

國台語無聲調拼音輸入法實作

An Implementation of Toneless Input for Mandarin and Taiwanese

余明興 Ming-Shing Yu

國立中興大學資訊科學與工程學系

Dept. of Computer Science and Engineering

National Chung-Hsing University

Taichung, Taiwan, 40227

msyu@nchu.edu.tw

蔡承融 Cheng-Rong Tsai

國立中興大學資訊科學與工程學系

Dept. of Computer Science and Engineering

National Chung-Hsing University

Taichung, Taiwan, 40227

s9556010@cs.nchu.edu.tw

摘要

目前的市面上很少有支援輸入台語、客語、原住民語等語言的輸入法。在拼音方案百家爭鳴的環境下，以及減少使用者因聲調的關係而造成聲調輸入錯誤，我們的目標是要提供一個支援多種拼音方案且無聲調拼音的國語及台語的輸入法。本文使用連續三個詞長詞優先法，在國語無聲調音轉字得到 89.590% 的正確率，混合國語多種拼音方案得到 88.854% 的正確率。

ABSTRACT

There are very few input systems supporting Taiwanese and Hakka on the market at present. Our purpose is to provide an input system supporting Mandarin and Taiwanese which is toneless and complies with multiple types of phonetic symbols, we hope that such an approach can resolve the problems resulting from tone and phonetic symbols. In this paper,

we use an algorithm based on three continuous longest word first whose precision is 89.59% in one type of phonetic symbols, and 88.54% in the combination of many types of phonetic symbols.

關鍵詞：無聲調輸入法，音轉字，連續三個詞長詞優先法

Keywords: toneless input system, phoneme to character, three continuous longest word first.

一、前言

由於電腦及網際網路的普及化，使得愈來愈多人透過網站的方式來取得最新的知識；使用網路即時通訊軟體、電子信箱聯絡親朋好友，而這些都必須經由文字的輸入與電腦溝通，因此輸入法是一個重要的議題。然而占台灣 73.3%人口的閩南人和 12%的客家人及 1.7%的原住民，在目前的市面上卻很少有支援其輸入台語、客語、原住民語等母語的輸入法，因此本文將針對人口比例最高的閩南族群開發出一個輸入法。

目前的拼音輸入法主要分為注音符號及羅馬拼音兩種。我們採用羅馬拼音，主因是利用二十六個英文字母，就可以正確的標示出國語、台語、客語及英語裡的所有字，另外全世界的電腦都可以輸入英文字母。

由於國語及台語的拼音方案百家爭鳴，沒有統一版本，因此我們打算製作出多種拼音系統共存的輸入法。另外在國語及台語會因為語意的關係而發生變調「如：老李 買好酒 和 老李 買好 酒」聲調分別為「23323」及「23223」，台語也會因各地發音腔調的差異而產生聲調的不同，例如：「台南」一詞，北部調為「dai3 lam5」，南部調為「dai7 lam5」。為了減少使用者的負擔，因此我們系統設計為不用輸入聲調。

二、拼音方案概述

(一) 國語拼音方案簡介

「國語」是台灣地區以北京話為基礎制定的中華民國國語，由於台灣已獨立發展了數十年，與北京話出現了差異，因此又稱為「台灣國語」。以下將簡介台灣常見的國語羅馬字拼寫方案及說明其優缺點。

1. 國語注音符號第二式

「國語注音符號第二式」簡稱「注音二式」，是台灣教育部為了因應外籍人士學習中文的需求，在民國七十三年成立專案研究小組所制定。民國八十五年行政院經建會決議以國語注音符號第二式翻譯地名。注音二式有符合英語發音習慣的優點，但其缺乏國際性。

2.漢語拼音

漢語拼音是中國在西元 1958 年正式通過的漢語普通話拉丁轉寫統一規範，也是國際承認的標準 ISO 7098。漢語拼音中沒有台語和客語的設計，其中有些不合台語使用，例如「iu」用來代表一又。比較好的選擇是用「iu」來代表「優」(台語)，而用來「iou」代表「優」(國語)。在這種情況下，「niu」(娘，台語)和「niou」(牛，國語)可以併存而且英語直覺性較佳。

3.通用拼音

通用拼音是民國八十七年台灣中研院副研究員余伯泉以 KK 音標為基礎而發展出的拼音系統。它盡量和漢語拼音相容，而且能同時標注國語和台語。去除漢語拼音中不符合英語讀的習慣(x、q)。台北市政府研發出國語、台語及客語三種通用的通用拼音。民國八十九年台灣教育部宣佈使用，取代國語注音符號第二式及威妥瑪拼音。另外自民國九十一年起台灣行政院全面推行以通用拼音為基礎的中文譯音政策。

4.美式拼音

美式拼音是中興大學語音與語言實驗室[9]所提出的拼音方案，它是依照通用拼音來修改的，修改之處如下。「chi」取代通用拼音的「ci」來標示ㄑ。「tz」取代通用拼音的「z」來標示ㄗ，用「ts」取代通用拼音的「c」來標示ㄘ，這些取代都是注音二式早已經發展出來，且是我們認為比較符合美式英語的發音方式。用「chi」取代「ci」是因為「ci」在美語中較常讀成「si」。用「tz」取代「z」之後，「z」可以在台語中使用，例如「zin」可拼「人」的音。表一是注音二式、漢語拼音、通用拼音及美式拼音的子音和母音對照表。

表一、國語拼音對照表。

注音符號	注音二式	漢語拼音	通用拼音	美式拼音	注音符號	注音二式	漢語拼音	通用拼音	美式拼音
ㄅ	b	b	b	b	ㄗ	tz	z	z	tz
ㄆ	p	p	p	p	ㄘ	ts	c	c	ts
ㄇ	m	m	m	m	ㄙ	s	s	s	s
ㄈ	f	f	f	f	ㄚ	a	a	a	a
ㄉ	d	d	d	d	ㄛ	o	o	o	o
ㄊ	t	t	t	t	ㄜ	e	e	e	e
ㄋ	n	n	n	n	ㄝ	ie	ie	ie	ie
ㄌ	l	l	l	l	ㄞ	ai	ai	ai	ai
ㄍ	g	g	g	g	ㄟ	ei	ei	ei	ei
ㄎ	k	k	k	k	ㄠ	au	ao	ao	ao
ㄏ	h	h	h	h	ㄡ	ou	ou	ou	ou
ㄐ	j(i)	j	j(i / y-)	j	ㄢ	an	an	an	an
ㄑ	ch(i)	q	c(i / y-)	ch(i)	ㄣ	en	en	en	en
ㄒ	sh(i)	x	s(i / y-)	s	ㄤ	ang	ang	ang	ang
ㄓ	j	zh	jh	jh	ㄥ	eng	eng	eng	eng
ㄔ	ch	ch	ch	ch	ㄜ	er	er	er	er
ㄕ	sh	sh	sh	sh	ㄝ	i,yi	i,yi	i,yi	i,yi
ㄖ	r	r	r	r	ㄨ	u,wu	u,wu	u,wu	u,wu
空韻	r,z	i	ih	ii	ㄩ	iu,yu	ü,yu	yu	yu

(二) 台語拼音方案簡介

「台語」為臺灣地區所使用的閩南語，又可稱為「台灣話」、「福佬話」、「河洛話」、「台灣閩南語」，是目前在台灣使用最多的語言之一。其由閩南人在明朝末年開始渡台，歷經了荷蘭、日本的統治，發展出獨特的閩南語。在所有的閩南語中，台灣閩南語與廈門話最為接近，主要的差異是在詞彙方面，據王育德、鄭良偉等人指出約有百分之十的不同。在本章節將簡介目前常用的台語拼寫方案。

1.教會羅馬字

「教會羅馬字」簡稱「教羅」，是 19 世紀時由基督教長老教會所創造，是以拉丁字母書寫的閩南語正字法，又稱可為「白話字」。

2.臺灣閩南語羅馬字拼音

臺灣閩南語羅馬字拼音簡稱「台羅拼音」，是台灣官方的閩南語拼音方案，它是整合原有的台灣閩南語音標(TLPA)及教會羅馬字而成的閩南語羅馬字拼音。

3.通用拼音

通用拼音和台羅拼音最大的不同有兩點。第一點是通用拼音有一併考慮到國語和客語，而台羅拼音只考慮到台語。第二點是關於ㄅ、ㄆ、ㄇ的拼音選擇。通用拼音使用 b、d、g 來標註，較合美式英語；而台羅拼音使用 p、t、k 來標註，較像國際音標。

4.美式拼音

除了前面的描述之外，美式拼音將國語ㄓ、ㄒ、ㄑ、ㄒ、ㄓ、ㄌ的空韻由加上「ih」改成加上「ii」。原因是「h」結尾的音為台語的入聲音，可能會有衝突發生。例如 *cih*(ㄓ)和 *sih*(ㄌ)也是台語的入聲音。台羅、通用和美式拼音的子音對照表如附錄一所示。

三、音轉字處理

音轉字的輸入方式可分為二種，一種是透過鍵盤輸入，另一種採用語音的輸入方式，而它們的主要問題是如何從一音多字裡選擇正確的字。在國語中字的音節有四百多個，而國語的字卻有一萬個以上，台語字的音節有七百多個，台語字也有八千個以上，平均一個音節會對應到一、二十個字。以下我們將探討此問題。

在處理中文音轉字的問題中，目前常見的方法有規則法及統計法。

- ◆ 規則法：從語言學中訂出規則，依據所訂的規則判斷出合理的結果，其缺點是需要大量的專業人士參與。

- ◆ 統計法：其中 N-Gram 語言模型[3]是目前最常用的方法，使用語料庫來訓練語言模型得到字、詞或詞性間的關係。

以下將介紹我們處理音轉字的方式。

(一) 國語音轉字

由於我們是要實作一個輸入法，所以不希望系統修改太久以前輸入的資訊造成使用者的煩惱，因此我們的組字視窗限制在十二個字內。系統只會修改組字視窗內的字，當組字視窗超過十二個字時，系統則會自動輸出第一個辭彙。在組字視窗內，我們依連續三個詞的長詞優先演算法[6][15]找出合理的結果。當連續三個詞的長詞優先演算法找出的結果有二組以上時，我們實驗二種方式來音轉字的計算分數，分別為公式 1 及公式 2。

W_m 是候選詞組的詞彙， $P(W_m)$ 是詞的機率。

$$T_1 = \prod_{m=1}^n P(W_m) \quad (1)$$

$$T_2 = \prod_{m=1}^n P(W_m) * \prod_{m=2}^n P(W_m | W_{m-1}) \quad (2)$$

(二) 台語音轉字

我們的台語音轉字採用上節所提的連續三個詞的長詞優先法，台語輸入採用長詞優先法的好處是較不需要訓練語料。台語輸入有一個要注意的情況是台語音 $S = S_1, S_2, \dots, S_N$ (N 是總音節數， S_i 是第 i 個音節) 對應至台語文 $X = X_1, X_2, \dots, X_T$ (T 是總字數， X_i 是第 i 個台語字) 時， N 與 T 的長度不一定是相同。在考量各種情況下，我們系統優先輸出音節數與辭彙字數相同的辭彙，使用者如果想要字數與音節數不相同的詞彙可以在修改模式中去選取。

(三) 智慧型處理

由於國語及台語的拼音方案眾多，因此對於拼音系統不熟悉的初學者往往會混合著不同的拼音方案輸入；所以我們希望系統能給與多種拼音方案相容的方式，以減少初學者拼音的錯誤率。我們採取的方法為當使用者輸入一個音串，系統會去評估使用者是輸入那一種拼音方案的那個音節，例如：使用者輸入「cyuan」音串，系統就評估它是輸入

通用拼音方案的「ㄍㄛㄛ」音節。

然而有些音串對應到的音節不只一種，例如：「niu」音串，可以對應到漢語拼音的「ㄋㄟ」音節及注音二式的「ㄋㄞ」音節，表二為國語拼音系統互相衝突的地方。對於此狀況，我們將可能的音節都送入至系統中，然後再依上下文估算合理的結果，例如：「liu ju mei guo yang ji duei wang jian min jhu tou liu ju wu shih fen , liu ju jin ji guan jyun jhan de si wang」，這一句中「liu」音串可以對應到漢語拼音的「ㄌㄟ」音節及注音二式的「ㄌㄞ」音節，「ju」音串可以對應到漢語拼音的「ㄐㄩ」音節及注音二式的「ㄐㄩ」音節。我們依上下文推斷合理的結果為「旅居美國洋基隊王建民主投六局無失分，留住晉級冠軍戰的希望」。

表二、國語拼音系統衝突表

音串	漢語拼音	注音二式
niu	ㄋㄟ	ㄋㄞ
liu	ㄌㄟ	ㄌㄞ
jiu	ㄐㄟ	ㄐㄞ
ju	ㄐㄩ	ㄐㄩ
juan	ㄐㄩㄛ	ㄐㄩㄛ
chi	ㄔ	ㄔ
shi	ㄕ	ㄕ

四、實驗與討論

在本章節中將說明我們訓練、測試語料庫和辭典的來源，及實驗結果的數據。

(一)語料庫

我們的國語語料庫來源有兩個。第一個是中央研究院平衡語料庫[17]，這是個約有五百萬詞的語料。另外一個是我們實驗室從各新聞網站上蒐集而來的新聞語料庫，這是一個含有約一千六百萬個詞(二千八百萬個字)的語料庫。

其中我們利用中央研究院的中文斷詞器[16]來訓練詞的 Bigram 模型，以及擷取約百分之十的語料庫，再透過[11]中文字轉音程式得到測試語料庫。

(二)辭典

在音轉字的過程中，辭典扮演著一個非常重要的角色，一個不錯的辭典可以提高正確率。在國語部份，我們的辭典來源有兩個。辭典 1 是論文[13]中使用的辭典，其辭典是以中研院八萬目詞為基礎，另外再從中研院平衡語料庫[17]抽取出未在八萬目詞內的詞，一共約 13 萬詞的辭典。辭典 2 是[11]從文章中擷取出約 44 萬個中文的常用字串，這些字串有對應的音，但沒有詞性可用。

在台語辭典部份，我們使用[10]所整理的辭典為基礎，其辭典約有 6 萬詞。

(三)實驗結果

1.實驗一

首先我們將實驗加入中文常用字串[11]對國語音轉字的影響，音轉字的演算法如上節所提的連續三個詞長詞優先詞。由實驗結果表三得知單獨使用[11]中文常用字串的辭典 2 就可以得到 87%的正確率，另外合併這二個辭典正確率可以提升至 88%。因此我們使用辭典 3 作為我們國語音轉字的辭典，後面的實驗也皆使用辭典 3。

表三、實驗一的結果

辭典	正確率
辭典 1	77.622%
辭典 2	87.500%
辭典 3	88.459%

註: 辭典 1 :[13]論文中使用的 13 萬詞辭典。

辭典 2 :[11]從文章中擷取出約 44 萬個中文的常用字串。

辭典 3 :合併辭典 1 與辭典 2。

2.實驗二

實驗 3.1 節所提的二種方法在國語音轉字的正確率，由實驗結果表四得知有加入 Bigram 的資訊的方法 2 所得到的正確率略高於只用到 Uigram 資訊的方法 1。

表四、實驗二的結果

方法	正確率
方法 1	88.459%
方法 2	89.590%

註：方法 1：候選詞組不只一組時，使用公式 1。

方法 2：候選詞組不只一組時，使用公式 2。

3.實驗三

最後我們將實驗在混合通用、漢語、注音二式及美式四種拼音方案時，國語音轉字的正確率。由實驗結果表五得知有加入 **Bigram** 的資訊的方法 2 所得到的正確率略高於只用到 **Uigram** 資訊的方法 1。另外這二種方法音轉字的正確率只略低一點點於無混合拼音方案的情況。

表五、實驗三的結果

方法	正確率
方法 1	87.879%
方法 2	88.854%

註：方法 1：候選詞組不只一組時，使用公式 1。

方法 2：候選詞組不只一組時，使用公式 2。

(四)實驗討論

根據實驗 1 的結果，得知合併辭典 1 及辭典 2 的辭典 3 可以得到 88% 國語音轉字的正確率，遠高於辭典 1 的 77% 正確率，推測原因可能是大部份的句子中都包含了常用的字串。因此我們的系統使用辭典 3 作為國語音轉字的辭典。

在實驗 2 的結果，發現加入 **Bigram** 資訊的方法 2 正確率有 89.590%，比方法 1 提昇 1% 的正確率。在實驗 3 的結果得知我們的方法在混合通用、漢語、注音二式及美式四種拼音方案的情況下，國語音轉字的正確率皆有 87% 以上。由於所訓練的 **Bigram** 資訊有約 120 萬筆，檔案龐大且正確率只提昇一點點，因此我們的音轉字採用方法 1。

五、系統設計考量

由於目前台語並沒有像中文一樣有正式且統一的文字，因此使用者對於系統的台語辭彙可能會不滿意，所以我們提供台語音對應至國語辭彙的功能。例如：若是使用者不滿意台語音「**bhin-a-am**」對應至台語詞「明仔暗」，可透過修改的方式，將台語音「**bhin-a-am**」對應至國語詞「明天晚上」或「明晚」。

當使用者選擇將台語音對應至國語辭彙時，我們的系統將會改變使用者原始輸入的音串，把使用者輸入的台語音改成國語辭彙可能的台語音(音節數與辭彙字數相同)。例如：使用者選擇將台語音「**bhin-a-am**」對應至國語詞「明晚」，我們將會把使用者輸入的台語音「**bhin-a-am**」轉成國語詞「明晚」最有可能的台語音「**bhin- bhuan**」。

六、結論與未來改進方向

雖然本論文完成了多種拼音方案相容的國台語無聲調拼音輸入法，但仍有些部份可以改進。在台語音轉字的部份由於欠缺語料庫，因此我們打算收集使用者輸入的語料，以改善台語音轉字的正確率。在國語音轉字雖然達到近九成的正確率，但未來我們可以加入構詞器及收集使用者輸入的語料，來提昇音轉字的正確率。另外我們希望未來能提供國語、台語及英語混合輸入的功能，讓使用者不必透過切換，就可以同時輸入國語、台語及英語這三種語言。

參考文獻

- [1] Amelia-Fong Lochovsky and Hon-Kit Cheung, "N-gram Estimates in Probabilistic Models for Pinyin to Hanzi Transcription", IEEE International Conference on Intelligent Processing Systems, Beijing, 1997, pp. 1798-1803.
- [2] Bing-Quan Liu and Xiao-Long Wang, "An Approach to Machine Learning of Chinese Pinyin-to-Character Conversion for Small-Memory Application", Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, 2002, pp. 1287-1291.
- [3] Frederick Jelinek, "Statistical Methods for Speech Recognition", The MIT Press, Cambridge Massachusetts, 1997.
- [4] Shuanfan Huang, "Language, Society, and Ethnic Identity (語言、社會、與族群意識)", Taipei Crane, 1993。
- [5] OpenVanilla, http://openvanilla.org/index-zh_TW.php。
- [6] T. H. Ho, K. C. Wang, J. S. Lin, and L. S. Lee, "Integrating Long-Distance Language Modeling to Phoneme-to-Character Conversion," Proceeding of ROCLING X, pp. 287-292, 1997.
- [7] Xiaolong Wang, Qingcai Chen, and Daniel S. Yeung, "Mining Pinyin-to-Character Conversion Rules from Large-Scale Corpus: A Rough Set Approach", IEEE Transactions on System, Man, and Cybernetics, 2004, pp. 834-844.
- [8] Xuan Wang, Lu Li, Lin Yao, and Waqas Anwar, "A Maximum Entropy Approach to Chinese Pinyin-to-Character Conversion", IEEE International Conference on Systems, Man, and Cybernetics, Taipei, 2006, pp. 2956-2959.
- [9] 余明興, "台灣共通語言", 第十九屆自然語言與語音處理研討會, 2007, pp. 319-333.
- [10] 蔡宗謀, "中文文句轉台語語音系統初步研究", 中興大學資訊科學與工程研究所碩士論文, 2008。
- [11] 林義証, "中文常用字串-一個優於傳統語言模型的新觀念", 中興大學應用數學系

博士論文，2002。

[12]林嘉信，“與多種拼音方法相容的國語輸入系統”，中興大學應用數學研究所資訊組碩士論文，2002。

[13]張唐瑜，“以大量詞彙作為合成單元的中文文轉音系統”，中興大學資訊科學研究所碩士論文，2005。

[14]許聞廉與陳克健，“自然智慧型輸入系統的語意分析脈絡會意法”，台灣中央研究院資訊所，1993。

[15]羅火嵐，“中文無聲調拼音輸入法及其實作”，中興大學資訊科學研究所碩士論文，2006。

[16]中央研究院中文斷詞系統，<http://ckipsvr.iis.sinica.edu.tw>。

[17]中央研究院平衡語料庫，<http://www.sinica.edu.tw/ftms-bin/kiwi1/mkiwi.sh>。

[18]世界台灣語通用協會，中研研究院，<http://abc.iis.sinica.edu.tw/>。

[19]國語注音符號第二式，教育部，
http://www.edu.tw/files/site_content/M0001/er/cmain.htm。

[20]漢語拼音方案，中國教育和科研計算機網，
<http://www.edu.cn/20011114/3009777.shtml>。

附錄一. 國台語拼音的子音表。台語羅馬字只列出教育部2006年公佈的部分。

注音符號	美式拼音	通用乙式	漢語拼音	注音二式	台語羅馬字
ㄅ	b	b	b	b	p
台帽	bh	bh			b
ㄆ	p	p	p	p	ph
ㄇ	m	m	m	m	m
ㄈ	f	f	f	f	
ㄉ	d	d	d	d	t
ㄊ	t	t	t	t	th
ㄋ	n	n	n	n	n
ㄌ	l	l	l	l	l
ㄍ	g	g	g	g	k
台鵝	gh	gh			g
ㄎ	k	k	k	k	kh
ㄏ	h	h	h	h	h
ㄐ	ji	ji	j	ji	
ㄑ	chi	ci	q	chi	
ㄒ	si	si	x	shi	
ㄓ	jh	jh	zh	j	
ㄔ	ch	ch	ch	ch	
ㄕ	sh	sh	sh	sh	
ㄖ	r	r	r	r	
ㄗ	tz	z	z	tz	ts
ㄘ	ts	c	c	ts	tsh
ㄙ	s	s	s	s	s
台字人如	z	zz			j
零韻	-ii	-ih	-i	-ih	
台姆(台ㄇ)	m(mh)	m(mh)			
台秧(台ㄋ)	ng(ngh)	ng(ngh)			ng