以最佳化及機率分佈判斷漢字聲符之研究

張嘉惠　Chia-Hui Chang
國立中央大學資訊工程學系
Department of Computer Science and Information Engineering
National Central University
chia@csie.ncu.edu.tw

李淑瑩, 林書彥, 黃嘉毅, 陳志銘
國立中央大學資訊工程學系
Department of Computer Science and Information Engineering
National Central University
{985202041, 965202104, 985202043, 985202018}@cc.ncu.edu.tw

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、九成的字是形聲字，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 7340 個形聲字中的判定聲符準確率爲 93.35%，而構件發聲分佈比較法則可達到 98.66%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

一、 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[1])，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，

Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING 2010), Pages 199-209, Puli, Nantou,Taiwan, September 2010.

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。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。爲此我們首先建立形聲字源標記系統，由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。不過由於此過程需耗費大量時間與人力投入（在 2009/11/10 至 2010/04/23 期間內共有 7340 字爲三位研究生共同標記完成），因此如何加速人工標記的速度是過程中必須解決的問題。在本篇論文中，我們提出二個不依靠人工標記而能自動判別形聲字聲符的方法。

1. 發音相似度比較法：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，因此我們經語言學專家的協助，分別制訂聲母、韻母之間發音相似度。進一步，爲了提升經由發音相似度比較法判斷聲符之準確率，我們採用限制性最佳化技術，求得新的發音相似度分數。
2. 構件發聲分佈比較法：通常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較爲集中。因此我們利用一個可以計算兩個機率分佈差距的公式 KL divergence，來計算每個構件的發聲分佈與所有漢字的發聲分佈 KL 值做爲構件做爲聲符的強度。

實驗結果顯示，發音相似度比較法在 7340 個形聲字中的判定聲符準確率爲 93.35%，而構件發聲分佈比較法則可達到 98.66%的準確率，顯示兩種方法做爲聲符判斷問題的可行性。

二、 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[2]。漢字構形資料庫不僅銜接古今文字以反映字形源流演，也記錄了記錄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[2,3]。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，順利讀出字的發音出來。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2010/04/23期間內共有7340字爲三位研究生共同標記完成），因此是否存在自動判定形聲字聲符的方法，則是本篇論文的重點。

三、 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。

漢字發音主要分爲聲母、韻母與調性三類。聲母隨著發音部位與發音方法而所有不同，如表 1 所示；而韻母則分爲單韻母、複韻母、聲隨韻母、捲舌韻母與結合韻母五種。因此我們在計算一個形聲字與其構件的發音相似度時也可依此三部份分別計算。

表 1：聲母的發音部位與發音方法。

發音部位 / 發音方法		上阻	上唇	上齒	齒背	上齒齦	前硬顎		軟顎
		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	ㄅ p			ㄉ t			ㄍ k
		送氣	ㄆ pʻ			ㄊ tʻ			ㄎ kʻ
塞擦	清	不送氣			ㄗ ts		ㄓ tʂ	ㄐ tɕ	
		送氣			ㄘ tsʻ		ㄔ tʂʻ	ㄑ tɕʻ	
擦	清			ㄈ f	ㄙ s		ㄕ ʂ	ㄒ ɕ	ㄏ x
	濁			(ㄪ) v			ㄖ ʐ		
鼻	濁		ㄇ m			ㄋ n		(ㄫ) ɳ	(ㄫ) ŋ
邊	濁					ㄌ l			

表 2：韻母種類

種類	注音符號
單韻母	(ㄭ)、ㄧ、ㄨ、ㄩ、ㄚ、ㄛ、ㄜ、ㄝ
複韻母	ㄞ、ㄟ、ㄠ、ㄡ
聲隨韻母	ㄢ、ㄣ、ㄤ、ㄥ
捲舌韻母	ㄦ
結合韻母	ㄧㄚ、ㄧㄛ、ㄧㄝ、ㄧㄞ、ㄧㄠ、ㄧㄡ、ㄧㄢ、ㄧㄣ、ㄧㄤ、ㄧㄥ
	ㄨㄚ、ㄨㄛ、ㄨㄞ、ㄨㄟ、ㄨㄢ、ㄨㄣ、ㄨㄤ、ㄨㄥ
	ㄩㄝ、ㄩㄢ、ㄩㄣ、ㄩㄥ

3.1 人工制定聲母和韻母發音相似度

依表 1 發音原則，我們制訂聲母與聲母之間發音相似度如下：

a. 國際拼音相同者：相似度爲 1。
b. 國際拼音相同，但是差一個上標號：相似度爲 a=0.9。
c. 國際拼音位於同行：相似度爲 b=0.8。
d. 國際拼音位於同列：相似度爲 c=0.5。

e. 國際拼音不同行也不同列：相似度爲 0.1。

另外，我們也制訂韻母與韻母的發音相似度如下：

a. 單韻母出現在結合韻母中的位置之後設相似度爲 x=0.8，若出現在前面則設相似度爲 y=0.5，未出現設相似度爲 0.1 。例如:ㄧ出現在ㄧㄚ之前設相似度爲 0.5 ；ㄞ出現在ㄧㄞ之後設相似度爲 0.8。

b. 單韻母、複韻母、聲隨韻母與捲舌韻母之間出現相同國際拼音部分，則設相似度爲 y，否則設相似度爲 0.1。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 0.5。

c. 結合韻母之間出現相同部分設相似度爲 y，否則設相似度爲 0.1。例如：ㄧㄚ和ㄨㄚ有相同部分ㄚ、ㄧㄢ和ㄨㄢ有相同部分，皆設相似度爲 x； ㄧㄞ和ㄨㄟ因無共同部分，設相似度爲 0.1。

根據上述規則所制訂的聲母與韻母相似度表分列於附錄一、二。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：

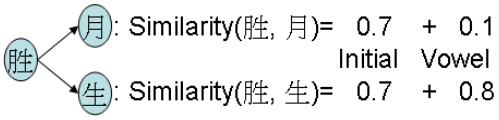
(1.) ㄈ與ㄅㄆㄇ的關係修正爲：合併爲同一行，隸屬唇音一大類。相似度定爲 b。

(2.) ㄗㄘㄙ、ㄓㄔㄕ與ㄐㄑㄒ三大類(不包括ㄖ)，除了同屬塞擦音、擦音外，在臺灣境內有許多人發音有彼此相混的現象，所以相似度定爲 d=0.7。

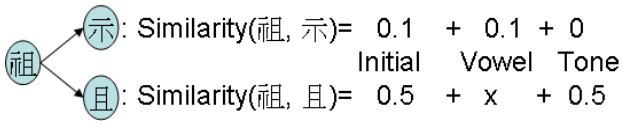
給定聲母與韻母相似表，我們定義兩個漢字發音相似度爲其聲母相似度與韻母相似度的總和：

Similarity(x,y)＝Initial(x,y)+Vowel(x,y)

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。舉例來說，漢字「胜」(ㄒㄧㄥ 1)的構件爲「月」(ㄖㄡ 4)和「生」(ㄕㄥ 1)，採用相似度比較公式表求算「月」和「胜」的分數，其中聲母同屬擦音，相似度爲 d＝0.7，而韻母相似度則爲 0.1，總和爲 0.7+0.1=0.8，同理「生」和「胜」的聲母亦同屬擦音，相似度爲 d，而「生」的單韻母出現「胜」的結合韻母之後，相似度爲 x＝0.8，總和爲 0.7+0.8=1.5，因此系統判定「生」爲「胜」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。舉例來說，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」的分數、「且」和「祖」的總和皆爲 0.7+0.1=0.8，因此我們再加上調性判別，由於「且」和「祖」的調性相同，因此我們預測「祖」的聲符爲「且」。



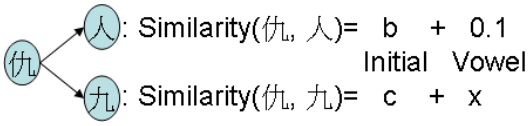
3.2 發音相似度自動制定法

由於前述發音相似度比較公式表是由人工制訂，假設我們不知聲韻相似度如何決定，但是已知某些字的聲符，是否能由已知資訊自動求出聲母和韻母發音相似度分數，從而決定聲符預測模型，則是本節所要探討的問題。我們嘗試以採用限制型最佳化方法計算聲母和韻母之發音相似度。最佳化的一般表示法如下：給定一組變數 $x \in R^n$，所希望最小化的目標函數 $f(x)$和某些約束條件 $h(x)=0, g(x)\le 0$，求得最小目標函數及當時的變數 x 解：

$$\min f(x),\ \text{s.t.} \begin{cases} h(x) = 0 \\ g(x) \le 0. \end{cases}$$

在我們的問題中，可以將聲母發音相似度分數 a, b, c 以及韻母發音相似度分數 x, y 做爲最佳化問題中的變數。給定一組已知聲符的漢字 T，依照發音相似度比較公式，我們可以爲已知訓練資料中的每一個字 $w \in T$ ，列出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。

舉例來說，漢字「仇」(ㄔㄡ 2)的構件爲「人」(ㄖㄣ 2)和「九」(ㄐㄧㄡ 3)，而其已知聲符爲「九」。針對仇字與其構件人、九，分別計算聲母和韻母之相似度，可得到 $b + 0.1 \le c + x$。



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 $\varepsilon_i \ge 0$，也就是 $b + 0.1 \le c + x + \varepsilon_i$，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：

$$\min \sum_i \varepsilon_i^p \ \text{s.t.} \ \begin{cases} b + 0.1 \le c + x + \varepsilon_1 \\ b + y \le b + 0.1 + \varepsilon_2 \\ \quad\quad \vdots \\ c + y \le b + 0.1 + \varepsilon_m \\ a, b, c, d, x, y, \varepsilon_i \ge 0. \end{cases}$$

其中 $p > 0$，代表錯誤聲符對系統的處罰程度的不同。

四、 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。首先，觀察一些常見的漢字，如下表 2：

表 2：構件「火」與構件「包」的漢字

包含構件「火」的漢字	包含構件「包」的漢字
伙(ㄏㄨㄛ 3)	泡(ㄆㄠ 4)
吹(ㄔㄩㄝ 1)	胞(ㄅㄠ 1)
灼(ㄓㄨㄛ 2)	跑(ㄆㄠ 3)
炊(ㄔㄨㄟ 1)	苞(ㄅㄠ 1)
煥(ㄏㄨㄢ 4)	砲(ㄆㄠ 4)

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 w，我們可以找出包含 w 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 w 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:

$$KL(\mathrm{P} \parallel \mathrm{Q}) = \sum_{\mathrm{i}} \mathrm{P(i)} \log \frac{\mathrm{P(i)}}{\mathrm{Q(i)}}$$

因此我們可以計算 KL(f(B) ‖ f(A))做爲構件 w 聲母強度，同理計算 KL(g(B) ‖ g(A)) 做爲 w 韻母強度，以及計算 KL(h(B) ‖ h(A)) 做爲 w 聲調強度。以下我們以調號爲例，計算構件「火」的聲調強度，我們從漢字構詞資料庫中找出所有標示注音的字共 14598 個， |A| = 14598，其調號分佈如下:

所有漢字之聲調分佈 h(A), |A|=14598				
調號 1	調號 2(ˊ)	調號 3(ˇ)	調號 4(ˋ)	調號 5(˙)
3418	4127	2411	4629	13

經正規化得 h(A)= (0.234, 0.283, 0.165, 0.317, 0.001)。我們同時統計含有構件「火」的字共有 259，而構件「火」的每個調號分佈如下:

含構件「火」的聲調分佈 h(B), |B|=259				
調號 1	調號 2(ˊ)	調號 3(ˇ)	調號 4(ˋ)	調號 5(˙)
53	68	37	101	0

經正規化可求得 h(B)= (0.205, 0.263, 0.143, 0.39, 0)。最後將 h(B)及 h(A)代入 KL-divergence 公式可得構件「火」的聲調強度:

$$KL(\mathrm{h(B)} \| \mathrm{h(A)}) = 0.205 \times \log \frac{0.205}{0.234} + 0.263 \times \log \frac{0.263}{0.2827}$$
$$+ 0.143 \times \log \frac{0.143}{0.165} + 0.390 \times \log \frac{0.390}{0.317} + 0 \times \log \frac{0}{0.001}$$

表 3：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

同理我們也可計算構件「火」的聲母 KL(f(B) ‖ f(A))及韻母的強度 KL(g(B)‖g(A))。應用時，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 3 分別列出聲母、韻母、聲調 KL 值與其構件出現次數|B|的乘積前十名構件。

五、 實驗

以下實驗探討前二章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 2010/04/23 期間內共有 7340 字爲三位中文系研究生共同標記完成。

5.1. 發音相似度比較公式表

首先我們進行發音相似度比較公式的預測效能。原始發音相似度方法，準確率約九成一，其中包含 368 筆無法判別的字；加入調號的判別，無法判別的字減少至 325，準確率約九成一八。採用專家建議的修正版之發音相似度比較公式表來測試，準確率提高至九成二一，再加入調號後則爲九成四的準確率。顯示以發音相似度比較公式進行判別聲符，有一定的效果。

表 4：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	6683	289	368	0.910
原始版＋調	6741	274	325	0.918
修正版	6761	268	311	0.921
修正版+調	6852	270	218	**0.934**

5.2. 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，在判別漢字聲符之效能。主要分成二項實驗進行。實驗一主要目的主要是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。我們取 p=1 做爲目標函數，隨機從 2000 筆訓練資料中抽取 n（=100, 150, 500, 750, 1000, 1500）個漢字產生 n 個限制條件，再由線性規劃方法計算出參數值，並利用這些參數來檢視 2000 筆訓練資料漢字聲符判別之準確率。我們反覆四次上述的實驗，用以繪製箱形圖(Box Plot)得到如圖 1 之結果。從圖 1 可以看出，當訓練資料筆數較少時，所得的準確率變化較大，而當限制條件個數大於 500 筆時，對於 2000 筆訓練資料漢字聲符判別之準確率已趨於平緩，顯示發音相似度最佳化可以在少數的已知聲符的漢字，即可得到相當不錯的結果。另外由於從 2000 中抽取 1500 筆時，會有相當多重覆取樣的情形，致使過度訓練，因此在 2000 筆中的準確率變異較大。

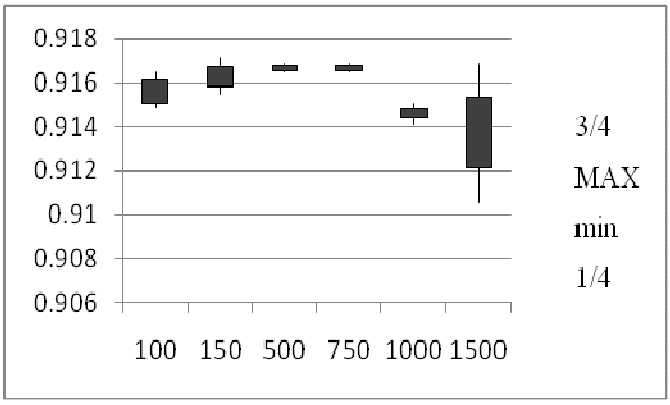


圖 1： 訓練資料數量對準確率之影響

實驗二目的則是要了解參數 p 的設定爲何，可以取得較佳的聲符判別之準確率。我們以隨機抽取 1000 漢字產生 1000 個限制條件，利用最佳化方法分別計算出其參數值，並測試漢字聲符判別之準確率，結果如表 5 所示。實驗顯示當 p 值爲 3 時測試資料中的準確率最高可達 0.9366。最後當 p=1/2 時，所得的參數爲極端，也就是除韻母相同設爲 0.9999 之外，其餘均設爲最小值 0.0001。

有趣的是這些訓練得到的聲母參數 a,b,c,d 均小於韻母參數值 x,y 值，顯示聲符韻母對於發音的影響遠大於聲母的影響。另外，原定同行聲母相似度分數略大於同列聲母相似度，與實驗結果相左，這些發音相似度分數是否有其意義，還需聲韻學方面的專家求證。

表 5：1000 筆訓練資料之相似度參數

Objective function	**p=1/2**	**p=1**	**p=2**	**p=3**	**p=4**
訓練資料平均準確率	0.9849	0.9935	0.9932	0.9939	0.9942
測試資料平均準確率	0.9169	0.9346	0.9255	**0.9366**	0.9200
a	0.1001	0.1951	0.3177	0.3363	0.3329
b	0.1001	0.1001	0.2410	0.3037	0.3294
c	0.1001	0.1002	0.3760	0.4371	0.4571
d	0.1001	0.1001	0.1001	0.1001	0.1001
x	0.9999	0.9999	0.6545	0.6281	0.6255
y	0.1001	0.1001	0.1001	0.1001	0.1001

5.3. 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。針對 7340 筆形聲字，其中有 7242 筆正確，98 筆錯誤，準確率 ：7242/7340 = 0.987。在三種方法中是準確率最高的一種方法。爲了解判定錯誤發生的可能原因，我們列出錯誤例子如表 6。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，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。因此未來我們將結合機率分佈比較法與發音相似度比較法，對於剩餘 14706-7340=7366 漢字進行聲符的判斷測試。

表 6：機率分佈比較法錯誤例子

字碼	注音	構件序	正確聲符	各構件機率分佈強度值
扣	ㄎㄡˋ	扌口	口	扌0.0756 > 口 0.0692
沐	ㄇㄨˋ	氵木	木	氵0.0284 > 木 0.0218
孟	ㄇㄥˋ	子皿	皿	子 0.6303 > 皿 0.2763
忝	ㄊㄧㄢˇ	天小	天	天 1.2455 < 小 1.8219
所	ㄙㄨㄛˇ	戶斤	戶	戶 0.7605 < 斤 0.9602
旺	ㄨㄤˋ	日王	王	日 0.1051 > 王 0.0904

六、 結論及未來研究

本篇論文是中央大學人文與數位整合計畫下，針對漢字學習形式的一項挑戰，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，我們所提出的兩種方式對於形聲字的聲符的判斷都有相當高的準確率。發音相似度最佳化方法雖然在判斷準確率上輸給機率分佈比較法，但是其所得的聲母，韻母相似度參數或許有助於未來漢字字音處理的研究，仍有相當的重要性。

未來本研究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。

七、 致謝

本論文的完成感謝李淑萍、廖湘美及孫致文教授，以及陳怡如、葉博榮、鍾哲宇、趙婕妤等人的幫助。

參考資料

[1] 許慎撰，段玉裁注，《說文解字注》，台北藝文印書館，1988年。

[2] 莊德明、謝清俊，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，2005年1月。

[3] 莊德明、鄧賢瑛，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，2008年8月。

附錄 A：聲母相似度比較公式表(右下角為修訂後的相似分數)。

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	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	0.5	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	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	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	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	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	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	0.1	0.1
ㄏ											1	0.1	0.1	0.5	0.1	0.1	0.5	0.1	0.1	0.5	0.1	0.1	0.5
ㄐ												1	0.9	0.8	0.5/0.7	0.1/0.7	0.1/0.7	0.1/0.7	0.5/0.7	0.1/0.7	0.1/0.7	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：韻母相似度比較公式表。

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