Speech input that does not meet a confidence threshold may be filtered out so as to not be used by the retrained model.

十、申請專利範圍：

1. 一種在一運算環境中的方法，至少包含以下步驟：

建構一聲音資料、一音素序列、一字素序列、以及該音素序列及字素序列之間的一排列，以提供一字音素模型；及

藉由使用該聲音資料來最佳化該字音素模型而再訓練語音辨識中可使用之一字素至音素模型。

2. 如請求項第1項所述之方法，其中最佳化該字音素模型之步驟包含以下步驟：

使用該聲音資料執行字音素模型參數之最大可能性訓練。

3. 如請求項第2項所述之方法，其中該最大可能性訓練包括使用一目前字音素模型以建立一特定字素序列之一最佳音素序列假設集合，以及依據聲音資料以及該目前字音素模型來重新評價該最佳假設之集合。

4. 如請求項第1項所述之方法，其中最佳化該字音素模型之步驟包含以下步驟：

使用該聲音資料執行字音素模型參數的辨別訓練。

5. 如請求項第1項所述之方法，其中再訓練該字素至音素模型之步驟包含以下步驟：

結合一發音字典以及聲音資訊。

6. 如請求項第5項所述之方法，其中結合該發音字典及該聲音資訊之步驟包含以下步驟：

將透過最大可能性訓練或者使用聲音資料之辨別訓練

所訓練字音素模型參數來內插使用一發音字典所訓練者。

7. 如請求項第5項所述之方法，其中結合該發音字典及該聲音資訊之步驟包含以下步驟：

取得對應至一聲音波形取樣之一音素序列以及對應至該相同聲音波形取樣之一字素序列以取得一字素-音素對，以及將一實質數量之此字素-音素對與該發音字典中的資料結合。

8. 如請求項第1項所述之方法，更包含以下步驟：

收集源自一說話者被接收作為語音輸入之一聲音波形的一字素標記，包括：藉由記錄該聲音資料，辨識該聲音資料為一潛在字素標記、從該說話者取得該潛在字素標記正確用於該語音輸入之確認、以及於確認時維持該語音資料與對應至該潛在字素標記之一實際字素標記相關聯。

9. 如請求項第8項所述之方法，更包含以下步驟：

執行複數個與該說話者間的互動以接收該聲音資料並取得該確認。

10. 如請求項第8項所述之方法，更包含以下步驟：

過濾掉不符合一確信臨界值之該聲音資料以及語音輸入之關聯字素標記。

11. 一種在一運算環境中的系統，至少包含：

一字素至音素模型；

一辨識器，連接至該字素至音素模型以辨識輸入語音為一對應的字素序列；及

一再訓練機制，連接至該辨識器，該再訓練機制依

據聲音資料以及一辨識系統所收集之關聯字素而再訓練該字素至音素模型為一再訓練的字素至音素模型。

12. 如請求項第11項所述之系統，其中該再訓練機制依據該聲音資料執行字素至音素模型的最大可能性訓練或辨別訓練。

13. 如請求項第11項所述之系統，其中該再訓練機制藉由將使用聲音資料所訓練之該些字素至音素模型參數來內插使用一發音字典所訓練者而結合一發音字典及聲音資訊，或者藉由取得對應至一聲音波形取樣之一音素序列以及對應至該相同聲音波形取樣之一字素序列以取得一字素-音素對，以及將一實質數量之此字素-音素對與該發音字典中的資料結合。

14. 如請求項第11項所述之系統，其中該辨識系統收集該聲音資料以及相關聯之字素，其藉由記錄聲音資料輸入作為一說話者之語音、辨識該聲音資料為對應至一字素標記之資料、以及在該字素標記正確用於該語音輸入而從該說話者取得確認時將該聲音資料與該字素標記相關聯。

15. 如請求項第14項所述之系統，其中該辨識系統包含一機制，該機制接收一名稱形式之語音、記錄對應該名稱之聲音資料、辨識該名稱為該字素標記、以及保持該聲音資料與該字素標記相關聯。

16. 如請求項第14項所述之系統，更包含一構件，用以對於不符合一確信臨界值之語音輸入過濾掉該聲音資料以

及該相關聯之字素標記。

17.一種具有電腦可執行指令之電腦可讀取媒體，該電腦可執行指令被執行時執行至少包含以下之步驟：

從一說話者接收聲音資料；

辨識該聲音資料為一結果以及相關聯之潛在字素序列；

與該說話者確認該結果是否正確符合該聲音資料，若是，則將該聲音資料與對應至該潛在字素序列之一實際字素序列相關聯；若否，則再與該說話者互動直到一結果被確認為正確符合該聲音資料，並關聯該對應的字素序列為該實際字素序列；以及

使用該聲音資料以及關聯之實際字素序列以供後續語音辨識。

18.如請求項第17項所述之電腦可讀取媒體，其中使用該聲音資料及相關聯之字素序列以供後續語音辨識之步驟包含以下步驟：

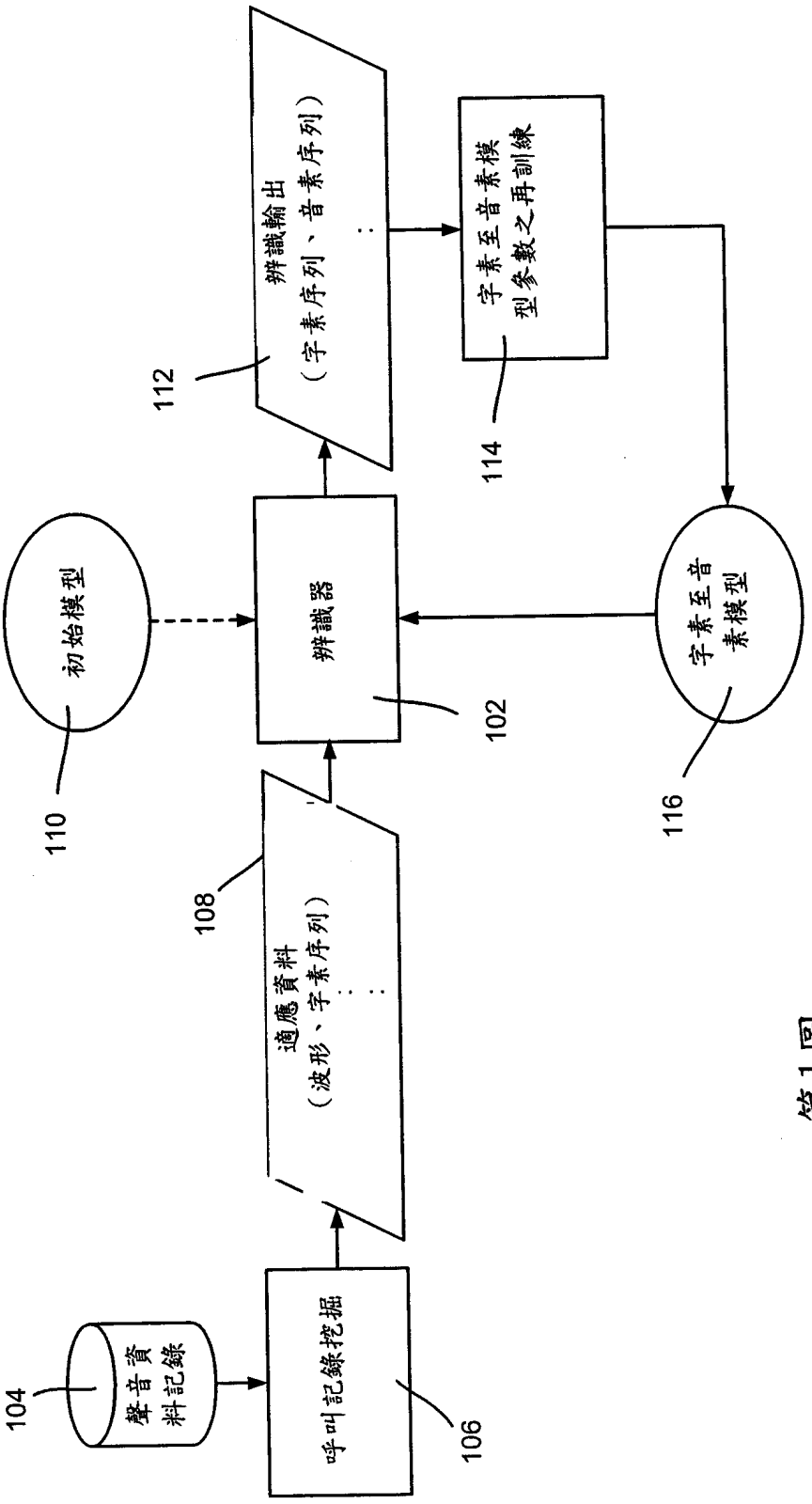
依據該聲音資料以及關聯之字素而再訓練映射於字素序列及音素序列間的一模型。

19.如請求項第18項所述之電腦可讀取媒體，其中該再訓練係使用最大可能性或辨別訓練而完成。

20.如請求項第17項所述之電腦可讀取媒體，其中再與該說話者互動包含再從一說話者接收聲音資料、使用該再接收之聲音資料以判定另一字素、與該說話者確認該其他字素正確符合該再接收之聲音資料、以及將該其他字素

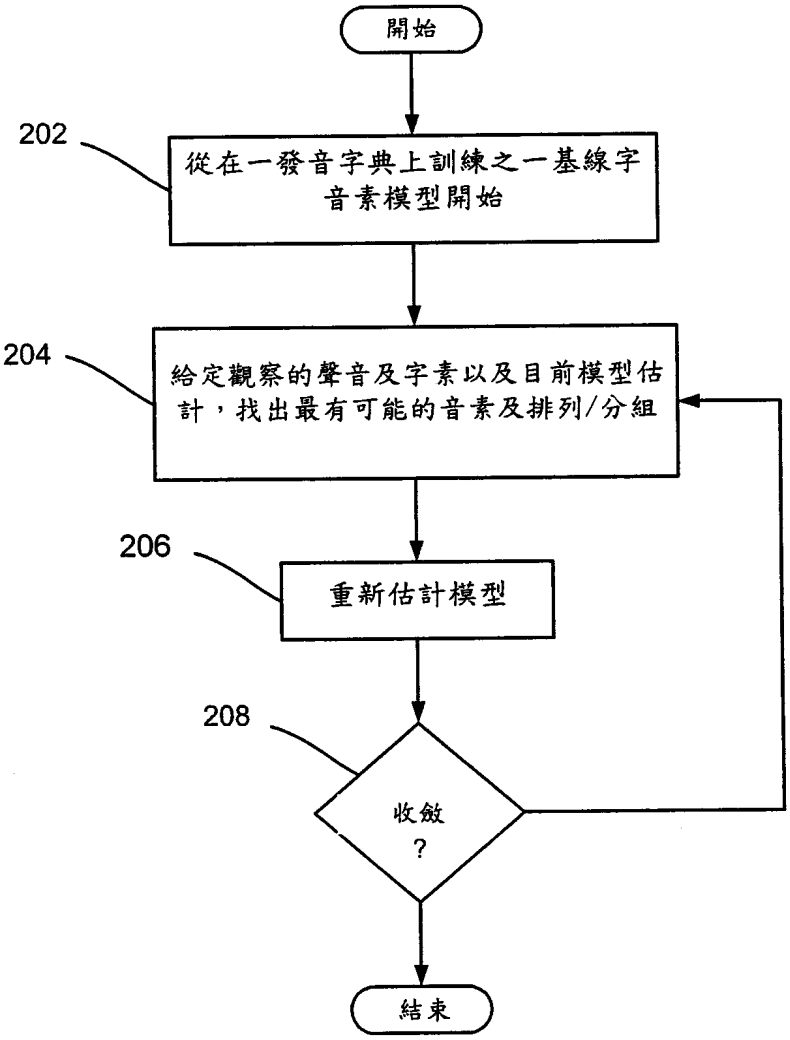
與該聲音資料以及該再接收之聲音資料相關聯。

200926139

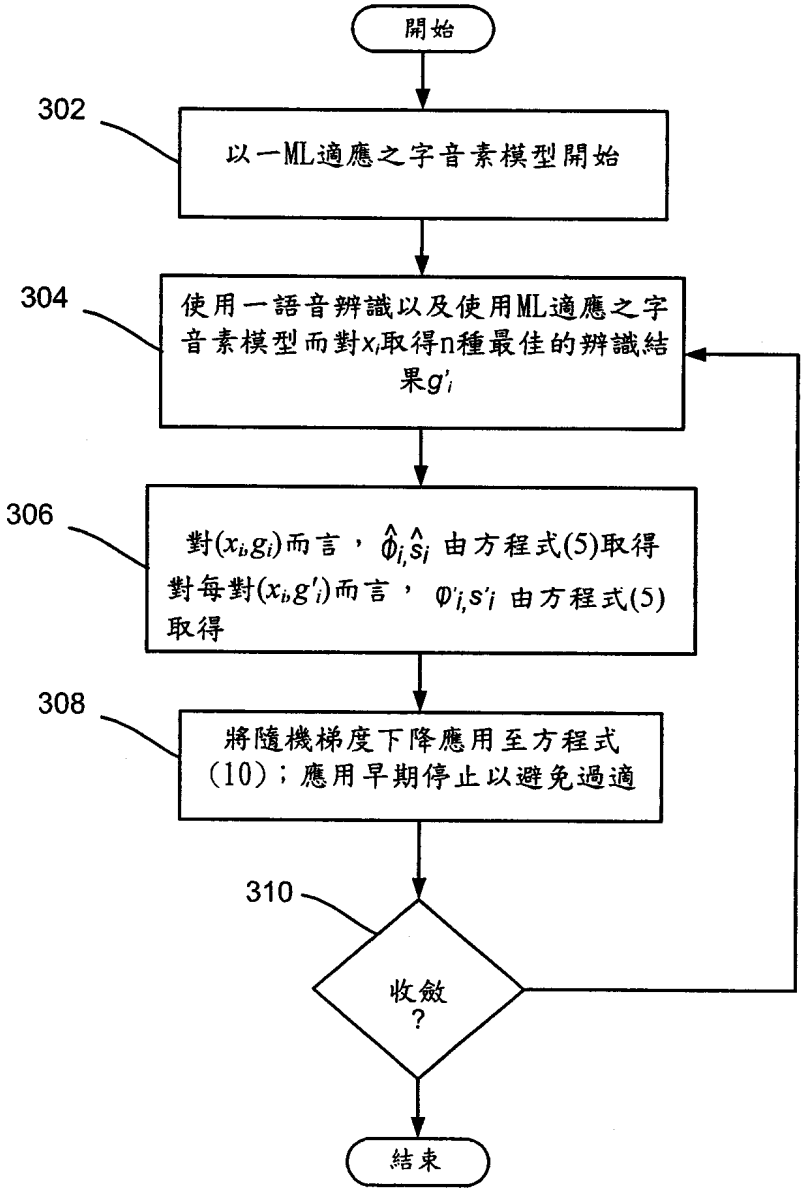


第1圖

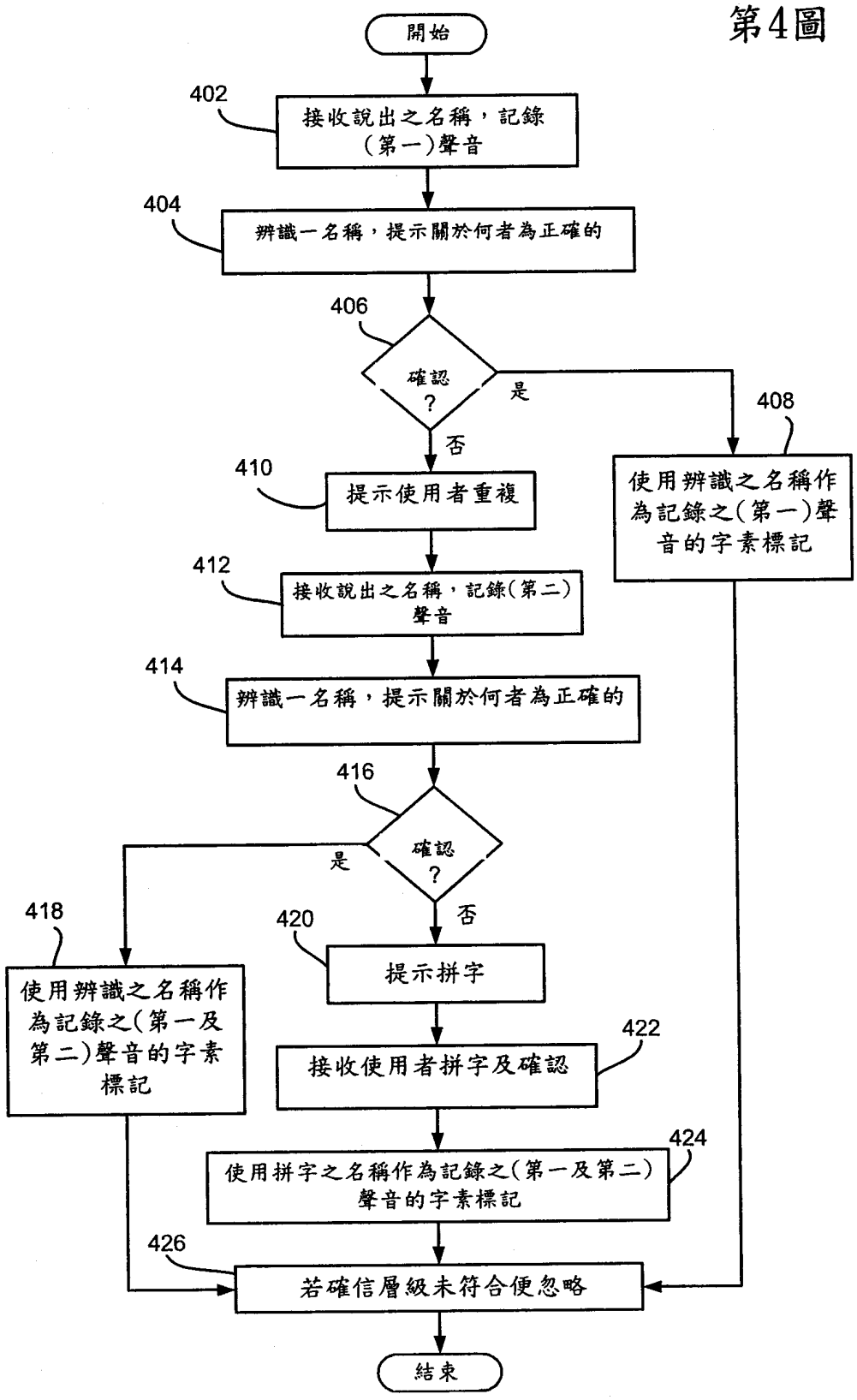
第2圖


開始
202
從在一發音字典上訓練之一基線字音素模型開始
204
給定觀察的聲音及字素以及目前模型估計，找出最有可能的音素及排列/分組
206
重新估計模型
208
收斂？
結束

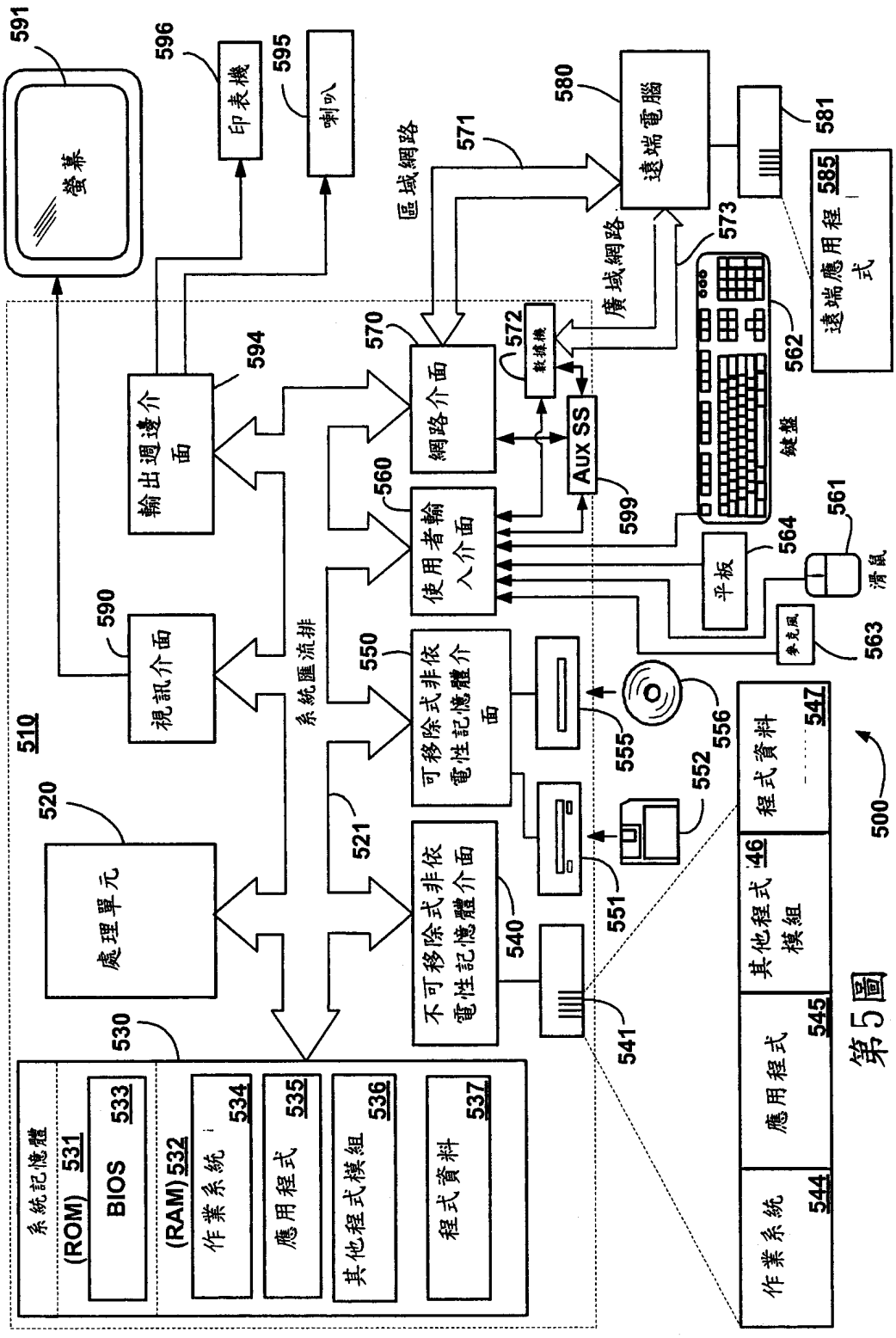
第3圖


開始
302
以一ML適應之字音素模型開始
304
使用一語音辨識以及使用ML適應之字音素模型而對x_i取得n種最佳的辨識結果g'_i
306
對(x_i, g_i)而言，$\hat{\phi}_i, \hat{s}_i$ 由方程式(5)取得
對每對(x_i, g'_i)而言，ϕ'_i, s'_i 由方程式(5)取得
308
將隨機梯度下降應用至方程式(10)；應用早期停止以避免過適
310
收斂？
結束

第4圖


開始
402
接收說出之名稱，記錄(第一)聲音
404
辨識一名稱，提示關於何者為正確的
406
確認?
是
否
408
使用辨識之名稱作為記錄之(第一)聲音的字素標記
410
提示使用者重複
412
接收說出之名稱，記錄(第二)聲音
414
辨識一名稱，提示關於何者為正確的
416
確認?
是
否
418
使用辨識之名稱作為記錄之(第一及第二)聲音的字素標記
420
提示拼字
422
接收使用者拼字及確認
424
使用拼字之名稱作為記錄之(第一及第二)聲音的字素標記
426
若確信層級未符合便忽略
結束

200926139



第5圖

七、指定代表圖：

(一)、本案指定代表圖為：第(1)圖。

(二)、本代表圖之元件代表符號簡單說明：

102 辨識器

104 聲音資料記錄

106 呼叫記錄挖掘

108 適應資料

110 初始模型

112 辨識輸出

114 字素至音素模型參數之再訓練

116 字素至音素模型

八、本案若有化學式時，請揭示最能顯示發明特徵的化學式：

無