中華民國計算語言學學會
通訊 ACLCLP
發行人：吳宗憲
主　編：王新民
執行編輯：黃　琪
The Association for Computational Linguistics and Chinese Language Processing 第二十二卷第三期
E-Mail：aclclp@hp.iis.sinica.edu.tw
地 址：台北市研究院路二段128號中研院資訊所
Website：http://www.aclclp.org.tw
電 話：(02)2788-3799 ext:1502
傳 眞：(02)2788-1638
劃 撥：19166251

本 期 要 目

ROCLING 2011

「第二十三屆自然語言與語音處理研討會」由國立台北科技大學電子系、電通所、及本會聯合主辦，謹訂於九月八日(週四)～九日(週五)假台北市台北科技大學科技大樓國際會議廳舉行。本次會議共邀請到三位國際知名學者，分別是：資訊檢索專家，現任 Google 台灣區總經理簡立峰博士、人工智能語言處理專家，現任新加坡科技研究局資訊通信研究院主任李海洲博士、及從事自然語言處理研究，東北大學信息學院朱靖波教授作專題演講，講題及主講人簡介請參閱本刊第二～三頁。ROCLING 會議論文已被 ACL Anthology 收錄(http://aclweb.org/anthology-new/)。會議相關訊息請參閱大會網頁：http://sites.google.com/site/rocling2011/。

博碩士論文獎 7/1 開始申請

名額及獎項：

1. 博士論文優等獎：一名，獎金二萬元，並頒給學生及指導教授獎狀各乙紙。
2. 博士論文佳作獎：一名，獎金一萬元，並頒給學生及指導教授獎狀各乙紙。
3. 碩士論文優等獎：一名，獎金一萬元，並頒給學生及指導教授獎狀各乙紙。
4. 碩士論文佳作獎：三名，獎金各伍仟元，並頒給學生及指導教授獎狀各乙紙。

申請資格：

1. 國內大專院校博碩士班應屆畢業生從事計算語言學相關研究方向者，由其指導教授推薦。
2. 參賽限制：每位指導教授以推薦一篇博士論文及兩篇碩士論文爲限。(含個人指導與共同指導)。

申請期間：7/1～7/31，申請手續請參閱本會網頁：http://www.aclclp.org.tw/doc/shipreg.htm。

一〇一年度會費開始繳交

民國一〇〇年度「個人會員」及「學生會員」有效期已於六月三十日到期，爲保障各位會員之權益，敬請如期繳交會費；若您個人的基本資料有所異動，亦請惠予通知。會員資格更新暨個人資料異動通知單及信用卡繳費單請參閱本訊第 15 頁。

獎助學生出席國際會議公告

- 會議名稱：ACM SIGIR 2011
- 論文題目：Collaborative Cyberporn Filtering with Collective Intelligence
- 獎助學生：李龍豪(台灣大學資訊工程所)
- 獎助金額：US$600 元

ROCLING 2011 Keynotes

Title: **Machine Transliteration - Translating the Untranslatables**

Speaker: ***Dr. Haizhou Li***, Institute of Infocomm, Singapore

Biography:

Dr. Haizhou Li is currently the Principal Scientist and Department Head of Human Language Technology at the Institute for Infocomm Research. Dr. Li has worked on speech and language technology in academia and industry since1988. He taught in the University of Hong Kong (1988-1990), South China University of Technology (1990-1994), and Nanyang Technological University (2006-). He was a Visiting Professor at CRIN/INRIA in France (1994-1995), and at the University of New South Wales in Australia (2008). As a technologist, he was appointed as Research Manager in Apple-ISS Research Centre (1996-1998), Research Director in Lernout & Hauspie Asia Pacific (1999-2001), and Vice President in InfoTalk Corp. Ltd (2001-2003).

Dr. Li's research interests include automatic speech recognition, natural language processing and social robotics. He has published over 150 technical papers in international journals and conferences. He holds five international patents. Dr. Li now serves as an Associate Editor of IEEE Transactions on Audio, Speech and Language Processing, ACM Transactions on Speech and Language Processing, and Springer International Journal of Social Robotics. He is an elected Board Member of the International Speech Communication Association (ISCA, 2009-2013), an Executive Board Member of the Asian Federation of Natural Language Processing (AFNLP, 2006-2010), and a Senior Member of IEEE since 2001. Dr. Li was the Local Organizing Chair of SIGIR 2008 and ACL-IJCNLP 2009. He was appointed the General Chair of ACL 2012 and Interspeech 2014. He was the recipient of National Infocomm Award of Singapore in 2001. He was named one of the two Nokia Professors 2009 by Nokia Foundation in recognition of his contribution to speaker and language recognition technologies.

Abstract:

Machine transliteration is the process of automatically rewriting the script of a word from one language to another, while preserving pronunciation. The last decade has seen a tremendous progress and a growth of interests from theory to practice of machine transliteration. In this talk, I will present an overview of the fundamentals, algorithms and applications, in particular, transliteration between English and Chinese. I will also report the findings in the most recent transliteration evaluation campaigns - NEWS 2009 and NEWS 2010 Machine Transliteration Shared Tasks.

Title: **Some Issues on Statistical Machine Translation Using Source and (or) Target Syntax**

Speaker: ***Prof. Jingbo Zhu***, Computer Science at the Northeastern University at Shenyang, China

Biography:

Prof. Jingbo Zhu is a full professor of Computer Science at the Northeastern University at Shenyang, China, and is in charge of research activities within the Natural Language Processing Laboratory (NEU-NLPlab, htttp://www.nlplab.com). He received his Ph.D. degree in computer software and theory from the Northeastern University in 1999. He was a visiting researcher at the City University of Hongkong (2004) and ISI, University of Southern California at Los Angeles (2006-2007), and was selected by the Program for New Century Excellent Talents in University, Ministry of Education (2005). His research interests include machine translation, syntactic parsing, sentiment analysis and text mining. He has published 100+ papers in many high-level journals and conferences including IEEE Transactions on Affective Computing, IEEE Transactions on Audio, Speech and Language Processing, ACM Transactions on Speech and Language Processing, ACM Transactions on Asian Language Information Processing, and ACL/EMNLP/Coling, etc.

Abstract:

Machine Translation (MT) is one of the oldest sub-fields in Natural Language Processing (NLP) and Artificial Intelligence (AI). During the last decade, syntax-based approaches have received growing interests in MT community, showing state-of-the-art performance for many language pairs such as Chinese-English. In this talk, I will present our recent work on syntax-based MT, and some approaches to performing translation using source and (or) target syntax, involving string-to-tree, tree-to-string and tree-to-tree SMT paradigms. Also, an empirical study is shown to compare the strengths and weaknesses among various syntax-based SMT approaches. Furthermore, several interesting issues are further addressed to investigate what the major problems in current (syntax-based) MT paradigm are. Finally, I will spend a little time to introduce a new open-source SMT toolkit (named NEUTrans) which was developed by the NLPLab of Northeastern University, and our current efforts on incorporating syntax-based SMT paradigms into this open SMT platform.

Title: To be determined

Speaker: ***Dr. Lee-Feng Chien***, General Manager, Google, Taiwan

Abstract:

N/A yet

Biography:

N/A yet

應用資料探勘於形聲字發音規則之探究

張嘉惠
國立中央大學資訊工程系

一、 研究背景

漢字是世界上最古老的文字之一，也是至今仍廣爲使用一種形系文字。近年來由於中國市場的興起，以華語做爲第二外語的學習也連帶地愈來愈受到重視，華語學習者的人數也倍數成長。對於未來華語文學習市場的龐大需求，台灣自然不能缺席。

對於華語作爲第二語言的學習者而言，從單字、單詞的學習開始，即是相當的挑戰。因爲漢語的字形與音調並無拼音文字連接關係，學習者要同時進行形、音、義三者的連結，造成學習曲線不佳；再者成人學漢字，著重在認知轉換，並善用其理解和歸納的能力（高柏園等人 2008），因此有故事性的解說一個字的起源，拆解其組成的部件，提供漢字發音的關聯規則，將有助於解決華語第一階段識字學習的難處。

過去這些知識與規則只能靠對文字學有專研的教師教導或是學習者慢慢累積經驗，不僅對於海外華語師資的培育緩不濟急，對於學習者而言更是一條漫長的路。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)，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。

爲有助於以華文做爲第二語言的學習者，在中央大學數據中心黃鍔院士及科教處處長陳國棟教授的催生之下，由中央大學資工系與中文系共五位教授組成團隊，從 2009 年 11 月開始透過建立的「形聲字源暨聲符標記系統」，由中文系師生協助形聲字及聲符標記，並由資工系師生利用資料探勘技術分析漢字得聲偏旁的規則，同時輔以統計資料分析形聲字與得聲偏旁間的轉變規律。我們以中研院文獻處理實驗室所建立的「漢字構形資料庫」爲基礎，希望透過資料的分析統計與資料探勘等技術，發掘串連漢字之間關係的形音關聯規則。我們的目標是提出一個以聲符部件教學爲主的漢字學習策略，用以提高學習曲線，讓漢字不是教一個字才學到一個字，而能搭配發音關聯規則「一舉數字」，發揮數位學習的優點。

爲達此項目標，我們擬定了三個階段的目標：第一階段是形聲字聲符的標記，第二階段是發音關聯規則的探勘與分析，第三階段則是部件教學教材編撰輔助系統。本文主要針對前兩階段所衍生的問題及解決方法做一介紹，期望更多對此有興趣的文字學者及研究人員共同參與。

二、 國內外相關研究

數位典藏與數位學習國家型計畫自 97 年開始徵求華語文作爲第二語言之數位學習研究計畫，徵求重點從學習歷程研究、華語文數位教學模式，延伸到學習策略，其中不乏數位教材課程設計、語料庫建構[1]、學習平台與測驗系統的研發，另外也有探討同步視訊[2]、雲端系統[3]、互動多媒體、學習載具等不同網路科技下，有關師資培訓、教學方法、線上學習成效的研究[4]，同時針對字詞教學模式、華語寫作、閱讀輔助、以及不同學習者如新移民[5]、以德語爲母語者的教材設計，也有相關研究進行。

與本計畫最爲相關的研究計畫是淡江大學中文系高柏園、郭經華、胡映雪教授所主持之"字詞教學模式與學習歷程研究"。其概念是藉由即時回饋的寫字練習（學文 Easy Go!），比較部件拆解做爲漢字教學策略成效（洪文斌 2010），輔以線上教學平台「 IWiLL Campus 」（郭經華 2010），進行「以字帶詞」之詞彙學習策略（高柏園 2010）。此計畫在美國加州地區 Saratoga High School 針對 26 名修習 AP 中文課程之學生，實施四週約八堂之主題課程，用以評估漢字部件教學之學習策略對於海外華語文學習者之成效。從國科會期中報告顯示，採用多媒體自習一組的學生在認字、書寫、及字的結構上，比傳統標示筆劃順序的習字方法呈現較佳的成果，顯示以部件拆解做爲漢字教學策略的可行性。

最早有關漢字構造的研究，因屬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[10]。漢字構形資料庫不僅銜接古今文字以反映字形源流演變，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。

然而漢字構形資料庫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；因此文獻處理實驗室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，讓華語學習者可以在聲符與規則的輔助下，順利讀出字的發音出來。

[1]師範大學國語教教學中心陳浩然教授"華語學習者中介語料庫之建構計畫"。
[2]師範大學華語文教學研究所信世昌教授所主持的"以同步視訊爲主體之數位學習研究"。
[3]東華大學資訊張瑞雄教授所主持的"華語文數位學習雲端系統之研發"。
[4]如師範大學資訊教育研究所張國恩教授所主持的"能力導向全方位華語學習"，鄭錦全教授所主持的"海外華語學習者之線上學習成效評估研究"。
[5]中央院資訊所許聞廉研究員"構建一個新移民有機成長的多元認同平台的整合研究"。

三、 形聲字聲符標記系統

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，從 2009/11/10 開始至 2011/4/30，耗費一年多時間，才標記完漢字構形資料庫中有注音標記的 14309 個漢字。因此在標記系統的過程中，我們也試圖發掘形聲字與聲符間是否有特定的關係，並檢驗這些關係對於聲符預測的有效性。

第一個方式是以部首來判定形聲字的聲符。漢字構形資料庫中雖沒有標記每一漢字的結構組合類型（象形、指事、會意、形聲、轉注及假借），但是卻有每一個漢字所屬的部首以及其組成的部件。例如：「詁」字的部首爲「言」，組成構件「古」及「言」，連接方式爲左右排列；「辜」字的部首爲「辛」，組成構件「古」及「辛」，連接方式爲上下排列。根據經驗法則，部首通常表意，因此非部首的部件通常爲形聲字的聲符。依此經驗法則遞迴拆解每一個漢字，在 9593 筆形聲字中約有 90.9%的準確率。

第二個方式則是以形聲字與其組成構件的發音相似度較高者做爲聲符判斷的依據。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。

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所有不同，而韻母則是一個音節中的元音（母音），也是押韻的主要部份，可分爲單韻母、複韻母、聲隨韻母、捲舌韻母與結合韻母五種 ；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。經由聲韻學家的協助，我們分別制訂聲母與聲母之間發音相似度，以及韻母與韻母之間發音相似度。例如依照聲母發音部位及方法的異同決定聲母之間的相似度分數，又如依照單韻母、複韻母、聲隨韻母與捲舌韻母之間出現相同國際拼音部分，來決定韻母之間相似度分數。

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，是否能由已知形聲字自動求出發音相似度分數，從而了解形聲字在多方演變下是否仍保有與聲符之間的高相似度，則是另一種聲符預測模型的建構方法。我們嘗試以限制型最佳化（Constrained Optimization）方法計算聲母之間發音相似度和韻母之間發音相似度應有的值。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「校」與「木」的發音相似度必須小於等於「校」與「交」的發音相似度。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 $\varepsilon_i \geq 0$，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。

第三種方式則是以每一個構件的發音強度做爲判斷聲符的標準。令 A 表示所有漢字所成的集合，我們以 $P_I(A)$、$P_V(A)$、$P_T(A)$分別表示漢字的聲母、韻母及調的分佈機率。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、韻母及調的分佈機率 $P_I(B)$、$P_V(B)$、$P_T(B)$。對於聲符而言，由於發音集中度較高，因此聲母分佈（$P_I(B)$與 $P_I(A)$）、韻母分佈（$P_V(B)$與 $P_V(A)$）以及聲調分佈（$P_T(B)$與 $P_T(A)$）就會有較大的差異。因此我們可以計算 $KL(P_I(B)||P_I(A))$做爲構件 w 聲母強度，同理計算 $KL(P_V(B)||P_V(A))$ 可以做爲 w 韻母強度，以及計算 $KL(P_T(B)||P_T(A))$ 做爲 w 聲調強度。

如表 1 所示，以非部首構件做爲聲符的經驗法則，準確率約 90.90%。利用聲韻學家建議的聲母間及韻母間發音相似度數值，對於聲符預測準確率約 90.70%，其中包含 244 筆無法判別的字；顯示以發音相似度比較公式進行判別聲符，有一定的效果。另藉由最佳化方法可以推出與聲韻學家建議相近的相似度參數，甚至於在少數已知聲符的漢字訓練資料，如 20 或 40 筆時，即有相當好的結果。事實上聲符預測準確率最高 90.03%，可由 40 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 及 1000 筆限制條件時，反而有下降趨勢。主要的原因在於聲符與形聲字發音相似度大於聲符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）。

表1：聲符預測方法比較

方法		正確	錯誤	無法判別	準確率
1. 非部首構件		8721	858	14	90.90
2. 發音相似度	**聲韻學建議**	8701	648	244	90.70
	限制型最佳化	8576	575	442	89.39
3. 機率分佈比較法	**聲**	8910	683	0	92.80
	韻	9362	231	0	97.50
	調	9207	386	0	95.90
	聲+韻+調	9413	180	0	**98.10**

最後有關機率分佈比較法對於在判別漢字聲符之效能結果如下：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，比起前述的兩種方法，有更高的準確度（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%。相關研究成果，請參考（張嘉惠等人 2010）。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。

四、 形聲字發音規則的探勘與分析

漢字主要源於圖形文字，並隨著時代的演進而成爲一個由複雜結構、多筆劃及大量字符所組成之文字系統。由於漢字並非拼音文字體系，其發音學習的門檻很高，無法如英文般藉由認識各字母的發音方式即可唸出所組成之字彙，使得本國學齡兒童及外國學習者在初學漢字發音時無一有效的發音規則可循，必須仰賴注音或漢語拼音輔助漢字的學習。因此如何探勘漢字發音規則，以提供一個有系統的學習模式，引導漢字學習者了解中文發音規則是本研究第二階段的主要目標。

爲了要產出易懂的發音規則，讓中文的學習則可以應用形聲字的特性來推測漢字的發音，在本計畫中我們應用關聯規則探勘，做爲探勘聲符與其形聲字發音所存在的規則。關聯規則探勘原本的目的是從超市購買交易記錄的資料庫中，利用機率來評估產品之間被購買的關聯程度，其主要依據爲支持度(support)及信賴度(confidence)。其中支持度代表一個規則的涵蓋率（全部交易資料中有多少百分比讓規則爲真），而信賴度則代表一個規則的準確率（前提爲真的情況下，有多少百分比資料讓結果也同時爲真）。

我們將每一個形聲字視爲一筆交易，分別記錄形聲字之聲韻調、以及其相對應聲符的聲韻調，再以部首、形聲字連接方式、聲韻調的轉變與否等，做爲中文形聲字發音規則探勘的項目。另外我們也將形聲字筆劃(Stroke)、聲符筆劃、兩者差值等列入特徵範圍。同時考慮到記憶的方便性，我們統計了漢字構型資料庫中所有的漢字筆劃數將其平分爲三類：L16 代表大於或等於 16 劃，s11 代表小於或等於 11 劃，居中則由 12-15 筆劃爲範圍（見表 2）。

有了形聲字的交易資料後，接著使用關連探勘軟體 weka 來進行常用形聲字探勘其發音規則。我們將最小支持度取 1%、0.5%、與 0.3%對應各種不同的最小信賴度(60%~100%)進行 Apriori 運算後，得到不同數量的規則數統計如表 3。

應用一般關聯規則探勘雖然相當方便，但是可能找到相當多不符合我們預期的規則，因此如何過濾並篩選重要的規則，是此處我們必須要解決的問題。過濾的方法主要是刪除對於形聲字發音並無幫助的規則，舉例而言，“若聲符位置在右，則形聲字連接方式爲左右連接”，這樣的規則對形聲字發音的推測其實並沒有幫助。又如“若形聲字聲母發音爲ㄅ，則其聲符聲母發音爲ㄅ”，這樣的規則也無助於推測發音。爲了讓學習者在具備基礎聲符的發音能力下，利用對聲符的相關認知，來推測出更多尚未認識的形

聲字發音，因此有用的規則的結論應該具備形聲字發音（項目 1-3）或是發音不變等內容（項目 15-17），同時規則前提則應不包括前述這些項目。根據上述的條件，對規則的結論與前提分別進行過濾之後，我們統計出表 4 的規則數。

表2：形聲字發音規則探勘項目表

No.	符號	意義	數值範圍	範例： 炮
1	INITIAL	聲母	{ø, ㄅ,ㄆ,…,ㄙ}	ㄆ
2	FINAL	韻母	{ø, ㄧ,ㄨ,…, ㄧㄥ}	ㄠ
3	TONE	調號	{1, 2, 3, 4, 5}	4
4	CONN	形聲字的連接方法	{單體字, 左右連接, 上下連接,包圍式,其他}	左右
5	RC	部首	部首	火
6	PC	聲符	聲符	包
7	PCLOC	聲符所在形聲字之位置	{左,右,上,下,內,其他}	右
8	PCI	聲符的聲母	{ø,ㄅ,ㄆ,…,ㄙ}	ㄅ
9	PCF	聲符的韻母	{ø,ㄧ,ㄨ,…, ㄧㄥ}	ㄠ
10	PCT	聲符的調號	{1, 2, 3, 4, 5}	1
11	WS	形聲字筆劃數	{L16, 14-15, s11}	s11
12	PCS	聲符筆劃數	{L16, 14-15, s11}	s11
13	RCS	部首筆劃數		
14	WS-PCS	形聲字與其聲符筆劃差值	{s2, 4-5, L6}	s2
15	IU	聲母不變	{false , true}	false
16	FU	韻母不變	{false , true}	true
17	TU	調號不變	{false , true}	false

表3：關聯規則數

Conf / Sup	60%	70%	80%	90%	100%
1%	304,330	217,346	143,301	87,324	50,054
0.5%	1,573,613	1,149,779	802,029	500,708	314,523
0.3%	6,625,518	5,144,742	3,879,619	2,809,951	1,810,585

表4：過濾後關聯規則數

Conf / Sup	60%	70%	80%	90%	100%
1%	13,470	6,340	1,889	505	42
0.5%	61,171	32,089	15,243	7,561	5,190
0.3%	368,810	272,957	195,735	152,152	106,740

爲了讓使用者能根據不同條件快速篩選發音規則，我們也製作了「形聲字發音規則查詢系統」及「形聲字查詢系統」。圖 1 爲形聲字發音規則查詢介面。網頁載入時，是採用動態呈現條件選單內容，因此第一次載入網頁時等待時間較長（約 20 秒），而後選擇條件時系統會透過 Ajax 的方式傳送搜尋條件至伺服器端撈取相關規則與分群。查詢過程中，左下角的「下載中…」字樣會表示資料正在回傳中，右上角兩個選項則是開啓「形聲字標記系統」及「構件發聲強度列表」的連結。

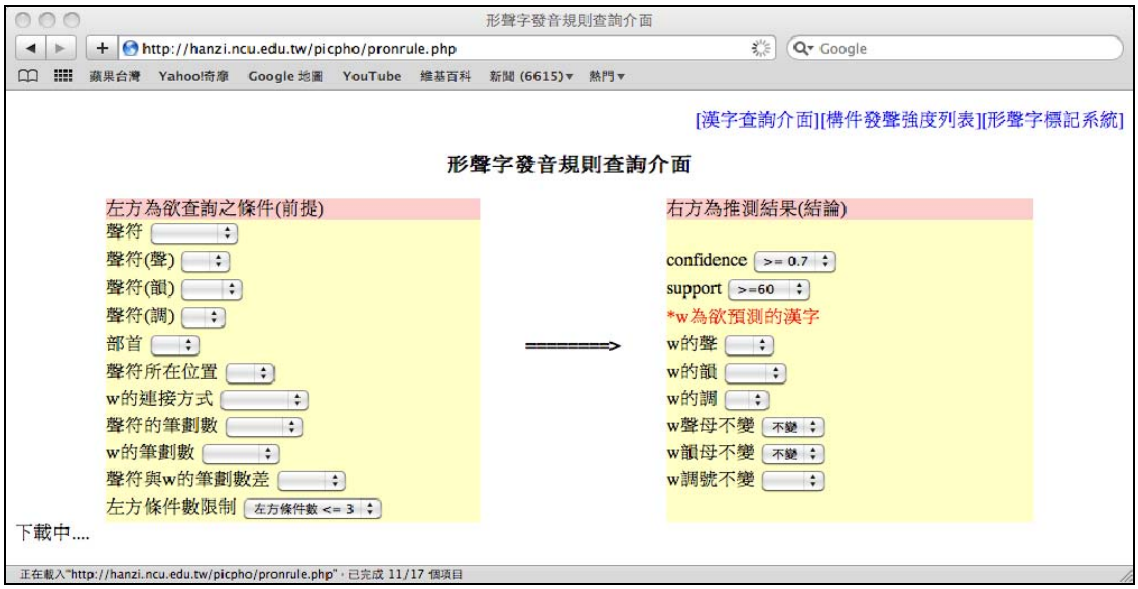


圖1 形聲字發音規則查詢介面(http://hanzi.ncu.edu.tw/picpho/pronrule.php)

查詢系統主要依據前述「形聲字發音規則探勘項目集」的特徵爲查詢條件，左邊是已知條件，右邊是推測結果，w 代表欲推測發音之形聲字。「左方條件限制」的功能則可篩選過長的規則。太長的規則通常不利於人們記誦，因此本系統預設條件數小於等於三。其他預設查詢條件爲「信賴度 >= 70%」,「支持度 >= 60」。當「下載完成」出現後，所有符合條件篩選的規則，經由分群後會顯示在介面的最下方，不同性質的規則會用不同底色做區隔（如圖2）。每條規則中都有可供細項查詢的連結。當我們選擇[查看]連結，則可顯示滿足整條規則（包含左方前提及右方結果）的形聲字；除此之外，[例外字]則可查出究竟有哪些形聲字是符合左方前提，但是不符合右方結果的形聲字。其他連結則可顯示符合單一條件的形聲字，如[部首=木]連結，可顯示出所有部首爲木的形聲字，[聲符的聲母=ㄌ]可找出所有聲符聲母是ㄌ的形聲字。

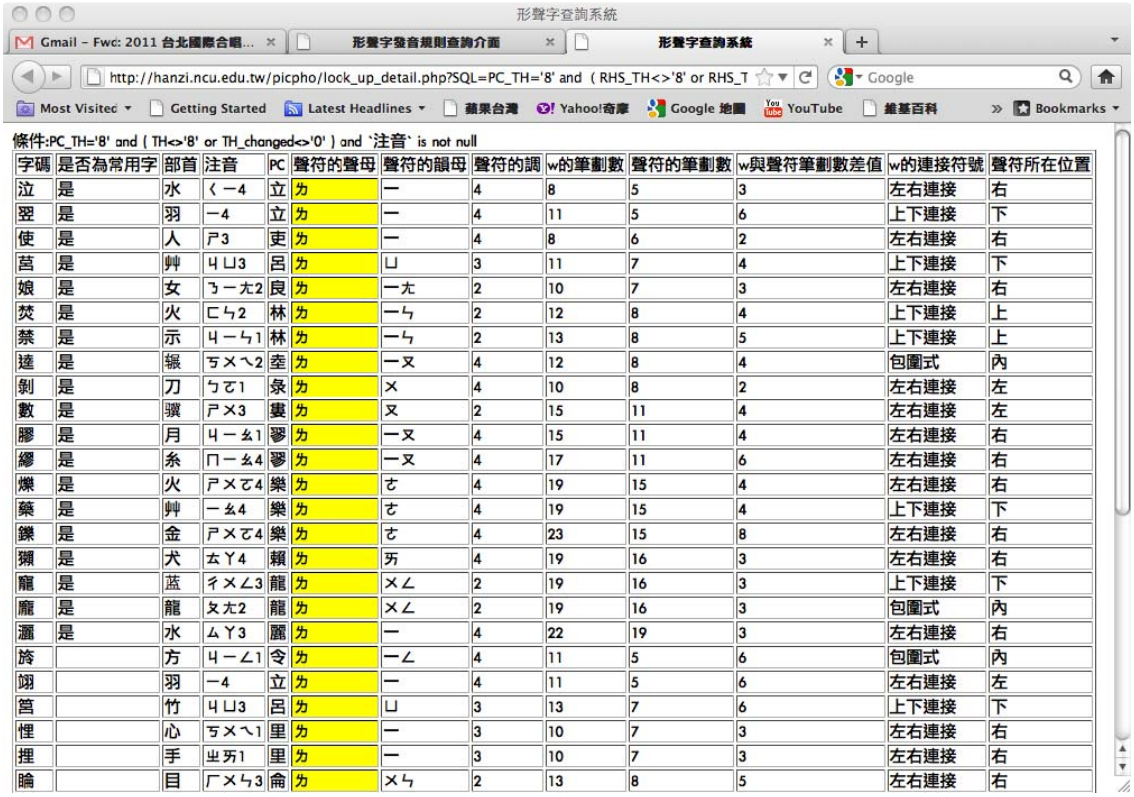
圖 2 形聲字發音規則分群結果

有了以上形聲字發音規則查詢系統，我們即可設定所需條件，找出相關發音規則。舉例來說，高支持度 3%、信賴度 80%且聲母發音不變的條件下的規則共有 15 條，共分成 3 組，如 R1-R3 所示。規則一說明聲符的聲母若爲ㄌ的前提下，形聲字的聲母發音將維持ㄌ（聲符範例：力令立列老利呂良里來侖兩戾拉林垒麥条刺郎栗留婁累連勞量廉虜雷豊劉閭厲慮樂閭魯畾歷盧賴龍闌羅麗蘭覽）；規則二顯示聲符的聲母若爲ㄇ的前提下，形聲字的聲母發音將維持 ㄇ（聲符範例：木末毛母民冊目矛名牟米免每孟明門冒某眉眇美苗面冥迷莽莫麻悶買閔滿蒙貌麼磨瞢彌）；規則三則敍述聲符的筆劃數若大或等於 16 以上，則形聲字的發音也多維持原本聲符的聲母發音（聲符範例：冀贏歷燕盧磨穌縣羲翰蕭謁賴頻龍瞢裹嬰彌毚爵襄闌隱霜鮮鐵瞿聶轉離魏確羅藝贊顛麗嚴藺蘇覺矍籣覽霸彎），不過第三個規則，由於筆劃數高，對於初學者來說幫助不大。另外使用者也可以查詢例外字，了解符合前提(聲符的聲母=ㄌ)但是聲母發音卻改變的形聲字（見圖 3）。

(R1) 聲符的聲母=ㄌ (supp:197) ======>聲母發音=不變 (supp:178, conf:0.9) [查看],[例外字]

(R2) 聲符的聲母=ㄇ (supp:128) ======>聲母發音=不變 (supp:105, conf:0.82) [查看],[例外字]

(R3) w的筆劃數=L16 , 聲符的筆劃數=L16 (supp:123) ======>聲母發音=不變 (supp:98, conf: 0.8)



條件:PC_TH='8' and (TH<>'8' or TH_changed<>'0') and `注音` is not null

字碼	是否為常用字	部首	注音	PC	聲符的聲母	聲符的韻母	聲符的調	w的筆劃數	聲符的筆劃數	w與聲符筆劃數差值	w的連接符號	聲符所在位置
泣	是	水	ㄑㄧ4	立	ㄌ	ㄧ	4	8	5	3	左右連接	右
翌	是	羽	ㄧ4	立	ㄌ	ㄧ	4	11	5	6	上下連接	下
使	是	人	ㄕ3	吏	ㄌ	ㄧ	4	8	6	2	左右連接	右
莒	是	艸	ㄐㄩ3	呂	ㄌ	ㄩ	3	11	7	4	上下連接	下
娘	是	女	ㄋㄧㄤ2	良	ㄌ	ㄧㄤ	2	10	7	3	左右連接	右
焚	是	火	ㄈㄣ2	林	ㄌ	ㄧㄣ	2	12	8	4	上下連接	上
禁	是	示	ㄐㄧㄣ1	林	ㄌ	ㄧㄣ	2	13	8	5	上下連接	上
達	是	辵	ㄎㄨㄟ2	奎	ㄌ	ㄧㄡ	4	12	8	4	包圍式	內
剝	是	刀	ㄅㄛ1	彔	ㄌ	ㄨ	4	10	8	2	左右連接	左
數	是	攴	ㄕㄨ3	婁	ㄌ	ㄡ	2	15	11	4	左右連接	左
膠	是	月	ㄐㄧㄠ1	翏	ㄌ	ㄧㄡ	4	15	11	4	左右連接	右
繆	是	糸	ㄇㄧㄠ4	翏	ㄌ	ㄧㄡ	4	17	11	6	左右連接	右
爍	是	火	ㄕㄨㄛ4	樂	ㄌ	ㄛ	4	19	15	4	左右連接	右
藥	是	艸	ㄧㄠ4	樂	ㄌ	ㄛ	4	19	15	4	上下連接	下
鑠	是	金	ㄕㄨㄛ4	樂	ㄌ	ㄛ	4	23	15	8	左右連接	右
獺	是	犬	ㄊㄚ4	賴	ㄌ	ㄞ	4	19	16	3	左右連接	右
龐	是	龍	ㄔㄨㄥ3	龍	ㄌ	ㄨㄥ	2	19	16	3	上下連接	下
龐	是	龍	ㄆㄤ2	龍	ㄌ	ㄨㄥ	2	19	16	3	包圍式	內
灑	是	水	ㄙㄚ3	麗	ㄌ	ㄧ	4	22	19	3	左右連接	右
旂		方	ㄐㄧㄥ1	令	ㄌ	ㄧㄥ	4	11	5	6	包圍式	內
翊		羽	ㄧ4	立	ㄌ	ㄧ	4	11	5	6	左右連接	左
筥		竹	ㄐㄩ3	呂	ㄌ	ㄩ	3	13	7	6	上下連接	下
悝		心	ㄎㄨㄟ1	里	ㄌ	ㄧ	3	10	7	3	左右連接	右
裡		手	ㄓㄞ1	里	ㄌ	ㄧ	3	10	7	3	左右連接	右
瞷		目	ㄏㄨㄣ3	侖	ㄌ	ㄨㄣ	2	13	8	5	左右連接	右

圖3 查看發音規則例外字

又如查詢高信賴度 100%、支持度 0.5%、且聲母與韻母均未改變的規則，可得 34 條符合條件的規則，分成 5 組，如 R4– R8 所示。規則左方的支持度表示滿足左方條件的常用形聲字，規則右方的支持度則爲滿足整個規則的常用形聲字。舉例來說 R7 說明聲符的聲母爲ㄒ、聲調爲一聲且聲符筆劃數小於等於 11 的時候，則衍生形聲字的聲母與韻母均不改變；符合這條規則的形聲字中包含的聲符包括「希」、「析」、「宣」、「星」、「相」、「胥」、「奚」等衍生的 16 個常用形聲字。不過使用者若是查看符合規則的形聲字，則同時可以看到其他符合條件的非常用形聲字，如「心」、「先」、「西」、「析」、「欣」、「香」、「悉」、「脩」等聲符所衍生的形聲字。規則 R8 則說明當聲符的韻母爲ㄤ、聲調爲一聲、聲符筆劃數小於等於 11 且聲符與部首爲左右連接的時候，則衍生形聲字的聲母與韻母均不改變；符合這條規則的形聲字中包含的聲符包括「方」、「邦」、「岡」、「昌」等衍生的 17 個常用形聲字。

(R4) 聲符的聲母=ㄌ , 聲符的調=3 , w與聲符筆劃數差值=s3 (supp:16)
=====>聲母發音=不變 , 韻母發音=不變 (supp:16, conf:1) [查看],[例外字]
(R5) 聲符的聲母=ㄌ , 聲符的調=2 , 聲符所在位置=右 , w的筆劃數=12-15 (supp:17)
=====>聲母發音=不變 , 韻母發音=不變 (supp:17, conf:1) [查看],[例外字]
(R6) 聲符的聲母=ㄌ , 聲符的筆劃數=L16, w與聲符筆劃數差值=4-5 (supp:16)
=====>聲母發音=不變 , 韻母發音=不變 (supp:16, conf:1) [查看],[例外字]
(R7) 聲符的聲母=ㄒ , 聲符的調=1 , w的筆劃數=12-15 , 聲符的筆劃數=s11 (supp:16)
=====>聲母發音=不變 , 韻母發音=不變 (supp:16, conf:1) [查看],[例外字]
(R8) 聲符的韻母=ㄤ , 聲符的調=1 , w的連接符號=左右連接 , w的筆劃數=s11 (supp:17)
=====> 聲母發音=不變 , 韻母發音=不變 (supp:17, conf:1) [查看], [例外字]

五、結語

本計畫的目標係利用以聲符爲主的部件教學，將構詞能力很強的部件放在課程的前面，發揮「以簡馭繁」、「快速掌握形聲字的結構」等部件教學的優點。當學習者明白能夠利用字形線索來學習新的生字時，其於漢字學習上的自學能力便能提升。在整字教學的部份，將透過部件的合體字進行引導學生瞭解從部件建立字形的標準，並且讓學習者得以快速掌握形聲字的結構。接著，以系統性的方式安排每課的整字教學與構詞活動，藉由字音聯想及以字代詞的方式來編排教材，未來也將實地進行部件教學學習效益的評估。

目前有關部件發音強度的計算，以及形聲字發音的關聯規則雖已完成，但是對於輔助以聲符爲主的部件教學教材編輯，仍有不足之處，如何能充份應用部件的組字功能及聲符部件的發音強度，做爲華語識字教學順序的參考，這也是第三階段計畫成敗的關鍵。幾個未來研究方向簡略如下：

- 由於關聯規則探勘可能找到相當多的規則，而且某些規則可由其他規則代表，因此如何進一步過濾關聯規則，並找出一組最重要的規則涵蓋愈多的常用字，是此處我們必須要解決的問題。
- 統計數字指出 7000 個通用字中總共有 246 個意符，5631 個形聲結構中包含了 1325 個不同的聲符。另外，246 個意符中的 54 個構字能力很強的意符，構成了 4898 個形聲結構，約佔形聲結構總數的 87%(陳原《現代漢語用字資訊分析》，上海教育出版社 1993)。如何折衷構字能力強與發音強度，篩選或排序聲符部件則是另一個重要的研究議題。
- 漢字教學步驟通常爲先教獨體字，再教簡單合體字，最後教複雜合體字。但並非每個部首和任何聲符都可組成合體字，對初學者而言，可能出現偏旁部首張冠李戴的情形。如何幫助學習者釐清這些差異，也是挑戰之一。

參考文獻

[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。

[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。

[4] 董鵬程，台灣華語文教學的過去、現在與未來展望. 2007多元文化與族群和諧國際研討會，台北教育大學。 http://r9.ntue.edu.tw/activity/multiculture_conference/memoirs.html。

[5] 許聞廉、呂明蓁、胡志偉、柯華葳、辜玉旻、呂菁菁、張智凱、莊宗嚴，構建一個新移民者有機成長的多元認同平台的整合研究（期中進度報告），2009– 2011。

[6] 高柏園、郭經華、胡映雪，華語文作爲第二語言之字詞教學模式與學習歷程研究，2009-2010。

[7] 洪文斌，華語文作爲第二語言之字詞教學模式與學習歷程研究– – 子計畫一：中文字部件拆解教學模式與電腦輔助學習系統之研發（期中進度報告），2010。

[8] 張嘉惠, 李淑瑩, 林書彥, 黃嘉毅, 陳志銘，《以最佳化及機率分佈判斷漢字聲符之研究》，ROCLING XXI，2010。

[9] 萬雲英，《兒童學習漢字的心理特徵與教學》，載於楊中芳、高尚仁主編，中國人、中國心－發展與教學篇，403-448。台北：遠流。

[10] 陳原（1993），《現代漢語用字資訊分析》，上海教育出版社。

[11] 盛繼豔，《華文教學中漢語的部件教學》。

[12] 梁彥民《漢字部件區別特徵與對外漢字教學》，《語言教學與研究》2004。

[13] 李思維、王昌茂編著，《漢字形音學》， 武漢：華中師範大學出版社，2000 年版。

[14] 中研院文獻處理實驗室，「漢字構形資料庫」網站。

中 華 民 國 計 算 語 言 學 學 會

一〇一年度會員資格更新暨個人資料異動單

一、 **姓名：**_____會員別：□終身　□個人　□學生

二、 **個人資料異動（請填寫異動部分）**

學歷：_____

現職：_____

地址：_____

電話：_____

E-mail：_____**（請務必提供）**

會員別：□ 個人會員轉終身會員　　□ 學生會員轉個人會員

三、 **會費繳交：**

1. 會費：□終身會員：10,000.-　□個人會員：1,000.-　□學生會員：500.-
2. 繳費方式：
   - □ 郵政劃撥（帳號：19166251 號，戶名：中華民國計算語言學學會）
   - □ 信用卡（請加填信用卡資料）
3. 繳費期限：一〇一年度會費請於 100 年 7 月 31 日前繳交

註：終身會員及今年新加入之會員，勿需再繳交會費。

四、 **101 年度會費有效期間：100 年 7 月 1 日至 101 年 6 月 30 日。**

五、 **信用卡繳費**

姓名：_____ (請以正楷書寫)

卡別：❑ VISA CARD　❑ MASTER CARD　❑ JCB CARD

卡號：_____—_____—_____—_____

有效日期：_____(M/Y) 卡片後三碼：_____發卡銀行：_____

金額：_____ 持卡人簽名：_____

六、 **上述資料填妥後，請傳真或郵寄至本會：**

會址：台北市 11529 南港區研究院路 2 段 128 號中研院資訊所（轉）
電話：(02)2788-3799 轉 1502，傳真：(02)2788-1638
聯絡人：黃琪 小姐 E-mail：aclclp@hp.iis.sinica.edu.tw

工研院資通所前瞻技術中心

文字轉語音合成技術

ITRI TTS@Web

～免費語音合成線上服務，助您輕鬆實現創意網頁～

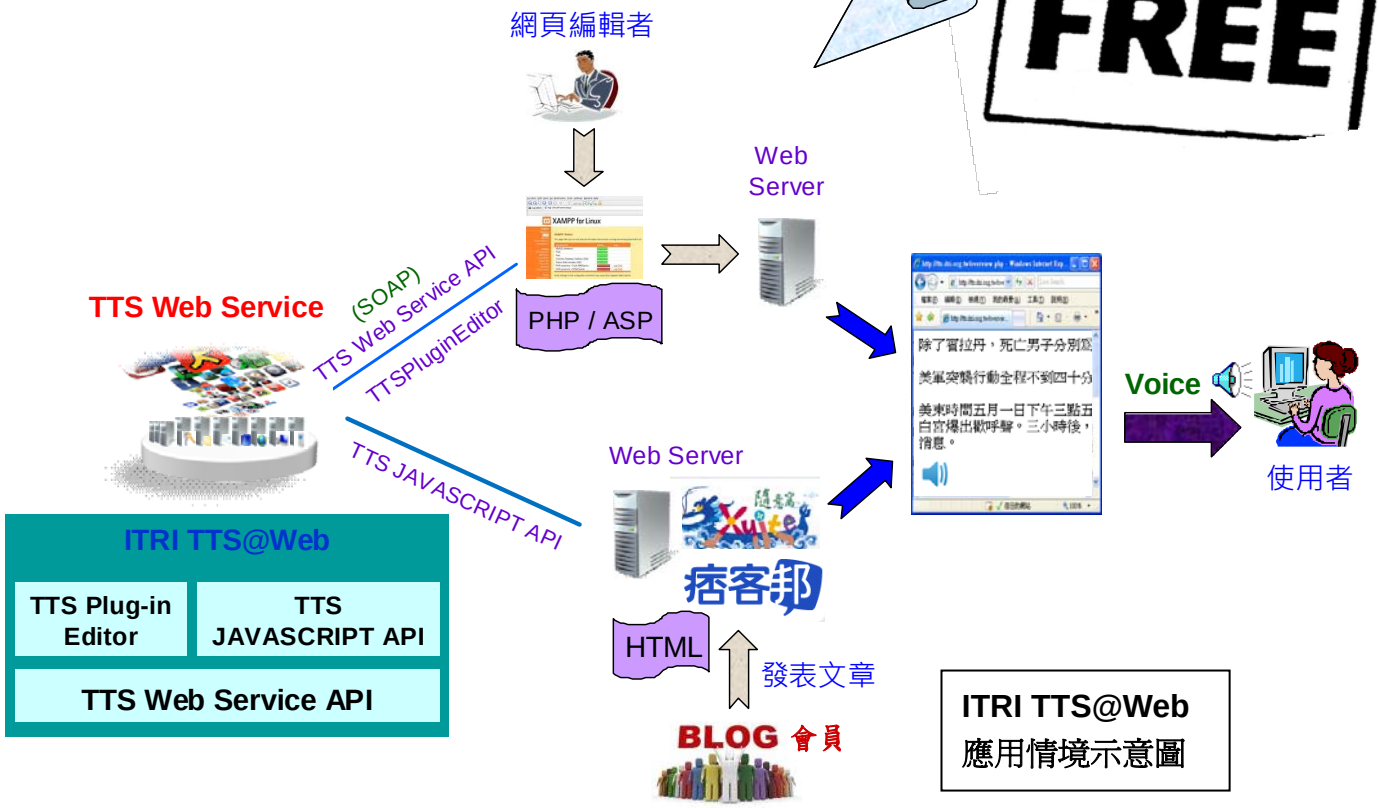
工研院文字轉語音（TTS: Text To Speech）web 服務（web service）可讓您在自己的網頁提供 TTS 體驗(服務網址 http://tts.itri.org.tw/)。

可使用 PHP、ASP(**透過 TTS Web Service API 與 TTS Plugin Editor**)與 Javascript(**透過 TTS JAVASCRIPT API**)等進行應用網頁撰寫。本服務使用 UTF-8 之格式，伺服器提供 SOAP（Simple Object Access Protocol）協議的 Web Service。

所謂的文字轉語音，是將所輸入的文字，轉換為合成語音進行輸出。早期的文字轉語音合成，會有機械音、或韻律不流暢的缺點。目前工研院所研發的中文文字轉語音技術，所合成的語音自然流暢、近似真人發音。

本服務提供之 TTS 有以下特點：

- 可輸入任意中文文字（正體、簡體）
- 可產生自然流暢的合成語音
- 可選擇不同語者聲音
- 可調整說話韻律：音調高低、速度快慢、音量大小
- 在線即時轉換，不需安裝程式



ITRI TTS@Web 應用情境示意圖

工業技術研究院
Industrial Technology Research Institute

For technology transfer about the TTS engine, please contact
程懷妥(Huai-Tuo Chen), TEL:+886-3-5914463, E-mail: Jethro@itri.org.tw

For technology support about the ITRI TTS@Web, please contact
彭先生, TEL:+886-3-5914070, E-mail: pengchikai@itri.org.tw
洪先生, TEL:+886-3-5914067, E-mail: tdy@itri.org.tw