線上國客雙語有聲詞典建置之研究

On-line Phonetic Dictionary for Chinese and Hakka Language

黃豐隆
聯合大學資工系，苗栗市
flhuang@nuu.edu.tw

余明興
中興大學資工系，台中市
msyu@nchu.edu.tw

吳俊毅
中興大學資工系，台中市
jalen2183@gmail.com

摘要

本文以自然語言處理技術為基礎，建構線上國客語的有聲詞典，作為國客雙語數位學習之平台。此系統具有五個模組，包含：輸入詞彙查詢，文字分析，韻律預估、選取合成單元與語音合成。我們錄製中文和客語四縣腔的基本語音合成單元，用以產生這二種語語之合成輸出，建立相關之語彙詞典與語音庫，並提供詞典中的詞義、例句、國客語用法對照，經線上實際測試，國客語具有真人之語音特性與清晰度，可有效協助使用者學習國客雙語，提升聽與說之能力。

關鍵詞：自然語言處理，有聲詞典，語音合成，客語連音變調，數位學習。

一、前言

詞典(Dictionary)是學習語言的重要工具，世界上多數人使用的語言，如中文、英文、日文與西班牙文等，均有內容完整的詞典。隨著網際網路之發展，近年來有許多線上(on-line)詞典出現[1][2][3][4][5]，除了較多世人使用的中文與英文之外，還有一些較少數人使用的詞典，含閩南語[2]與客家語[5]等，這些線上的詞典具有不受時空限制與重覆學習的優點，提供數位學習(e-Learning)的環境，使語言的學習邁向更高、更有效率的層次。

1.1 語言學習－聽說讀寫

基本上，語言的學習需考量「**聽說讀寫**」等能力，一個良好線上詞典需具有這些功能，具備語言「**聽說**」基礎後，進而培養讀寫能力，將使語言學習達到事半功倍的目的。因此，具有數位學習的線上有聲詞典一直是國內外許多研發單位努力追求的目標。

台灣目前所使用的本土語言，包含有國語、閩南語、客語與原住民語言，十分多元，國語是官方語言，民間日常生活之溝通則普遍使用國語、閩南語與客語等語言。在台灣國語已成為一種相對強勢的語言，根據台灣客語之**調查報告**，有礙客語傳承之主要因素為「**不太會講**」，具有「**聽、說**」客語能力逐年下降，因而保存純正客家語音與語文的工作益加重要。如何有效學習國客語，成為認識、傳續文化的基礎工作。學習者具有國客語「**聽、說**」能力，即可進一步達到「**讀、寫**」例句與文章的層次。

在台灣使用國語的人口較廣，客語則屬相對弱勢的語言，其使用人口亦較少，因此，以國語搭配客語，有助於客語之學習，進而相得益彰，提昇語言之學習成效。本系統期望建立一整合國語(Chinese)與客語(Hakka)之有聲詞典(Phonetic Dictionary)，主要目標即：在網際網路(Internet)上，建置一個學習國客語有聲詞典之數位平台，提供學習「**聽、說**」國、客語詞彙的環境。此數位學習平台可以提供國語之學習與準備客委會客語認證者、以及想認識中文與客語家文化之人士，而國際人士能使用拼音標注功能，瞭解客語

之發音與語彙、語義。此外，本系統亦提供國客語之語文資訊，如國語聲調變化、客語拼音與連音變調之規則， 將有助於學習國客語。

語言裏的詞(Word)是文句中有意義且可自由使用的最小語彙單位(Unit)，因此學習語言時，可以詞彙開始，再配合相關之例句，可達事半功倍的效 果。本系統提供國語與客語之詞彙查詢，這二種語言均提供文句之語彙資訊，如拼音與注音，並有英文與例句。為了達到「**聽、說**」的目的，國語與客語詞彙均具有由語音輸出，可以反覆聆聽語言，以提昇「**聽、說**」的語言能力，進而認識中文與客家文化之美。

1.2 線上有聲詞典

行政院客委會客家語言能力認證網站提供的客語字典查詢，其發音為事先錄製之詞彙與例句，並非使用語音合成。此方法有個缺點，系統管理員需要把詞典中所有詞彙都用人工錄音一遍，若加入新詞或詞典修改，都要重新再錄，詞典收錄詞數越多，情況會越嚴重。如此一來維護此系統變成相當棘手的工作。另外，若要發展成用單音合成的文字轉語音系統，則這些語音檔案都無用武之地。至於，教育部網站的臺灣客家語常用辭典，僅顯示拼音，並未提供線上之語音功能。另一個常用線上的翻譯軟體 Dr.eye，雖有中、英文詞彙的語音，惟沒有提供中英文例句之語音。本文所建置的系統期望能改善上述這些系統之缺失，針對國客語之語言學習，提供使用者「聽說」之學習功能，表 1 列出相關系統之比較。

表 1：線上詞典的特性比較

	運用語音合成技術	中文有聲例句	整合國客雙語	擴充性
教育部[1]	無	無	無	較低
客委會[5]	無	有	無	較低
Yahoo 字典	無	無	有	較低
Cambridge[3]	無	無	無	無
本系統	有	有	有	較高

我們的系統使用單音語音合成，若加入新詞或是修改詞典，系統都能自動合出發音，不必再人工錄音，甚至發展成文字轉語音系統也非常方便，故擁有較佳的擴充性。

由於客語之腔調有四縣、海陸、永定與大埔腔等，本系統以台灣使用人口最多的四縣腔為主，未來將再擴充至其它的客語腔調。

本文架構如下;第二節說明自然語言處理技術與線上有聲詞典簡介，第三節介紹我們所建置的國客雙語有聲詞典的環 境與主要的功能模組，第四節說明系統實作與測試，最後為結論與未來研究方向。

二、自然語言處理

2.1 自然語言處理簡介

電腦科技日新月異，經過數十年發展，已成為生活中不可或缺的一部分。在分秒必爭的社會中，人們總希望能更方便快速地操作電腦，而最方便的方法就是直接用語音操控。若要達成這個目標，必需讓電腦具備語音辨識的能力，並且能夠了解語意，然後才能執行命令。另一方面，想要讓電腦說話，要靠文字轉語音系統。電腦不像人類，可以直接讀出語言的讀音，推想其語意，甚至詞性。若要讓電腦具備上述能力，必需先進行自然語言處理。

自然語言處理的範圍相當廣，其中與本系統相關研究包括文句分析與韻律處理[6][7][8][9]。文句分析包括斷詞與未知詞辨識，即一般常說的構詞。韻律處理則包括產生韻律訊息與連音變調兩部分。本系統首先將使用者輸入的詞送至國客對照詞典，查出客語詞之發音與例句。再對其例句做文句分析與韻律處理。

2.2 語音合成技術

本系統所使用的四縣客語帶聲調單音總共

約有 2500 個不同的音檔，聲調有六種，即高、低、升、降，加上入聲的高、低。每個音檔以客語拼音開頭，接上阿拉伯數字音調。目前實驗室尚無足夠的客語語料庫來訓練韻律訊息，且語音合成方面僅用語音單元做接合。單音合成方面，最常用的語音合成技術是基頻同步疊加法(PSOLA)，PSOLA 分為三類 TD-PSOLA、FD-PSOLA，與 LP-PSOLA。PSOLA 常使用於調整音長與音調，因為 TD_PSOLA 方法簡單且效果不差，常被實作於電腦應用與嵌入式系統。

三、系統架構

3.1 系統架構

本系統網際網路線上之架構圖，請參閱圖 1，使用者經由網際網路登入系統，使用線上七聲詞典。

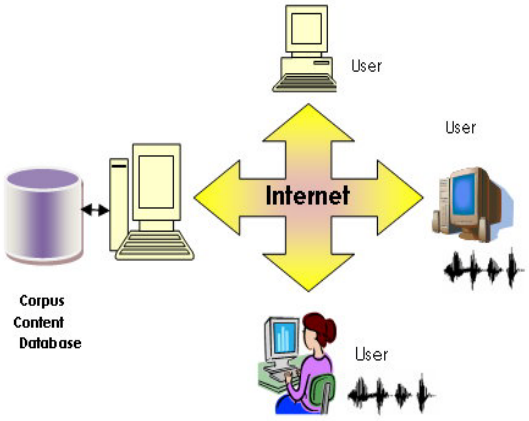

圖 1: 線上國客語有聲詞典網路環境

3.2 功能模組

本系統的模組參見圖 2 所示，包含有下列五大模組：

(A)詞彙輸入、查詢(Input, Query):

(B)文句分析(Text Analysis)：含線上中文斷詞功能

(C)韻律預估(Prosody Prediction)：含標點符號與語音時長分析。

(D)選取合成單元 (Selection of speech Uuits) ：選用合適之語音檔

(E)語音合成(Speech Generation)：合成產生正確之詞彙語音

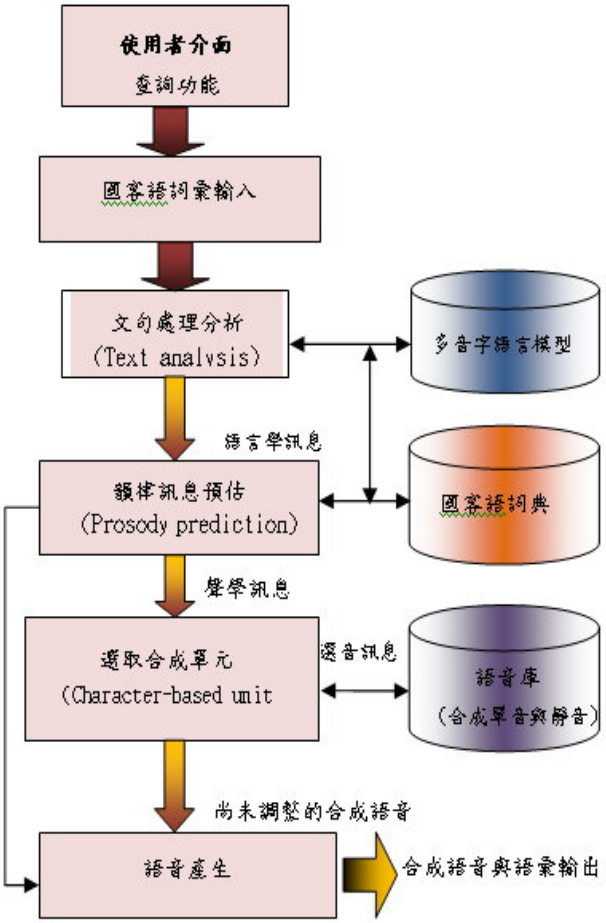

圖 2: 架構模組

3.3 雙語詞典

本系統共有兩種詞典，分別為國語十三萬目詞，與客語詞彙資料庫。以下為其介紹：

(A)國語十三萬目詞(ASCED)：

此為中研究之國語詞典，含有近 13 萬國語之詞目，以及其詞性與詞頻等訊息，中文例句作國語斷詞(Segmentation)時之比對依據，並

可取得詞彙之注音資訊。

(B)客語詞彙資料庫：

目前約有 3 萬詞目，主要來源含教育部、行政院客委會、台北市客委會等單位，並有一部份自建。每一詞目含有四縣腔與海陸腔資訊，及其拼音與例句。本系統所使用之客語拼音與聲調表示，係依據教育部 98 年修正之版本。

3.4 語音庫

本系統包含了三種語料庫，分別為多音字語料庫、國語合成單元語音庫、客語合成單元語音庫。以下為其介紹。

(A) 多音字語料庫：

訓練語料後建立語料庫(Corpus)，以統計式之語言模型(Language Models, LM)方法預測國客語多音字(Polyphones)之正確發音類別，作為合成語音時選取基本合成語音單元之依據。

(B) 國語合成單元語音庫：

錄製國語基本之合成單元，以中文單音節之語音為單位，含四聲之變化，共有 1400 個基本單元，此外還錄製相關之靜音之語音約 50 個。

(C) 客語合成單元語音庫：

錄製國語基本之合成單元，以中文單音節之語音為單位，含四縣腔之 6 種聲調變化，共有 2500 個基本單元。此外，還錄製相關之語音約 50 個，以及不同時長之靜音檔。

3.5 斷詞簡介

斷詞(Word Segmentation)是文句分析中第一個步驟。斷詞的方法有很多，有長詞優先、少詞優先與機率等等。本系統目前使用基本的長詞優先來作為斷詞模型。方法為：把例句從頭到尾切成許多字串，將字串一一傳入詞典比對，若一字串存在於詞典中，就認定該字串可能為詞。然後由句首開始，每次看兩個詞計算詞長，將較長詞的第一詞輸出。依序往後做至句尾，最後所有詞皆輸出，即為斷詞結果。

抓出而未知詞與構詞都是不存在辭典中的詞彙，所斷出來的詞找不到對應之客語音後，還是必需拆成單字去決定讀音。現階段對於構詞與未知詞(Known word)的部分，本系統還未處理。

3.6 客語連音變調

此部分先介紹客語四縣、海陸調值與調號表示。四縣音有六種聲調，海陸音則有七種聲調(多「陽去」聲調)。以四縣腔為例，家話的聲調有六種，即:高、低、升、 降，加上入聲的高、低。

客語保留古代的入聲(國語則沒有入聲)，入聲發音較短促，惟只有高低二種。

上揚(陰平)：如「夫」， 類似國語第二聲，由國語的第一聲變來。

低平(陽平)：如「扶」， 類似國語第三聲，由國語的第二聲變來。

下降(上聲)：如「虎」， 類似國語第四聲，由國語的第三聲變來。

高平(陰去)：如「富」， 類似國語第一聲，由國語的第四聲變來。

低入(陰入)：如「福」。(入聲，低) 。

高入(陽入)：如「服」。(入聲，高)。

本系統採用調形符號表示客語之聲調，參見表 2，調號採用趙元任所創立之「5 等級制」，如圖 3 所示。

表 2：客語四縣調值與調號

調類	陰平	陽平	上聲	陰去	陰入	陽入	陽去
調號	24	314	31	55	21	5	
調形	fu↗	fu↘	fu↖	fu	fug↖	fug	
例字	夫	扶	虎	富	福	服	
國語	2聲↗	3聲↘	4聲↖	1聲			

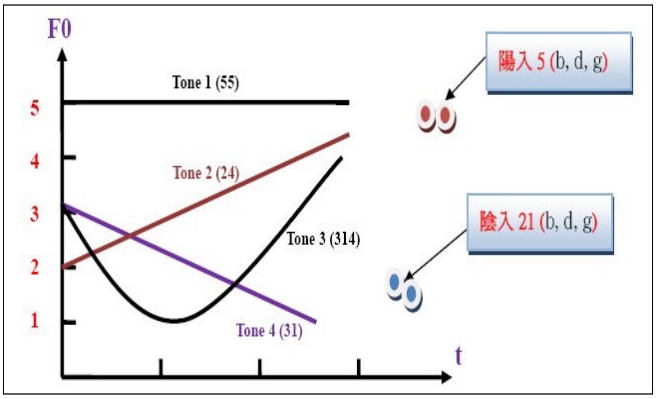

圖 3: 國客語不同聲調之音高。

同樣地，兩種腔調的連音變調規則亦有所有不同。客語四縣變調規則可規納出三個情況，如表 3 所示：

表 3：四縣客語之連音變調規則

規則 1：由兩個陰平（調號 ✓ ）字構成的字彙，讀時前字變調讀陽平（ ∨ ） 陰平 + 陰平 → 陽平 + 陰平			
範例	**詞彙**	**變調前之拼音**	**變調後之拼音**
	新衫	sin ✓ sam ✓	sin ∨ sam ✓
	買新衫	mai ✓ sin ✓ sam ✓	mai ∨ sin ∨ sam ✓
規則 2：陰平與陰去構成的詞彙，讀時前字變調讀陽平（ ∨ ） 陰平 + 陰去 → 陽平 + 陰平			
範例	**詞彙**	**變調前之拼音**	**變調後之拼音**
	針線	ziim ✓ sien	ziim ∨ sien
	拿針線	na ✓ ziim ✓ sien	na ∨ ziim ∨ sien
規則 3：陰平字與陽入字構成的詞彙，讀時前字變調讀陽平（ ∨ ） 陰平 + 陽入 → 陽入 + 陰平			
範例	**詞彙**	**變調前之拼音**	**變調後之拼音**
	音樂	im ✓ ngog	im ∨ ngog
	聽音樂	tang ✓ im ✓ ngog	tang ∨ im ∨ ngog

本系統提供四縣腔客語發音。故只處理四縣客語的連音變調部分。至於海陸音與其它腔調的部分，未來錄音工作完成之後進行處理，再擴增至系統中。

四、系統實作與測試

4.1 系統實作環境

本系統開發環境(使用之軟體)如表 4 所示，Context server 建置在網路上的主機。

表 4：系統開發環境

Software	Version
Operating System:	Windows XP
Web Server:	Apache Server v2.2.4
Database:	MySQL v5.0.45
DBMS platform:	phpMyAdmin v2.10.2
Web Programming:	PHP5
FTP Server	FileZilla Server

4.2 查詢之畫面說明

首先，在瀏覽器網址列輸入以下網址(http://203.64.183.226/public2/hakka-edu/index.html)連到線上國客語有聲詞典網站。看過開頭動畫後，會自動導向到有聲詞典查詢頁面。如圖 4：

圖 4：有聲詞典查詢頁面

畫面中央的文字欄位，用於輸入欲查詢的字詞。輸入文字可為國語、客語、英文，需自行選擇，預設值為國語。下方的選擇方塊可讓使用者自行選擇想要的資料，預設值為全選，即符合查詢的結果全部輸出。例如，輸入「音

樂」，按下 送出 鈕開始查詢。結果如圖 6：

圖 5：查詢「音樂」

圖 6：查詢「音樂」的結果

查詢結果畫面中，有數個按鈕， PLAY 按下後可以聽到合成出的聲音。

接下來，試試輸入客語詞。在查詢前，請先把欲查詢文字上方的語別選項，選 客語，再按 送出 進行查詢。以查詢「伯公」為例。如圖 7。

圖 7：查詢客語詞「伯公」

查詢結果與輸入國語詞相同，都會顯示出四縣客語、四縣拼音、海陸拼音、國語用詞、英文用語、客語例句、客語例句拼音、國語例句、國語注音與國語斷詞。如圖 8。

圖 8：客語詞「伯公」的查詢結果。

同樣地，輸入本系統也提供輸入英文詞查詢的功能。文字方塊上方選 英文 ，接著輸入英文詞，按下 送出 進行查詢。以查詢「sleep」為例。如圖 9

圖 9：查詢英文詞「sleep」

查詢結果也與輸入國語詞和客語詞相同，都會顯示四縣客語、四縣拼音、海陸拼音、國語用詞……等等，結果如圖 10。

圖 10：英文詞「sleep」的查詢結果

除了精確查詢，本系統還提供模糊查詢的功能。所謂模糊查詢，就是把所有與欲查詢字串相關的資料都輸出，如圖 11。

圖 11：進行模糊查詢

假設輸入中文詞彙「音樂」，輸出資料一樣預設全選，則輸出結果如圖 12。

圖 12：「音　樂」的模糊查詢結果

4.3 線上測試與未來改進方向

本系統建置完成後 經由三位具有國語與客語聽說流利之語文背景之專家實際測測，並提出相關之缺失，作為改進之依據，合成之國客語語音輸出具有真人之語音特性與清晰度，至於語音之可理解度(Comprehension)已達實用階段。

我們將持續改善相關的缺點，尚待改進、加強的部分有下列三項：

1、目前海陸腔客語部分，尚未錄完全部的音，故系統暫不支援海陸腔客語。未來錄音工

作完成之後，將加入至系統中。再配上原有的連音變調和語音合成功能，也能輸出聲音，使此詞典在「有聲」的方面更完善。

2、詞典內容仍需再充實。目前本系統詞典來源包含教育部、行政院客委會與台北市客委會，共兩萬九千餘個客語詞。其中並非所有詞典內容都包含四縣與海陸兩種拼音，有些沒有例句，或是沒有對應的英文詞。這些詞典資料不足之處，皆會造成使用上的不方便。未來會多整合數種詞典，使本系統的詞典更為充實。

3、目前本系統的語音合成功能沒有加入韻律訊息與停頓，故合成完的語音聽起來仍不像人類說話。未來加入韻律訊息後，將改善此情形，讓使用者能聽到滿意的語音為本系統的目標，並朝著此方向努力。

五、結論

本論文主要目標為;在網際網路上建置一個客語數位學習平台，提供學習「**聽、說**」客語詞彙的系統。此平台包含建置線上有聲詞典功能以及客語語音之數位學習平台，具有中文、客語與英文之檢索功能，並有中文、客語與英文例句之輸出與其語音合成輸出，我們自然語言處理技術為基礎，產生任意文句之合成語音。在客語之音韻方面還需進一步改進，且目前我們所建之詞典內容還要繼續增加，預估未來能增加至4 萬筆以上，將更具有實用性。

此數位學習平台可以提供準備客委會客語認證者以及想認識客家文化者，亦可有效學習客語之聲母與韻母發音與拼音，國際人士能使用拼音標注功能，瞭解客語之發音與語彙、語義，進而認識客家文化之美。

本系統的功能進一步可作為**客語文轉語音**之「**語音合成平台**」，提供相關客家語譯使用技術，如104查號台，有聲書與客語導覽系統等應用。未來研究項目將包含有:

1. 客語語義岐異處理(多音字、多音詞)。
2. 客語單字、拼音與詞典之擴充。
3. 客語斷詞處理與詞性標註。
4. 以客語詞為單位錄音與音韻處理。
5. 海陸腔之語音處理。
6. 建置任意文句之客語合成系統。

致謝

本文由客委會計劃補助、國科會部份經費補助，教育部、台北市客委會、與苗栗市文化局均提供國客語之語料，特此致謝。

參考文獻

[1] 重編國語辭典修訂本，中華民國教育部。http://dict.revised.moe.edu.tw/
臺灣閩南語常用詞辭典 http://twblg.dict.edu.tw/tw/index.htm，中華民國教育部

[2] Cambridge Advanced Learner's Dictionary http://dictionary.cambridge.org/

[3] Yahoo 奇摩字典 http://tw.dictionary.yahoo.com/

[4] 客家語言認證網站 http://kaga.hakka.gov.tw/ct.asp?xItem=43076&ctNode=1375&mp=100

[5] 中文文句轉臺語語音系統初步研究，蔡宗謀、余明興，2008 年。

[6] 陳信希，2001, 自然語言處理技術之應用：國台客語機器翻譯與語音合成系統，科學發展月刊, Vol. 29, No. 11, pp. 824.

[7] 林東毅，2006，客語文句翻語音系統之實作，交通大學碩士論文。

[8] 李雪貞，2001，客語語音合成之初步研究，台科大碩士論文。

[9] Gu, Hung-Yan, Yan-Zuo Zhou and Huang-Liang Liau, 2007, "A System Framework for Integrated Synthesis of Mandarin, Min-nan, and Hakka Speech", International Journal of Computational Linguistics and Chinese Language Processing, Vol. 12, No. 4, pp. 371-390.