行政院國家科學委員會補助專題研究計畫 □ 成 果 報 告 ☑期中進度報告

客語文句轉語音及語音辨認之研究(2/3)

計畫類別：☑ 個別型計畫　　□ 整合型計畫
計畫編號：NSC 96－2221－E－009－030－MY3
執行期間 97 年 8 月 1 日至 98 年 7 月 31 日

計畫主持人：陳信宏
共同主持人：余秀敏、羅烈師
計畫參與人員：蕭希群、楊智合、江振宇

成果報告類型(依經費核定清單規定繳交)：☑精簡報告　□完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
□出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：可公開查詢

執行單位：國立交通大學

中　華　民　國　98　年　5　月　30　日

摘要

本計畫使用中文語音合成及辨認技術，將現有的客語文字分析系統(Text Analysis)、語音合成系統與辨認模組做進一步的改進，改進項目包括：詞典資訊的擴充、客語 Text Analysis 斷詞正確性的改進、韻律產生器的改進、語音合成器的改進、語料庫的收錄與整理、以及建立以 word n-gram 為基礎的客語辨認用 Language Model，以強化客語語音合成與辨認的效能。

關鍵詞：客語語音合成、客語語音辨認

一、前言

現有的客語語音合成與辨認系統所共同面臨的狀況，就是詞彙資訊量及資料量的不足所造成的各種問題。在語音合成方面，詞典資訊量不足所造成的不良結果除了在針對文章作斷詞分析(Text Analysis)時無法得到正確的分詞結果外，對於合成聲音所需的韻律(Prosody)也會造成不順暢的預估。另外，訓練韻律產生器(Prosody Generator)以及實現 Corpus-based TTS 系統所需的語料若不足，亦會造成合成效果不如預期。在語音辨識方面，現有的客語語音辨認系統受限於語料庫的不足，無法建立良好的聲學模型(Acoustic Model)及語言模型(Language Model)，導致辨認效能無法提升。

二、研究目的

基於上述論點，本計畫除致力於建立客語語音合成及辨認所需的語料庫外，亦希望取用目前中文語音合成與辨認技術與資源[1-8]，輔助客語系統的建立。技術部分包含引用中文定量複合詞規則、詞綴構詞，以及專有名詞(地名、機構等)的結構化資訊，來建立相對的客語構詞規則，以改進客語 TTS 文句分析的效能。另外，我們也進行以大量的中文語言模型，利用語言模型調適方式(Language Model Adaptation)，調整成客語能夠使用的語言模型，以提升客語語音辨認的效能。

三、研究方法

本年度計畫執行可分為三大工作項目，包括(1) Text Analysis 的改進，(2)語音合成系統的改良，以及(3)語音辨認模組的改進，分述如下：

1. Text Analysis 的改進

(1) 客語詞典的擴充

文字處理不論在語音辨認或語音合成上都佔有重要角色，對於語音辨認方面，好的詞典搭配語言模型能夠增加詞彙辨認率，而在語音合成上，斷詞結果將影響整體語音韻律的

流暢度。

我們希望將目前國語分類詞典及文字分析技巧擴充至客語文字分析中。目前發展中的國語階層式語言模型將詞彙分成一般詞、詞綴詞、定量複合詞、功能詞及其他特定詞類詞、人名、專有名詞等類別(表一)。不同類別各自訓練不同的統計式語言模型，並分析各類別的局部結構。針對詞綴詞部分，我們將國語詞綴詞部分做拆解，定義詞幹(stem)與詞綴(affix)部分，建立詞綴表(表二)，並依據國國語-客語對照詞典找出詞綴詞的交互對應，經由人工校正及拼音修正，並加入客語特有詞綴後建立出客語詞綴詞對應表(表三)。定量複合詞部分，針對量詞類別、數詞結構、前後接詞類別及定量複合詞整體結構做分類，建立國語-客語定量複合詞對應表，加入客語特有量詞，並修正數詞、量詞等發音(表四)。功能詞及其他特定詞類部分，則針對特定詞類如介詞、連接詞等建立國語-客語對照表(表五)。人名部分目前共有華人、日本、歐美三類人名，華人部分又有古、今、暱稱等三類。我們先針對華人姓氏作分析，決定姓氏的客語發音，以減少一字詞發音混淆度。另外對於預估人名所需要的頭銜、稱謂兩類詞彙，則建立國語-客語對應表。未來將針對其他類別的專名做處理，處理方式為先將專名分類後擴充至客語詞典中，並將專名分成可拆解與不可拆解兩類，並針對可拆解部分建立詞組機率模型。

表一：國語分類詞典統計

	前/後詞綴詞	量詞	人名/人名相關	一般專有名詞	功能詞及其他
類別數	13/9	14	5/2	62	22
詞數	2960/14577	362	412814 / 787	5122	871

表二：國語、客語詞綴

	國語詞綴	客語詞綴
前詞綴	正、副、前、代、大、小、老、最、超、總、主	阿、𡃉、細、歸、逐
後詞綴	人、隊、會、者、性、率、室、科、制、法	仔、嫲、牯、公、个、娘、這兜、等、片

表三：國語-客語加詞綴後的詞對應範例

國語詞綴詞	淡褐色	烤香腸	長工	血腥味
客語對應	甜粄色	焙煙腸	承勞仔	臭血腥
客語詞綴詞	𡃉叔	兄弟仔	半路仔	刺泡仔
國語對應	小叔	好兄弟	中途	野草莓

表四：國語-客語量詞對應表

量詞種類	國語量詞	客語對應
數量單位(後可接受詞)	束、把、片	撮、抱、幷
金錢單位	元、分錢	銀、點錢

表五：國語-客語功能詞及其他特定詞類對應範例

詞類	國語詞	對應客語詞
Dfa	很	已、真、異、閑、當
DE	的	个、介
Cbb	而、但	毋過、總係
Da	才、只	正、單淨、單單

(詞典來源：台北市客委會、行政院客委會中級詞彙、台大客家社客語詞典、教育部台灣客家語常用詞詞典、網路資料等)

(2) 客語 Text Analysis 模組的改進

目前的客語 Text Analysis 仍以詞典斷詞加入斷詞規則為主，對於具有特殊結構的詞種，如定量複合詞、詞綴詞等並沒有特別處理，以至於斷詞結果不盡理想。除了詞典的廣泛收納外，對於有規則的詞種，我們希望以現有的國語構詞規則，建立國語-客語對應後，產生新的客語構詞規則。目前國語定量複合詞構詞規則已接近完整，除了將大部分規則直接套入客語構詞模組外，針對客語特殊構詞規則，我們也特別將國語-客語規則對應建表，以使客語構詞更精確完整(表六)。

表六：國語-客語定量複合詞構詞規則對應範例

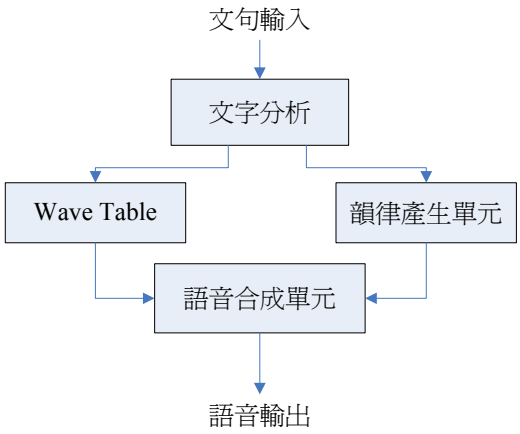
數詞種類	國語規則範例	客語對應
一開頭(數字不含零)	一百(千、萬)+數字	百(千、萬)+數字
數字含零	一百零五	一百空五

2. 語音合成系統的改良

(1) 四縣客語文句轉語音基本雛型系統

我們已建立一個基本的四縣客語文句轉語音(Hakka Text-to-Speech, HTTS)系統[5,6]，其基本模組如圖一所示，它係採用我們過去發展國語文句轉語音子系統相同的架構，系統包含四個模組：文句分析(Text analysis)、以遞迴式類神經網路(Recurrent neural network, RNN)為架構之韻律產生單元(Prosody generation)、基本波形單元(Wave table)、以 PSOLA 為基礎之語音合成單元(Speech synthesis)，其中 RNN 韻律產生器雖然可以使用簡單之語言參數，便可產生流利的韻律參數，但由於 RNN 內部之參數及架構難以分析，當產生不適當之韻律參數時，很難改善其表現，另外，語音合成單元是以 PSOLA 使用韻律信息來及調整基

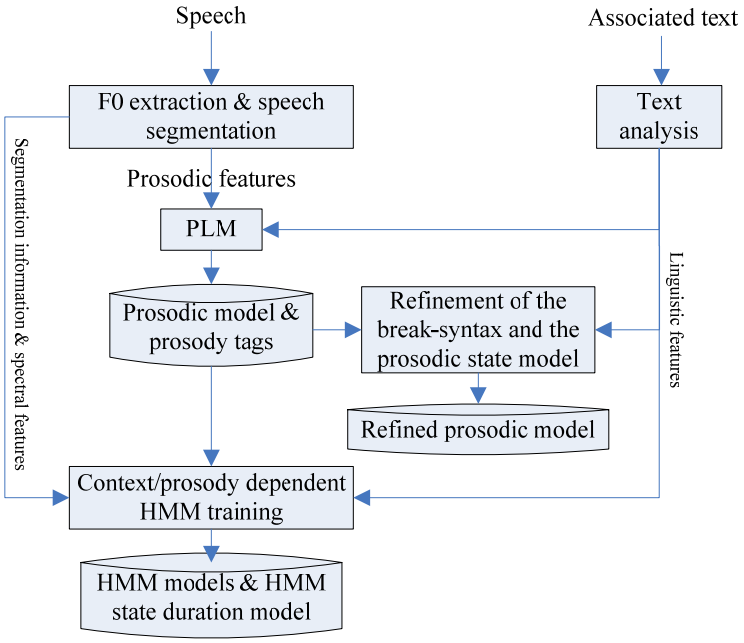
本音節波形，以串接的方式產生合成語音，此方法合成之語音較不自然，尤其是在目標韻律參數和基本音節波形的韻律參數差異較大時，容易造成音節內音素成份比例不適當以及雜音等問題。



圖一：客語文字轉語音雛型系統之基本架構圖

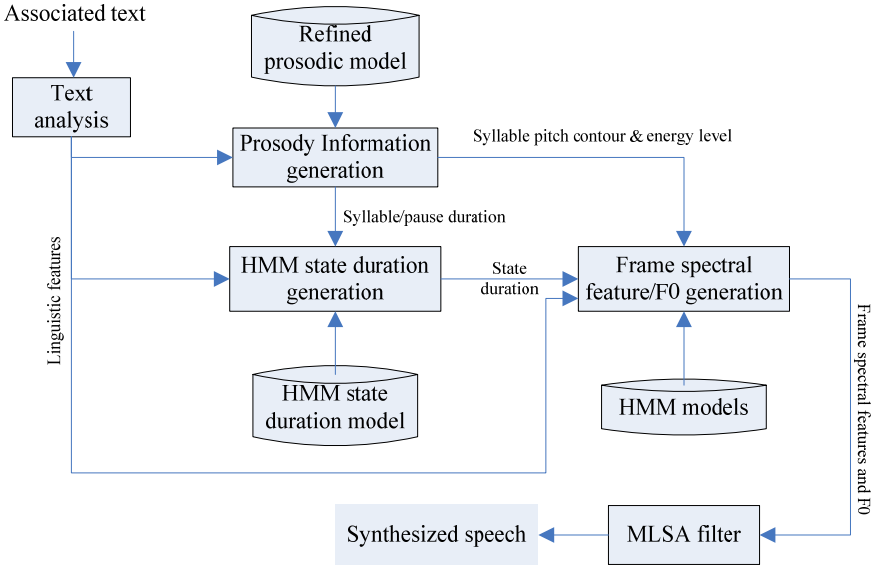
(2) 韻律模組及合成器之改良

為了改善以上缺點，韻律產生器將要改以我們最近提出之非監督式韻律標記及模式(Unsupervised joint Prosody Labeling and Modeling method, PLM)[7]為基礎之統計式韻律模型，而語音合成器將改用以隱藏式馬可夫模型(HMM)為基礎之韻律產生器[8]，其中 PLM 為基礎之韻律模型(prosodic model)可以描述韻律參數、語言參數以及韻律階層的關係，以 PLM 標記出之韻律標記(prosody tags)可以表示韻律單位的邊界(break)以及上層韻律單元的變化(prosodic state)，以此方法建構韻律模型，各韻律變化的影響因素可以清楚地被分析；另外，以 HMM 為基礎之語音合成器可以平滑地描述語音 spectral 上的變化，同時也可模擬前後音素以及 break 的影響。圖二為韻律模型及韻律產生器之訓練流程，首先，語音經由 F0 的偵測與切割至聲母及韻母之後，可以得到訓練韻律模型所需之韻律參數，由 text analysis 依據語音對應的文章內容得到語言參數，接下來使用 PLM 方法訓練出該語料語音之韻律模型，且同時對此語料庫標記 prosody tags，由於 PLM 方法是同時利用韻律參數和語言參數來標記 prosody tags，所以不需使用細緻的語言參數，然而在 TTS 的應用上，prosody tags 的產生只能由語言參數預估，因此 prosodic model 必須經過 refinement 的過程，使語言參數預估 prosody tags 的能力更強健。最後，我們利用決策樹方法，建構 context/prosody dependent 的 HMM 聲學模型以及 HMM state duration 模型，其中決策樹所使用的問題集，除了傳統的前後 phonetic structure 相關的語言參數外，也使用韻律參數相關的參數，如音節前後之 break type 以及音節所在的 prosodic state 等。



圖二：韻律模型及韻律產生器之訓練

圖三為客語語音合成系統的方塊圖，其工作方法如下所述。輸入的文字首先經由 text analysis 得到對應之語言參數，再以訓練好的 refined prosodic model 由語言參數產生韻律參數(或是 prosody tags)，這些韻律參數包含音節間的靜音時長、音節時長、音節音高軌跡以及音節能量 level，接下來利用 HMM state duration 模型在給定音節時長條件下，由語言參數和韻律參數預估音節內 HMM 各 state 的時長，然後在給定 state duration 和音節音高軌跡下，以訓練好之 HMM model 產生 frame spectral feature 以及 F0，最後以 MLSA filter[]產生合成語音。



圖三：客語語音合成系統方塊圖

(3) 語音合成語料之處理

在上年度，針對四縣腔，我們請一位發音純正的龔萬灶老師錄製，使用的文本為龔老師所出刊的客語散文集「阿鷙箭介故鄉」。錄音文章數為 42 篇，共 72064 音節，目前已錄製完成，所錄製的音檔均已自動切割，正著手以人工方式修正至更為正確切割位置。由於語音合成系統對於語音切割位置精確度的要求極高，這項工作必須以極細緻且正確的方式進行。另外，在文字檔部分，其對應發音均已由龔老師本人修正完畢，文章皆已自動斷詞且標記上詞類，需進一步檢查其對錯。

3. 語音辨認模組的改進

(1) 錄製客語語料部分

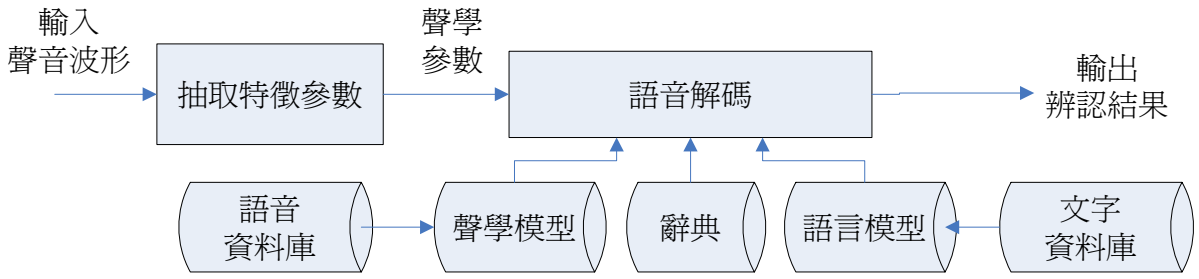
在增加客語語料的部分，目前新增了約 7 萬個音節，使得目前客語語料庫的總共音節數量達到 16 萬個音節左右(表七)。至於海陸腔客語的錄製方面，目前錄製的進展為 27965 個音節，包含了 2 男 4 女。

表七：目前收錄到的總音節數及語者數目於四縣腔與海陸腔

	四縣腔	**海陸腔**
總共的音節	160742	27965
語者數目	97 人(41 男，56 女)	6 人(2 男，4 女)

(2) Baseline Speech Recognizer

客語辨認器的改善部分，基本的客語語音辨認系統如圖四所示，本計畫基於此架構之上來改善聲學模型與語言模型，我們利用 HTK toolkit[9]建立聲學模型，使用的語料庫是由 97 個語者(含 41 男性及 56 女性)所錄製，總共有 160742 個音節，然後再將語料庫的十分之九當作訓練語料，語料庫的十分之一當作測試語料。



圖四：客語語音轉文字雛型系統之基本架構圖

本計畫使用了兩種辨認單元，第一種是使用和國語語音辨認器相同的次音節 HMM 模型，包括 87 個右相關聲母模型和 71 個韻母模型(簡稱 RCD)；其中右相關聲母模型狀態數目為 3 而韻母模型狀態數目為 5，用來與第二種辨認單元作比較，其辨認結果於表八所示；

另一種辨認單元為 phone HMM 模型，這些 phone set 大部分是與中文的 phone 共用的；總共 38 個 phones，其中每個 phone HMM 的狀態數目為 3。

語言模型的部分，由於文字資料仍不足訓練語言模型，因此我們使用 syllable tri-gram 當作 baseline LM，其文字資料來自於訓練部分，而這個 baseline LM 的參數是使用 HTK toolkit 所估計出來的。

表八：四縣腔 ASR Syllable Accuracy Rates

Method	**Syllable Accuracy (%)**
RCD model	52.93
RCD + syllable bigram LM	71.24
RCD + syllable trigram LM	75.70
phone model	55.55
phone + syllable bigram LM	62.34
phone + syllable trigram LM	75.26
triphones model (tri)	64.55
triphones + syllable bigram LM	77.07
triphones + syllable trigram LM	79.65

表九：模型大小

模型	**狀態數量**	**每個狀態的高斯數量**	**總共高斯數量**
RCD	716	16	11456
phone	114	100	11400
triphones	702	16	11232

(3) 未來目標：調適以解決訓練語料不足的困境

聲學模型部分：

因為客語語音是相對少數的語音於中文，因此使用中文聲學模型當作種子模型，調適出較 robust 的客語聲學模型，因為目前客語語料庫仍是偏少，所以藉由中文聲學模型來調適客語聲學模型會比較 robust。

語言模型部分：

已經處理 20102 個客語詞條可以對應到國語詞條，未來目標是完成國語辨認詞典的六萬詞可以有相對應的客語詞條，另一方面是建立客語的階層式詞典，將定量複合詞、詞綴、專有名詞、人名等有特殊結構的詞，抽取其結構，並以 Finite State Machine 等方式建立統計式的機率模型，並在語音解碼步驟採取兩階段的方式，將詞彙的結構資訊加入語音辨認的流程。。除了嘗試國語語言模型調適成客語語言模型外，也同時持續收集客語文章，也同時必須處理客語文章上的一個大問題，就是客語的文字不並像國語有統一的字詞，時常出現造字的問題。

四、結果與討論

本計畫正進行客語 TTS 及辨認系統的改進，除由基本的詞典擴增、新語料錄製上著手外，亦使用中文語音合成及辨認技術及資源，對 TTS 的文句分析及辨認的語言模型進行改進，在本年度結束時將可完成初步的系統改進，開發可用的四縣客語 TTS 系統及四縣客語辨認系統。

成果發表之論文：

[1] Chiang, Chen-Yu, Yu, Hsiu-Min, Wang, Yih-Ru and Chen, Sin-Horng, "Exploration of High-level Prosodic Patterns for Continuous Mandarin Speech," to be presented at *ICASSP* 2008, Las Vegas, Nevada, USA, March 30-April 4, 2008.

[2] Chen-Yu Chiang, Sin-Horng Chen, Hsiu-Min and Yu, Yih-Ru Wang, "Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech," J. Acoust. Soc. Am. 125, No. 2, pp. 1164-1183, 2009.

文獻探討

[1] S. H. Chen, S. H. Hwang, and Y. R. Wang, "An RNN-based Prosodic Information Synthesizer for Mandarin Text-to-Speech", IEEE Trans. Speech and Audio Processing, Vol.6, No.3, pp.226-239, May 1998.

[2] Min Chu, Hu Peng, Yong Zhao, Zhengyu Niu and Eric Wang, "Microsoft Mulan – A Bilingual TTS System," ICASSP 2003,

[3] C. H. Lee, H. Li, L. S. Lee, R. H. Wang and Q. Huo, *Advances in Chinese Spoken Language Processing*, World Scientific Publishing Co., 2006

[4] R. Sproat, Multilingual Text-to-Speech Synthesis: The Bell Labs Approach, Kluwer Academic Publishers, 1998.

[5] Hsi-Chun Hsiao, Hsiu-Min Yu, Yih-Ru Wang and Sin-Horng Chen, "Multilingual Speech Corpora for TTS System Development", Int. Symp. on Chinese Spoken Language Processing, Dec. 2006, Singapore; and Lecture Note in Computer Science, Vol. 4274/2006, Chinese Spoken Language Processing, Springer, pp.748-759 (SCI)

[6] Hsiu-Min Yu, Hsin-Te Hwang, Dong-Yi Lin and Sin-Horng Chen, "A Hakka Text-to-Speech System", Int. Symp. on Chinese Spoken Language Processing, Dec. 2006, Singapore; and Lecture Note in Computer Science, Vol. 4274/2006, Chinese Spoken Language Processing, Springer, pp.241-247 (SCI)

[7] Chen-Yu Chiang, Sin-Horng Chen, Hsiu-Min and Yu, Yih-Ru Wang, "Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech," J. Acoust. Soc. Am. 125, No. 2, pp. 1164-1183, 2009.

[8] T. Yoshimura, Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems, Ph.D thesis, Nagoya Institute of Technology, Jan. 2002.

[9] Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P. C., 2006. The HTK Book, version 3.4. Cambridge University Engineering Department, Cambridge, UK.