Computational Linguistics and Chinese Language Processing
Vol. 15, No. 2, June 2010, pp. 145-160

以最佳化及機率分佈標記形聲字聲符之研究

Annotating Phonetic Component of Chinese Characters Using Constrained Optimization and Pronunciation Distribution

張嘉惠*、林書彥*、李淑瑩*、蔡孟峰*、李淑萍+、廖湘美+、孫致文+、黃鍔#

Chia-Hui Chang, Shu-Yen Lin, Shu-Ying Li , Meng-Feng Tsai,

Shu-Ping Li, Hsiang-Mei Liao, Chih-Wen Sun, and Norden E. Huang

摘要

一般說來，漢字乃圖形文字，無法像英文等拼音文字一樣，一旦學會拼音方法，即有基本的閱讀能力。相對的，漢字讀寫的學習進展則相當緩慢，而且必須搭配注音符號或是其他拼音方法，才可知道每個漢字的發音。事實上漢字中有八成的字是形聲字，形聲字不僅可由形旁表意，又可以聲符表音，因此即使沒見過的字也可以由偏旁推論其音及義。不過主要的困難在於聲旁未必一定同音，可能是相近的發音，之間的演變規則尚未有人探究過，例如：泡、抱、飽三個字同樣與『包』的發音相近，然而發音如何由『包』的發音轉變成其他三個字的發音，則仍待研究。本論文首先嘗試以自動化方式判定漢字聲符，做爲研究形聲字發聲規則的第一步。實驗顯示，我們所提的兩種方式，發音相似度比較法在 9593 個形聲字中的判定聲符準確率爲 90.7%，而構件發聲分佈比較法則

*國立中央大學資訊工程所，台灣桃園縣中壢市中大路 300 號
Dept. of Computer Science and Information Engineering, National Central University, Taiwan
E-mail: chia@csie.ncu.edu.tw
The author for correspondence is Chia-Hui Chang.
+國立中央大學中文系
Dept. of Chinese Literature, National Central University
#國立中央大學數據中心
Research Center for Adaptive Data Analysis, National Central University

可達到 98.1%的準確率，可以加速形聲字聲符標記所需的大量人力工作與時間。

關鍵字: 形聲字、聲符、發音相似度、最佳化、機率分佈、KL divergence

Abstract

Generally speaking, Chinese characters are graphic characters that do not allow immediate pronunciation unless they are accompanied with Mandarin phonetic symbols (zhuyin) or other pinyin methods (e.g. romanization system). In fact, about 80 to 90 percents of Chinese characters are pictophonetic characters which are composed of a phonetic component and a semantic component. Therefore, even if one had not seen the character before, one can make a logical guess at the character's pronunciation and meaning from its phonetic and semantic symbols. In order to analyze such relations, we start by analyzing the characteristics of phonetic components. We found two interesting features that could automatically identify the phonectic components of Chinese characters. One is pronunciation similarity, the other is pronunciation distribution. Experiments show that these two methods have high accuracy (90.8% and 98.1% for 9593 pictophonetic characters) in predicting the phonetic components of pictophonetic characters. These methods can save a lot of time and effort during the annotation of phonetic symbols in the early stage.

Keywords: Picto-phonetic Compounds, Phonetic Component, Pronunciation Similarity, Pronunciation Distribution, Optimization.

1. 簡介

漢語字形及音讀的繁複，向爲初學者及外籍人士所苦，即使會說華語的海外華人對於漢字的認識也可能相當有限。最主要的原因在於漢字乃圖形文字(pictograph system)，無法像英文等拼音文字(alphabet system)一樣，一旦學會拼音方法(phonetic representation)， 即有基本的閱讀能力。相對的，漢字讀寫的學習進展則相當緩慢，而且必須搭配注音符號(Chinese phonetic symbols)或是其他拼音方法，才可知道每個漢字的發音。這樣的限制，對於漢字的學習相當不利，這也是爲什麼二十世紀初期許多中國革命家意欲將漢字拉丁化的主要原因。

漢字的構成包含象形、指事、會意、形聲、轉注、假借(總稱六書(許慎，1988))，其中象形、指事是「造字法」，會意、形聲是「組字法」，轉注、假借是「用字法」。事實上形聲字所占的比例相當高，約佔八、九成。形聲字不僅可由形旁表意，又可以聲符表音，因此即使沒見過的字也可以由偏旁推論其音及義，所謂『有邊讀邊沒邊讀中間』即是此意。不過主要的困難在於聲旁僅代表相近的發音，之間的演變規則尚未有人探究

過，例如：泡、抱、飽三個字同樣與『包』的發音相近，然而發音如何由『包』的發音轉變成其他三個字的發音，則仍待研究。

爲了解漢字中形聲字與其聲符之間發音規則的轉變，我們必須先知道每一個形聲字的聲符。由於形聲字與聲符發音相近，因此我們憶測可用形聲字與其構件之間的發音相似度做爲聲符預測方法，因此我們一方面尋求聲母、韻母之間發音相似度，一方面也建立一個形聲字源標記系統，由中央大學中文所四位研究生與三位教授參與人工標記漢字構形資料庫中 14706 有注音標示的漢字是否爲形聲字以及其聲符構件，做爲預測方法的驗證。另外爲了提升經由發音相似度比較法判斷聲符之準確率，我們也採用限制性最佳化技術，自動求得新的發音相似度分數。

不過發音相似度比較法僅參考單一漢字的資訊，事實上做爲聲符構件的漢字，其衍生字的發聲分佈比非聲符構件的漢字發聲分佈更爲集中。因此我們進一步提出構件發聲分佈比較法，經由計算每個構件的發聲分佈與所有漢字的發聲分佈的 KL 值，做爲此構件做爲聲符的強度代表。實驗結果顯示，發音相似度比較法在 9593 個已標示的形聲字中判定聲符準確率爲 90.8%，而構件發聲分佈比較法則可達到 98.1%的準確率，顯示兩種方法做爲聲符判斷問題的可行性。

雖然形聲字的聲符預測並非計畫最終目的，而且在聲符標注完成後，預測聲符的需求即消失不在，但是透過發音相似度最佳化方法所得的聲母，韻母相似度參數或許有助於未來漢字字音處理的研究，同時部件發音強度也可做爲漢字學習順序之參考，仍有相當的重要性。

2. 相關研究

漢字的使用從殷商的甲骨文算起已達 3,400 年之久，由於結構複雜，因此文字學在中國特別發達。文字學的研究包括文字起源、發展、性質、體系，文字的形、音、義的關係，正字法以及個別文字演變的情況等等議題。爲有系統的研究，中央研究院資訊科學研究所文獻處理實驗室從 1993 年開始，陸續建構古今文字的源流演變、字形結構及異體字表，做爲記錄漢字形體知識的資料庫，也就是漢字構形資料庫(莊德明、謝清俊，2005)。漢字構形資料庫不僅銜接古今文字以反映字形源流演進，也記錄了不同歷史時期的文字結構。另外也由於開發漢字部件檢字系統，得以解決缺字問題。

然而過去的研究著重在字形知識的整理，尚未涉及字音與字義的處理；因此近年來開始文字學入口網站建置計畫(莊德明、謝清俊，2005；莊德明、鄧賢瑛，2008)。一如其文所述：

“漢字構形資料庫目前只著重在字形知識的整理，尚未涉及字音與字義；建立一個形、音、義俱備的漢字知識庫，仍是我們長遠的目標”，因此本計畫“漢語系統音源脈絡之分析”的目的即是以挑戰漢字的發音規則知識庫爲出發，除了了解漢字發音規則外，也希望藉由此項研究找出一套形聲字發音轉換規則，讓華語學習可以在聲符與規則的輔助下，順利讀出字的發音。爲達成此目的，第一步我們必須了解每個漢字是否爲形聲字，

以及了解形聲字聲符的部件，進而解析聲符與最終發音之間轉換的規則。因此我們首先設計一個“形聲字聲符標記系統”，由中文系研究生與教授的協助，進行形聲字與其聲符的標記。不過由於此過程需耗費大量時間與人力投入（在 2009/11/10 至 2010/11/23 期間內共有 9593 字由至少三位研究生標記相同聲符），因此是否存在自動判定形聲字聲符的方法，變成本計畫第一個挑戰。

3. 發音相似度比較公式表

一般說來，聲符構件通常與原字的發音相似度高於非聲符構件與原字的發音相似度，舉例來說，“詁”字與其聲符構件”古”發音相同，而與其非聲符構件“言”發音較不相似，又如“校”字與其聲符構件”交”發音相近，而與其非聲符構件“木”發音較不相似。因此發音相似度可以做爲我們判定一個形聲字聲符的重要依據。爲此我們必須有一套漢字發音相似度的計算方法。

每一個單獨的漢字雖爲單音節發音，但是就聲韻學上來看可分爲聲母、韻母與調性三類。聲母是使用在韻母前面的輔音，隨著發音部位與發音方法而有所不同，如表一所示；而韻母則是一個音節中的元音（母音），也是押韻的主要部份；再者由於漢語本身是具備聲調系統的語言，因此我們在計算一個形聲字與其構件的發音相似度時即可以聲母、韻母、及聲調的相似度做爲判斷發音相似度的依據。

表一、聲母的發音部位與發音方法

發音部位		上阻	上唇	上齒	齒背	上齒齦	前硬顎		軟顎
發音方法		下阻	下唇		舌尖		舌尖後	舌面前	舌面後
狀態	聲帶	簡稱 氣流	雙唇	唇齒	舌尖前	舌尖中	舌尖後	舌面前	舌根
塞	清	不送氣	ㄅb [p]			ㄉd [t]			ㄍg [k]
		送氣	ㄆp [pʰ]			ㄊt [tʰ]			ㄎk [kʰ]
塞擦	清	不送氣			ㄗz [ts]		ㄓzh [tʂ]	ㄐj [tɕ]	
		送氣			ㄘc [tsʰ]		ㄔch [tʂʰ]	ㄑq [tɕʰ]	
擦	清			ㄈf [f]	ㄙs [s]		ㄕsh [ʂ]	ㄒx [ɕ]	ㄏh [x]
	濁						ㄖr [ʐ]		
鼻	濁		ㄇm [m]			ㄋn [n]			
邊	濁					ㄌl [l]			

3.1 人工制定聲母和韻母發音相似度

如表一所示，聲母依發音部位的不同，可分爲雙唇、唇齒、舌尖前、舌尖中、舌尖後、舌面前、舌根七種音；若依發音方法的不同，則可分爲塞音、塞擦音、擦音、鼻音、邊音等五類，加以聲帶的清濁不同，我們制訂如下之聲母與聲母之間發音相似度：

相同的聲母，相似度訂爲 ub=1。

國際拼音相同，但是差一個上標號，相似度訂爲 a=0.9。

發音部位相同（同列），相似度訂爲 b=0.8。

發音方法相同（同行），相似度訂爲 c=0.5。

其他情形（如不同行不同列或是沒有聲母的情形），相似度訂爲 0.1。

同時我們也尋求語言學專家的協助，針對此版本的制訂規則提供意見，專家們對聲母部分提出以下兩項的建議：

ㄈ與ㄅㄆㄇ的關係合併爲同一行，隸屬唇音一大類。相似度訂爲 b。

ㄗㄘㄙ、ㄓㄔㄕ與ㄐㄑㄒ三大類(不包括ㄖ)，除了同屬塞擦音、擦音外，在發音上有許多人有彼此相混的現象，所以相似度定爲 d=0.7。

表二、韻母種類

種類	注音符號
單韻母	(ㅍ)、ㄧ(yi)、ㄨ(wu)、ㄩ(yu)、ㄚ(a)、ㄛ(o)、ㄜ(e)、ㄝ( )
複韻母	ㄞ(ai)、ㄟ(ei)、ㄠ(ao)、ㄡ(ou)
聲隨韻母	ㄢ(an)、ㄣ(en)、ㄤ(ang)、ㄥ(eng)
捲舌韻母	ㄦ(er)
結合韻母	ㄧㄚ(ya)、ㄧㄛ(yo)、ㄧㄝ(ye)、ㄧㄞ(yai)、ㄧㄠ(yau)、ㄧㄡ(you)、ㄧㄢ(yan)、ㄧㄣ(yin)、ㄧㄤ(yang)、ㄧㄥ(ying)
	ㄨㄚ(wa)、ㄨㄛ(wo)、ㄨㄞ(wai)、ㄨㄟ(wei)、ㄨㄢ(wan)、ㄨㄣ(wen)、ㄨㄤ(wang)、ㄨㄥ(weng)
	ㄩㄝ(yue)、ㄩㄢ(yuan)、ㄩㄣ(yun)、ㄩㄥ(yong)

接著我們制訂韻母與韻母的發音相似度。雖然每個韻母在注音符號表中只是單一個符號，但是我們可以進一步將韻母分爲單韻母、複韻母、聲隨韻母、捲舌韻母與結合韻母五種。其中複韻母包括兩個母音，聲隨韻母包含韻母後的輔音，而結合韻母則包含兩個韻母。依據以上分類，我們制訂韻母發音相似度如下：

韻母相同則相似度訂爲 1。

若兩個韻母同爲結合韻母，若兩者有共同尾音則設相似度爲 x=0.8，若兩者有共同第一個音則設相似度爲 y=0.5，否則設相似度爲 lb=0.1。例如：ㄧㄚ和ㄨㄚ有共同尾音ㄚ因此相似度爲 x、ㄨㄤ和ㄨㄢ有相同第一個韻母因此相似度爲 y；ㄧㄞ和ㄨㄟ因無共同部分，設相似度爲 lb。

若兩個韻母均非結合韻母（也就是屬於單韻母、複韻母、聲隨韻母或捲舌韻母），則相似度以國際拼音決定。若拼音出現相同字母則設相似度爲 z=0.5，否則設相似度爲 lb。例如：ㄧ(i)和ㄞ(ai)有相同部分(i)、ㄠ(au)和ㄡ(ou)間有相同部分(u)皆設相似度爲 z。

若一則爲結合韻母，一則非結合韻母，則給分方式如下所示：若單韻母出現在結合韻母中的後面位置設相似度爲 x=0.8，若出現在前面則設相似度爲 y=0.5，否則相似度爲 lb。例如:ㄧ出現在ㄧㄚ前面位置，則設相似度爲 y=0.5；而ㄞ出現在ㄧㄞ後面位置，設相似度則爲 x=0.8。

其他情形，相似度訂爲 lb。

根據上述規則所制訂的聲母與韻母相似度表分別列於附錄 A 及 B。給定以上相似度，我們定義兩個漢字發音相似度爲其聲母相似度、韻母相似度的總和：

$$\text{Similarity}(c1,c2) = \text{Initial}(c1,c2) + \text{Vowel}(c1,c2) \quad (1)$$

因此給定一個形聲字，我們依據漢字構詞資料庫所拆解成的二至三個構件，分別計算這些構件與原本漢字的發音相似度，查閱聲母與韻母的相似度比較公式表，求算聲母與韻母的總和，取相似度大者構件，做爲聲符的預測。因此形聲字 w 的聲符即可選取與 w 發聲最爲相似的構件 c。

$$PC(w) = \arg\max_{c \in w} Similarity(w,c) \quad (2)$$

舉例來說，漢字「校」(ㄒㄧㄠ 4)的構件爲「木」(ㄇㄨ 4)和「交」(ㄐㄧㄠ 1)，採用相似度比較公式表求算「木」和「 校 」的分數，其中聲母不同行不同列，相似度爲 0.1，而韻母無共同音因此相似度則爲 lb，故相似度總和爲 0.1+lb =0.2；同理「交」和「校」的聲母發音部位相同，相似度爲 b，且韻母同爲結合韻母，相似度爲 1，故相似度總和爲 b+1 =1.8，因此系統判定「交」爲「校」的聲符。若各構件的總和皆相同，則加入調性進行校正，選擇與原字調性相同的構件做爲預測聲符。舉例而言，漢字「祖」(ㄗㄨ 3)的構件爲「示」(ㄕ 4)和「且」(ㄑㄧㄝ 3)，採用相似度比較公式表求算「示」和「祖」的相似度分數爲 d+0.1=0.8，「且」和「祖」的總和爲 d+lb=0.8，兩者相似度分數相同，因此我們再加上調性判別，預測與「祖」聲調相同的「且」爲聲符 。

3.2 發音相似度最佳化

由於前述發音相似度比較公式表是由人工制訂，這些值是否能有效的做爲聲符預測的參數，還是有更佳的值可以推測形聲字聲符？另外，聲母、韻母及聲調三者所占的比重爲何？則是本節所要探討的問題。我們嘗試採用限制型最佳化方法計算聲母和韻母之發音相似度。假設一組已知聲符的形聲字 T，依照發音相似度比較公式，我們可以爲每一個形聲字 w∈T 列出 w 的聲符構件與原字發音相似度必須大於非聲符構件與原字發音相似度的限制條件。以前例漢字「校」來說，其構件爲「木」和「交」，而其已知聲符爲「交」，因此 Similarity(木,校)≤Similarity(交,校)，也就是 0.1+lb ≤ b+1。

在我們的問題中，可以將聲母發音相似度參數 ub, a, b, c, d 以及韻母發音相似度參數 x, y, z, lb， 拿來做爲最佳化問題中的變數。由於當限制條件多於變數個數時，系統可能無解，因此我們對每個不等式的聲符部份加上一個額外的變數 $\varepsilon_i \geq 0$，也就是 d+0.1≤ b+1+ ε_i，再以 $\sum_i \varepsilon_i^p$ 做爲最小化的目標函數，確保聲符與原字的發音相似度大於非聲符構件

與原字的相似度。舉例而言，若是聲符與原字的相似度小於非聲符構件與原字的相似度，則 ε_i 必須大於 0 才足以讓條件成立，反之若聲符與原字的相似度已大於非聲符構件與原字的相似度，則 ε_i 在最小化的目標下自然會是 0。因此若有 m 個已知聲符的漢字，則可化爲以下最佳化問題：

$$\min \sum_i \varepsilon_i^p \text{ s.t. } \begin{cases} 0.1 + lb <= ub + 1 + \varepsilon_1 \\ 0.1 + y <= ub + 1 + \varepsilon_2 \\ \text{M} \\ b + y <= a + 1 + \varepsilon_{|T|} \\ a, b, c, d, x, y, z, ub, lb, \varepsilon_i >= 0 \end{cases} \tag{3}$$

其中 p > 0，代表錯誤聲符對系統的處罰程度的不同。在最極端的情形下，我們可以針對 21 個聲母之間的相似度設定 22*21/2 個變數（含聲母空的情形），同理也可對 39 個韻母（含空韻）之間的相似度設定 39*38/2 個變數，再以限制型最佳化的方法來找出對聲符預測最有利的相似度分數。但由於會有將近 1000 個變數，所花的時間相對也比較長。

機率分佈比較法

除了前述兩項發音相似度比較公式表與最佳化分析的方法，我們也從另一個角度觀察漢字的發音，我們發現某些漢字構件有較強的發音強度，常常做爲聲符，而屬於部首的構件，則通常代表字的形意。於是我們假設漢字的發音有可能是由其構件發音強度較高的構件所支配。因此，如何制定構件的發音強度而又不耗費大量的人力是我們的首要目標。我們觀察一些常見的漢字，如表三所示。從中不難發現包含構件「包」的漢字的發音不管在聲母、韻母或聲調的表現一致性較高，而包含構件「火」的漢字則較低，一致性較高的構件我們也可以說它是發音集中在某幾種發音上，反之較爲分散。集中度較高的構件就很有可能支配著包含此構件的漢字發音，也就是構件發音強度較強。

假設 S 代表某些漢字所形成的集合，f(S)、g(S)、h(S)分表示其聲母、韻母及聲調的分佈機率。令 A 表示所有漢字所成的集合，則f(A)、g(A)、h(A)分別表示漢字的聲母、韻母及聲調的分佈機率。同理對於一個漢字構件 b，我們可以找出包含 b 的所有漢字 B，同時求得其聲母、韻母及聲調的分佈機率f(B)、g(B)、h(B)。若是 b 發音集中度較高，則其聲母分佈f(B)與f(A)就會有較大的差異。因此我們採用 Kullback–Leibler divergence 的方法來計算兩個分佈的距離。Kullback–Leibler divergence 的公式如下:

$$KL(\text{P}\|\text{Q}) = \sum_i \text{P(i)} \log \frac{\text{P(i)}}{\text{Q(i)}} \tag{4}$$

因此我們可以計算 KL(f(B) || f(A))做爲構件 b 聲母強度，同理計算 KL(g(B) || g(A)) 做爲 b 韻母強度，以及計算 KL(h(B) || h(A)) 做爲聲調強度。以下我們以調號以下我們以聲母爲例，計算構件「火」及「包」的聲母強度，我們從漢字構詞資料庫中找出所有標示注音的字共|A| = 14598；我們同時統計含有構件「火」的字共有|B|=259，含有構件「包」的字共有|C|=32，其聲母分佈如下:

表三、全部漢字A 及包含構件「火」B 的聲母分佈狀況

	ㄅ	ㄆ	ㄇ	ㄈ	ㄉ	ㄊ	ㄋ	ㄌ	ㄍ	ㄎ	ㄏ
所有漢字 |A|=14958	660	445	564	445	642	630	312	1009	587	380	742
f(A)	0.04	0.03	0.03	0.03	0.04	0.04	0.02	0.06	0.04	0.02	0.05
含構件火 |B|=259	8	5	10	7	6	13	2	18	7	6	25
f(B)	0.03	0.01	0.03	0.02	0.02	0.05	0.007	0.06	0.02	0.02	0.09
含構件包 |C|=32	15	16	0	1	0	0	0	0	0	0	0
f(C)	0.46	0.5	0	0.03	0	0	0	0	0	0	0

	ㄑ	ㄒ	ㄓ	ㄔ	ㄕ	ㄖ	ㄗ	ㄘ	ㄙ	空
所有漢字 |A|=14958	719	931	835	540	567	239	371	294	356	2120
f(A)	0.04	0.06	0.05	0.04	0.03	0.01	0.02	0.02	0.02	0.14
含構件火 |B|=259	8	27	14	11	8	8	8	4	1	45
f(B)	0.03	0.1	0.05	0.04	0.03	0.03	0.03	0.015	0.003	0.17
含構件包 |C|=32	0	0	0	0	0	0	0	0	0	0
f(C)	0	0	0	0	0	0	0	0	0	0

將|A|與|B|正規化得 f(A)與 f(B)兩機率分佈。最後將 f(B)及 f(A)代入 KL-divergence 公式可得構件「火」的聲母強度。同樣方式可計算「包」的聲母強度。

當我們要判斷某個漢字的聲符時，只需要將此漢字的所有構件的聲母、韻母及聲調三種 Kullback–Leibler divergence 的值做加總，那麼擁有最大的加總值的構件便會被我們判定爲此漢字的聲符。這個方法的好處在於方法簡單而且不需訓練過程，後續實驗將可看出這個方法對形聲字聲符判斷相當有效。表四分別列出依聲母、韻母、聲調 KL 值與其構件出現次數|B|的乘積值排序前十名構件。這樣子的排序可以做爲聲符學習的參考順序。

表四、漢字構件發音強度表

字碼	聲母 KL 值	字碼	韻母 KL 值	字碼	聲調 KL 值
分	0.9768	非	1.3864	皇	0.4851
莫	1.3439	分	1.1621	盧	0.5487
非	0.9789	令	1.4116	令	0.2977
令	1.0335	票	1.4535	會	0.4988
元	1.7167	莫	1.439	夷	0.5487
票	1.1263	屯	1.6778	希	0.5458
卑	1.0746	龍	1.1123	余	0.3386
弗	1.4473	皇	1.6968	吉	0.3332
方	0.9731	包	1.3036	肖	0.2747
俞	1.0632	同	1.4166	世	0.4988

4. 實驗

以下實驗探討前二節所提出的三種分析方法：發音相似度比較公式表、發音相似度最佳化與機率分佈比較法，分別在自動判別形聲字之聲符的效能爲何。實驗中使用的測試資料集，是取自漢字構形資料庫中有注音標示的漢字，在 2009/11/10 至 2010/11/23 期間內共有 9593 字爲四位中文系研究生共同標記完成。

4.1 發音相似度比較法

首先我們進行發音相似度比較公式的預測效能。如表五所示，原始發音相似度方法，準確率約 88.67%，其中包含 518 筆無法判別的字；加入聲調的判別，無法判別的字減少至 291，準確率約 90.21%。採用專家建議的修正版之發音相似度比較公式表來測試，準確率提高至 89.39%，再加入聲調後則爲九成零七的準確率。顯示以發音相似度比較公式進行判別聲符，有一定的效果。也因爲考慮聲調可以提升一個多百分點，因此我們將調的相似度直接納入發音相似度的計算，將計算式(1)改爲如下計算式：

$$\text{Similarity}(c1,c2) = \text{Initial}(c1,c2)+\text{Vowel}(c1,c2)+\text{Tone}(c1,c2) \quad (5)$$

其中聲調相似度依兩個發音的聲調是否相同，給予δ及 0 的相似度。參數δ的值則由分析每一個字的 pcdiff 的分佈圖而得。方法如下：對於每一個字 w，我們先用方程式(1)計算 w 與其聲符構件 pc_w 的發音相似度，以及 w 與其非聲符構件的發音相似度（若有多個非

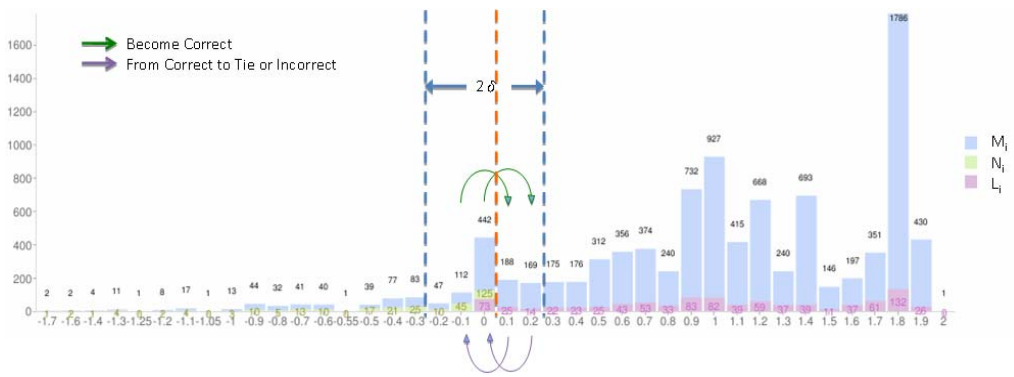
聲符構件則取發音相似度值較大者），並定義 pcdiff(w)爲兩者的差：

$$pcdiff(w) = Similarity(w, pc_w) - \underset{c \in w \setminus pc_w}{similarity(w, c)} \quad (6)$$

表五、發音相似度比較法判別準確率

	正確	錯誤	無法判別	準確率
原始版(no d)	8507	568	518	88.67
原始版＋調	8654	648	291	90.21
修正版(d=0.7)	8576	575	442	89.39
修正版+調	8701	648	244	90.70
整合版(δ=0.2)	8707	618	268	90.86

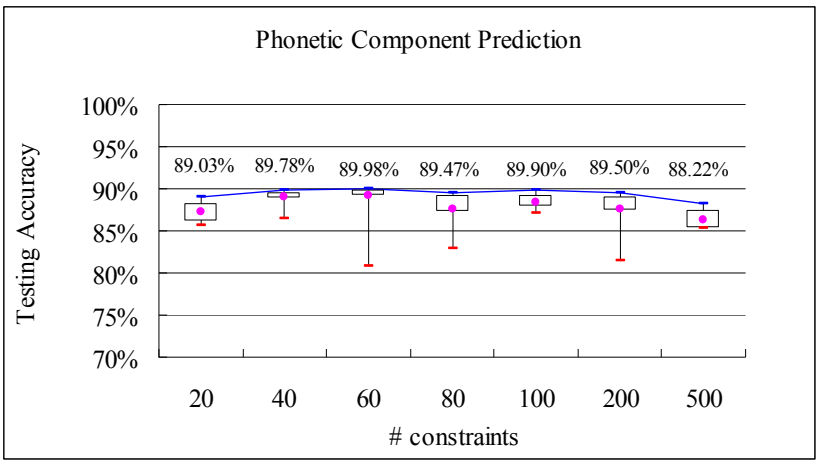
圖一爲所有字的 pcdiff 分佈圖（Histogram）。每一個藍色長條圖 M_i 代表 pcdiff 爲 i（i=…, -0.2, -0.1, 0, 0.1, 0.2, ….）的字的個數（黑色字），同時我們也統計每一藍色長條中有多少個字的聲調與其聲符構件相同但與非聲符構件不同，我們用 N_i 來表示（綠色字）；另外我們也統計有多少個字的聲調與其聲符構件不同但與非聲符構件相同，我們用 L_i 來表示（紫色字）。這兩部份的字集分別代表的是在採用方程式(5)來計算發音相似度時，pcdiff 會增加或是減少的字數。因此如果我們採用方程式(5)來計算發音相似度，將會有新增加 $N_0+N_{-0.1}+\cdots+N_{-\delta+0.1}$ 可被正確預測的字（橘色轉換），但是同時會有 $L_{0.1}+L_{0.2}+\cdots+L_{\delta}$的字會從正確預測轉爲錯誤預測或無法判斷（藍色轉換）。因此若 N_0=125, L_0=73, $N_{-0.1}$=45, $N_{-0.2}$=10, $L_{0.1}$=25, $L_{0.2}$=14（如圖二），當我們設δ=0.2，則我們可新增 125+45=170 個可正確判斷的字，但失去 25+14=39 原可正確判斷的字。整體來說，我們可多預測 170-39=131 個正確的字，而無法判別的字則減少 125+73-10-14＝174 個字(442-174=268)。



圖一、聲符發音相似度與非聲符發音相似度差分佈圖

4.2 發音相似度最佳化

第二部份實驗主要是了解最佳化方法計算出的聲母與韻母相似度參數在判別漢字聲符之可能性。實驗主要目的在了解需要多少訓練資料筆數才能達到大約九成的準確率，以及最佳化方法的學習曲線。由於變數不多，理論上我們所需要的訓練資料筆數並不需要太多，但是若是取太少訓練資料，則預測準確的效果差異則會變大。我們取 p=1 做爲目標函數，隨機抽取 1000 筆資料做爲訓練資料，剩餘 8593 筆資料做爲測試資料。每次從訓練資料中隨機抽取 m（=20, 60, 80, 100, 200, 500）個漢字產生 m 個限制條件（公式 1），再由線性規劃方法計算出參數值，並利用這些參數來檢視所有 8593 筆測試漢字聲符判別之準確率。我們反覆十次上述的實驗，用以繪製盒鬚圖(Box-Whisker Plot)得到如圖二之結果。



圖二、以發音相似度最佳化作爲聲符預測準確圖

如圖二所示，我們可以藉由最佳化方法找到與前節所得準確率相當的相似度參數，甚至於在少數已知聲符的漢字訓練資料，如 40 或 60 筆時，即有相當好的結果，超越修正版 89.39%的準確率，所得參數值與人工定義的參數設定方式接近，如聲母參數 ub>a>b>c,d>0.1 等關係，韻母參數 1>x>y,z>lb 等關係（如表六所示）。事實上聲符預測準確率最高 89.98%，可由 60 筆已知聲符的形聲字所產生的限制條件求出，然當訓練資料（限制條件）增加時，準確率並無上升的情形，甚至在 500 時，反而有下降趨勢（最佳測試準確率降至 87%及 81.5%）。推測主要的原因在於聲符與形聲字發音相似度大於非聲符構件與形聲字發音相似度僅是一個通則，仍有相當多例外的情形。例如，冶、洛、債、時、枸、荼等字，前二者產生 $y<=lb+\varepsilon$ 的限制，中二者則產生 $b<=d+\varepsilon$ 的限制，後二者則產生 $z<=lb+\varepsilon$ 的限制條件。在少量的訓練資料時，這些限制條件的影響可能只是影響一二個變數大小關係，但是在較多的訓練資料時，將可能同時影響數個變數，最終是聲母參數均相同（ub=a=b=c=d），韻母參數均相同（x=y=z=lb）時，反而可以讓目標函數的值較小。因此我們看到當訓練資料 500 筆時，即使是最佳一組測試準確率，也僅得聲母參數均相同的結果。

表六、不同筆訓練資料所得最佳測試準確率及其聲母、韻母參數

訓練資料準確率	90.0%	98.0%	98.0%	96.0%	94.0%	98.0%	97.0%
測試資料準確率	89.0%	89.8%	90.0%	89.5%	90.0%	90.0%	88.0%
ub	0.97	0.82	0.94	0.93	1.00	0.66	0.93
a	0.73	0.55	0.63	0.73	0.84	0.89	0.10
b	§0.10	§0.10	0.60	0.76	§0.37	0.50	0.10
c	§0.46	§0.67	0.00	0.10	§0.42	0.10	0.10
d	§0.10	§0.40	0.55	0.19	§0.58	0.10	0.10
x	0.63	0.70	0.72	0.83	0.64	0.89	0.93
y	0.10	¶0.22	¶0.38	¶0.33	¶0.16	0.18	0.10
z	0.10	¶0.34	¶0.51	¶0.52	¶0.41	♮0.10	♮0.10
lb	0.10	0.16	0.10	0.13	0.10	♮0.13	♮0.12

Affected by constraints § b<=c or d, ¶ y<=z, ♮z<=lb

4.3 機率分佈比較法

第三部份的實驗目的在於了解機率分佈比較法對於在判別漢字的聲符之效能。如表七所示，以聲母分佈、韻母分佈以及聲調分佈個別強度做爲聲符預測，都有一定的準確度（分別爲 92.8%、97.5%及 95.9%），其中又以韻母的分佈是三者當中最爲有效方式。整體來說，三種分佈一起考量的結果有最佳的效果，針對 9593 筆形聲字，其中有 9413 筆正確，180 筆錯誤，準確率 98.1%。

表七、機率分佈比較法判別準確率。

	正確	錯誤	準確率
聲	8910	683	92.80
韻	9362	231	97.50
調	9207	386	95.90
聲+韻+調	9413	180	**98.10**

爲了解判定錯誤發生的可能原因，我們列出錯誤例子如表八。這些例子顯示，機率分佈比較法發生錯誤多在在於構件間的強度太過接近，尤其是兩個構件均爲部首的情形，然而若就發音相似度比較法而言應該可以正確判斷其聲符。因此也可以考慮結合機率分佈比較法與發音相似度比較法，對於形聲字進行聲符的判斷測試。

表八、機率分佈比較法錯誤例子

word	Phonetic notation	Elements	Correct Phonetic Symbol	Sum of KL divergence of each element
扣	ㄎㄡ4 (kou)	扌口	口	扌 0.0756 > 口 0.0692
沐	ㄇㄨ4 (mu)	氵木	木	氵 0.0284 > 木 0.0218
孟	ㄇㄥ4 (meng)	子皿	皿	子 0.6303 > 皿 0.2763
忝	ㄊㄧㄢ3 (tian)	天小	天	天 1.2455 < 小 1.8219
所	ㄙㄨㄛ3 (suo)	戶斤	戶	戶 0.7605 < 斤 0.9602
旺	ㄨㄤ4 (wang)	日王	王	日 0.1051 > 王 0.0904

5. 結論及未來研究

本篇論文主要目的是藉由對形聲字的分析研究，找出漢字與其聲符構件原字之間的關係。在第一階段我們針對漢字形聲字聲符的標記，除了採用中文系研究生的人力標記之外，同時也提出三種自動判別的方式，用以加速形聲字聲符的標記工作。實驗顯示，不論是人工制訂的相似度參數或是最佳化方式所得的參數，預測準確率大約可以做到九成。這是否是本篇論文所使用的最佳化方法的不足，又或是發音相似度比較法對於聲符預測的極限，仍尚待進一步的研究。另外一方面，以每個部件發音的分佈集中與否做爲聲符的判斷，則有高達九成八的準確率，則是相當有用的資訊。

雖然形聲字在聲符標注完成後，預測聲符的需求即消失不在，但是透過發音相似度最佳化方法所得的聲母，韻母相似度參數或許有助於未來漢字字音處理的研究，同時部件發音強度也可做爲漢字教學順序參考，仍有相當的重要性。未來我們將以持續以挑戰漢字的發音規則知識庫爲出發，除解析漢字發音規則外，也希望以此項研究爲出發，發展一套漢字學習的順序，讓使用者可用較少的學習時間，有效率認識更多漢字。

致謝

本論文的完成感謝陳怡如、葉博榮、鍾哲宇、趙婕妤等人的幫助。

參考資料

許慎撰，段玉裁注(1988)。《說文解字注》，台北藝文印書館。

莊德明、謝清俊(2005，1 月)。漢字構形資料庫的建置與應用，漢字與全球化國際學術研討會，台北。

莊德明、鄧賢瑛(2008，8 月)。文字學入口網站的規畫，第四屆中國文字學國際學術研討會，山東煙台。

附錄A:聲母相似度比較表

聲	ㄅ	ㄆ	ㄇ	ㄈ	ㄉ	ㄊ	ㄋ	ㄌ	ㄍ	ㄎ	ㄏ	ㄐ	ㄑ	ㄒ	ㄓ	ㄔ	ㄕ	ㄖ	ㄗ	ㄘ	ㄙ
ㄅ	1	0.9	0.8	0.1/0.8	0.5	0.1	0.1	0.1	0.5	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄆ		1	0.8	0.1/0.8	0.1	0.5	0.1	0.1	0.1	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄇ			1	0.1/0.8	0.1	0.1	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄈ				1	0.1	0.1	0.1	0.1	0.1	0.1	0.5	0.1	0.1	0.5	0.1	0.1	0.5	0.1	0.1	0.1	0.5
ㄉ					1	0.9	0.8	0.8	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄊ						1	0.8	0.8	0.1	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄋ							1	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄌ								1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄍ									1	0.9	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄎ										1	0.9	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄏ											1	0.9	0.1	0.1	0.5	0.1	0.1	0.1	0.1	0.1	0.5
ㄐ												1	0.9	0.8	0.1/0.7	0.1/0.7	0.1/0.7	0.1/0.7	0.5/0.7	0.1/0.7	0.1/0.7
ㄑ													1	0.8	0.1/0.7	0.5/0.7	0.1/0.7	0.1/0.7	0.1/0.7	0.5/0.7	0.1/0.7
ㄒ														1	0.1/0.7	0.1/0.7	0.5/0.7	0.1/0.7	0.1/0.7	0.1/0.7	0.5/0.7
ㄓ															1	0.9	0.8	0.8	0.5/0.7	0.1/0.7	0.1/0.7
ㄔ																1	0.8	0.1/0.7	0.1/0.7	0.5/0.7	0.1/0.7
ㄕ																	1	0.8	0.1/0.7	0.1/0.7	0.5/0.7
ㄖ																		1	0.1/0.7	0.1/0.7	0.1/0.7
ㄗ																			1	0.9	0.8
ㄘ																				1	0.8
ㄙ																					1

附錄B:韻母相似度比較表

韻母		ㄓ 空韻	ㄧ i	ㄨ u	ㄩ y	ㄚ a	ㄛ o	ㄜ ɤ	ㄝ e	ㄞ ai	ㄟ ei	ㄠ au	ㄡ ou	ㄢ an	ㄣ ən	ㄤ aŋ	ㄥ əŋ	ㄦ ər	ㄧㄚ ia	ㄧㄛ io	ㄧㄝ ie	ㄧㄞ iai	ㄧㄠ iau	ㄧㄡ iou	ㄧㄢ ian	ㄧㄣ iən	ㄧㄤ iaŋ	ㄧㄥ iəŋ	ㄨㄚ ua	ㄨㄛ uo	ㄨㄞ uai	ㄨㄟ uei	ㄨㄢ uan	ㄨㄣ uən	ㄨㄤ uaŋ	ㄨㄥ uəŋ	ㄩㄝ ye	ㄩㄢ yan	ㄩㄣ yən	ㄩㄥ yəŋ
ㄓ	空韻	1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄧ	i		1	0.1	0.1	0.1	0.1	0.1	0.1	0.5	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄨ	u			1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.5	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.1	0.1	0.1	0.1
ㄩ	y				1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.5	0.5	0.5	0.5
ㄚ	a					1	0.1	0.1	0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄛ	o						1	0.1	0.1	0.1	0.1	0.1	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄜ	ɤ							1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄝ	e								1	0.1	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1
ㄞ	ai									1	0.5	0.5	0.1	0.5	0.1	0.5	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄟ	ei										1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄠ	au											1	0.5	0.5	0.1	0.5	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄡ	ou												1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄢ	an													1	0.5	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.8	0.1	0.1
ㄣ	ən														1	0.1	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.8	0.1
ㄤ	aŋ															1	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.1	0.1
ㄥ	əŋ																1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.8
ㄦ	ər																	1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄧㄚ	ia																		1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄧㄛ	io																			1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.1	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄧㄝ	ie																				1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄧㄞ	iai																					1	0.5	0.5	0.5	0.5	0.5	0.5	0.1	0.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄧㄠ	iau																						1	0.5	0.5	0.5	0.5	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄧㄡ	iou																							1	0.5	0.5	0.5	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
ㄧㄢ	ian																								1	0.5	0.5	0.5	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.8	0.1	0.1
ㄧㄣ	iən																									1	0.5	0.5	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.8	0.1
ㄧㄤ	iaŋ																										1	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1
ㄧㄥ	iəŋ																											1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.8
ㄨㄚ	ua																												1	0.5	0.5	0.5	0.5	0.5	0.5	0.1	0.1	0.1	0.1
ㄨㄛ	uo																													1	0.5	0.5	0.5	0.5	0.5	0.1	0.1	0.1	0.1
ㄨㄞ	uai																														1	0.5	0.5	0.5	0.5	0.1	0.1	0.1	0.1
ㄨㄟ	uei																															1	0.5	0.5	0.5	0.1	0.1	0.1	0.1
ㄨㄢ	uan																																1	0.5	0.5	0.1	0.8	0.1	0.1
ㄨㄣ	uən																																	1	0.5	0.5	0.1	0.1	0.8	0.1
ㄨㄤ	uaŋ																																		1	0.5	0.1	0.1	0.1	0.8
ㄨㄥ	uəŋ																																			1	0.5	0.5	0.5	0.5
ㄩㄝ	ye																																				1	0.5	0.5	0.5
ㄩㄢ	yan																																					1	0.5	0.5
ㄩㄣ	yən																																						1	0.5
ㄩㄥ	yəŋ																																							1