

客語文句轉語音及語音辨認之研究(3/3)

A Study on Corpus-based Text-to-Speech and Speech Recognition for Hakka Language

計畫類別：☒ 個別型計畫 ☐ 整合型計畫

計畫編號：NSC-95-2221-E-009-057-MY3

執行期間：96年8月1日至99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：江振宇、楊智合、蕭希群、黃信德、蔡依玲、蔡財祿

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共五份

☐ 國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、
列管計畫及下列情形者外，得立即公開查詢

☐ 涉及專利或其他智慧財產權，☐ 一年 ☐ 二年後可公開查詢

執行單位：國立交通大學電信工程學系

中 華 民 國 99 年 10 月 31 日

中文摘要

本計畫進行客語語音處理之研究，包含三個研究主題，一為建構高音質客語文句轉語音系統，二為建構實用的客語及國客夾雜語音辨認系統，三為客語音韻學研究。以四縣腔客語為研究主要對象，研究內容包含：(1) 基本資料的收集及建構，如詞典、文字語料、語音語料等；(2) 客語構詞規則及文句分析；(3) 客語特殊詞之音韻研究；(4) 客語文字轉語音系統發展；(5) 語言模式之建立；(6) 客語/國客語夾雜語音辨認系統發展。

關鍵詞：客語文句轉語音、客語語音辨認、四縣客語、韻律模式、語言模式

Abstract

In this project, we studied three speech processing topics for Hakka language: (1) the development of a high-quality Hakka text-to-speech (HTTS) system, (2) the implementation of a Hakka speech recognizer, and (3) the study in prosodic rules of special Hakka words. The study was conducted mainly on the major sub-dialect – “Si-xian Hakka” spoken in Taiwan. Several important issues were investigated, including (1) The collection and processing of lexicon, text databases and speech databases; (2) The realization of a high-performance text analyzer; (3) Prosody modeling of special Hakka words; (4) The implementation of the HTTS system; (5) Language modeling; and (6) The realization of the Hakka/mixed Hakka-Mandarin speech recognizer.

Keywords: Hakka text-to-speech system, Hakka speech recognition, Si-xian Hakka, Prosody, Language modeling

目錄

中文摘要.....	I
Abstract	I
目錄.....	II
一、緒論	1
1.1 研究背景及目的.....	1
1.2 先前研究	1
1.3 研究方向	5
1.4 具體成果概述.....	6
二、客語音韻之介紹.....	9
2.1 客語的音節結構.....	9
2.2 客語的聲母	10
2.3 四縣客家話之韻母.....	10
2.4 四縣客家話之聲調.....	11
2.5 四縣客家話之音節.....	12
三、客語文字轉語音系統.....	13
3.1 斷詞系統概述.....	14
3.2 中文分詞與詞性標記.....	15
3.3 客語分詞與詞性標記.....	18
3.4 HMM-based 語音合成器.....	21
3.5 實驗結果	29
四、客語及國客雙語語音辨認系統之建立	38
4.1 實驗語料介紹.....	38
4.2 客語與國語之間的語文特性	42
4.3 聲學模型	44
4.4 客語語言模型.....	51
4.5 雙語語言模型.....	57
4.6 實驗結果及討論.....	59
五、音韻學研究 - 客語「啊」時貌之間的關係探討	68
六、結論與未來發展.....	69
6.1 結論	69
6.2 未來展望	69
七、參考文獻	70
附錄一.....	73
附錄二.....	81
附錄三.....	82
附錄四.....	83
附錄五.....	85

一、緒論

1.1 研究背景及目的

文字轉語音系統(Text-to-speech system, TTS)及語音辨認的研究已有很長的歷史[14,15]，也都有很好的進展，近年來TTS 已進展到使用大型speech corpus 來選取適合的語音片段直接連接而不做韻律調整，因而獲得相當好的音質，研究重點在於prosody modeling[23-25]以及如何選取最適合的音節片段做連接，以中文而言，台大[16]、交大[26,29]、中研院、北京微軟[19-21]、中國電子科大、Infotalk 等都已具有Mandarin corpus-based TTS system。在語音辨認方面，已進展到實用階段，許多應用相繼推出，而研究方向則相當廣泛，如speaker adaptation、noisy speech recognition、spontaneous speech recognition 等；在中文ASR 的研究上，國內外有許多單位如台大、交大、清大、成大、中山、北科大、北京微軟、北京清華、中國科學院等在進行，也都有很好的進展。

然而，有關客語語音處理的研究則相當少，過去國內語言學家進行了一些客語相關的研究[1-13]，但在TTS 及語音辨認方面的研究則很少[27]，因此研究所需的相關的資料也不多，包括Lexicon、annotated text、speech corpora 都很零散。因此，為了增加研究的深度及廣度，在這三年期的計畫，除了建立文字轉語音系統及語音辨認系統之外，在蒐集及統整語音資料庫、文字資料庫、建立文字分析上付出相當的努力。另外，在客語的音韻學研究上，也有初步的成果。希望本研究建立的文字轉語音系統及語音辨認系統能應用於建立客語語言學習系統方面，輔助客語語言教學，除此之外，建立客語語音辨認系統的經驗，也可應用於建立其他常用漢語方言的語音辨識系統，如閩南語、不同腔調之客語等等，因此，本計畫對於方言之語音辨認技術、文字轉語音技術及自然語言處理是有其必要性及重要貢獻。

1.2 先前研究

我們在過去受客委會贊助組成一個跨語音處理(陳信宏)及客語語言處理(余秀敏、羅烈師)的團隊，共同發展了一套四縣客語文句轉語音(Hakka Text-to-Speech, HTTS)系統，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)、韻律信息產生(Prosody generation)、合成單元波形表(Acoustic inventory)、語音合成(Speech synthesis)，文句分析是將輸入文句做斷詞分析，以產生詞串、詞類串、音節串；韻律信息產生是由文句分析的結果抽取一些語言參數，來產生適當的韻律信息，用以合成自然流利的語音；聲音波形資料庫儲存所有的基本音節波形訊號，它使用文句分析的結果，來組合出對應輸入文句音節串的基本音節波形串；語音合成是使用韻律信息來調整基本音節波形串的韻律，使其成為自然流利的語音。各模組之工作說明如下：

1. 文句分析

文句分析為HTTS 系統重要的一個部份，文句分析結果的優劣直接影響合成聲音的

品質和正確性，所以在做文句分析時，必須使用正確的資訊和方法，以提供後級好的各項語言資訊。以下對(a) 詞典的收集狀況、(b)文句分析程序、(c)文句分析模組輸入與輸出之格式的三個部分加以說明：

(a) 詞典的收集狀況

詞典是文句分析器最重要的資料之一，以詞典的各項資訊可決定斷詞結果的輸出。我們的詞典共含36,258 個詞條，主要來源為台北市客委會以及台大客家社所提供的客語電子辭典，此外，也從《客語實用手冊》(龔萬灶著)、《初級客語認證詞庫》(行政院客委會)、以及陳碧娥老師撰寫之客語文稿等資料中，補充辭典所缺列的詞彙。

(b) 客語文句分析器處理程序

客語文句分析器主要分成三個步驟來決定斷詞結果，分別為前處理步驟、長詞優先斷詞、斷詞結果加入詞類標記及音碼標示等步驟，作法說明如下：

- 針對文章的前處理：客語文章如前所述，其特性為漢羅夾雜，在分析時將造成困難，必須先將羅馬拼音部份轉換成為程式可接受的內碼格式，才能作正確斷詞。前處理能夠將部份拼音連寫的現象作區隔，譬如拼音文字間以“-”作區隔，或者拼音和漢字連寫在一起時，此程序可以正確將拼音部份單獨提取出來，以達到正確斷詞。
- 長詞優先斷詞：以查詢詞典的方式，配合長詞優先原則做斷詞的依據。以詞典查詢方式將輸入文句分析成為多種可能的斷詞結果，再套入長詞優先原則，以決定最後的斷詞結果。長詞優先原則係依據中央研究院詞庫小組所定義的長詞優先標準，以連續三詞長度總和及詞長變異數最小的分詞結果為優先考慮。
- 將斷詞結果標上音碼及詞類標記：由詞典中各詞所擁有的音節碼以及詞類資訊，做斷詞結果的拼音及詞類標示。其中拼音部份只選用唯一的拼音，詞類標記的原則是以使用頻率最高的詞類作為詞類標示結果的依據。此客語文句分析器雖已加入一些詞綴相關的構詞規則，但尚不完整。

(c) 文句輸入與輸出格式

客語文章的特性為漢羅夾雜，尤其是當文章用字遣詞方式越接近口語、文章內容越屬於鄉土文學，漢羅混用的情形就越明顯，例如『當時介竹筴，就像這下介hiap4 錢箱樣仔』就是一個混用漢字和羅馬拼音的句子。以客語文句分析器分析以上句子，可得以下輸出資訊：

- 斷詞結果：將原本未斷詞文句分析成為以詞為單位的詞串；
- 詞長：由斷詞結果詞串計算而得的各詞長度；
- syllable 在詞中的位置
- 音節碼：各詞所擁有的客語音節碼(目前客語音節碼共有671 類)；
- 詞類標記。

以下為輸入文句『當時介竹筴，就像這下介hiap4 錢箱樣仔』得到的斷詞結果輸出形式：

例句一：當時(dong1 sii5, Nd) 介(ge2, De) 竹筴(zuk4 gong5, Na) ， 就像(ciu3 ciong3, P) 這下(it4 ha3, Nd) 介(ge3, De) hiap4 錢(cien5, Na) 箱(siong1, Na) 樣仔(iong3 e2, Na)

其斷詞效能之評估如表1.1其效能仍有進步的空間，所以在本計畫執行的三年，我們對斷詞系統重新設計，利用國語和客語之間相近的語言特性進行改善，在第三章有所詳述。

表1.1：斷詞效能評估

N1(hand-segmented word number)	25306
N2(TA-segmented word number)	26039
N3(correct TA-segmented word number)	20910
Recall (N3/N1)	82.63%
Precision (N3/N2)	80.3%

2. 韻律信息產生

韻律信息產生是要由語言參數來產生包含音節基週軌跡、音節能量軌跡、音節長度(或聲韻母長度)、及音節間停頓長度等韻律信息，本系統使用的方法是遞迴式類神經網路(Recurrent Neural Network, RNN) [18] (見圖一)，它由一個大的語料庫來學習語言參數到韻律參數的對應關係，訓練時藉調整類神經網路的神經元間連結的加權(weights)，在合成時，將語言參數輸入來產生韻律參數。

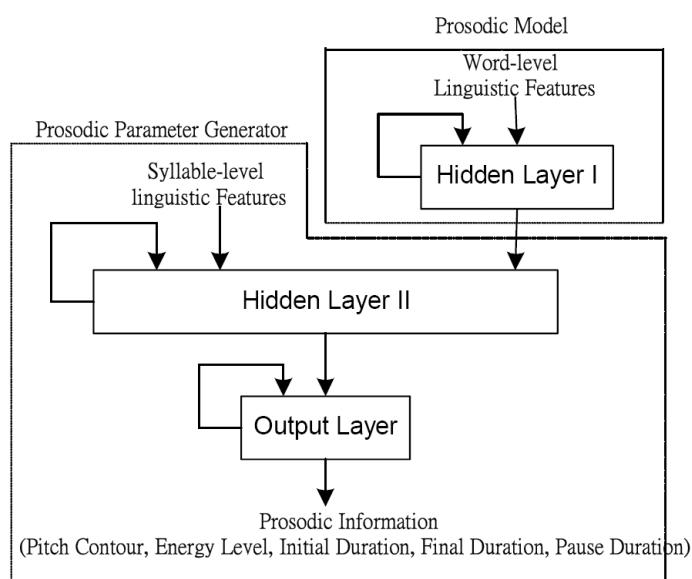


圖1.1：遞迴式類神經網路韻律信息產生器架構圖

在過去我們錄製了一份女性朗讀語料庫，共有57 篇文章(故事、諺語)，包含42,337 個音節、12,293 個詞(不包含一字詞)及325 個語音檔，求取音檔的基週軌跡參數、音長、能量，以提供RNN韻律訊息產生器訓練用。RNN 之輸入語言參數包含兩類，一類是音節層次(syllable-level) 的參數包括前後音節的tones、現在音節的17 類聲母及47 類韻母(客語韻母共有72 類、以及現在音節在詞中的四種位置資訊(單音節詞、多音節詞的詞首、詞中及詞尾)等。另一類是詞層次(word-level) 的參數包括前後詞的詞類(Part of speech, POS)、詞長、詞前後是否有標點符號等。在完成訓練後，此韻律信息產生器之效果尚佳，產生之韻律信息可用以合成流利自然的語音，表1.2種韻律參數之均方根誤差，但仍有少許的較大誤差，在本計畫三年中，我們將RNN韻律產生器，改以HMM模擬，內容詳述於第三章。

表1.2 RMSEs of the five synthesized prosody parameters.

	Inside Test	Outside Test
F0 Contour	1.9ms	2.2ms
Pause Duration	56.8ms	65.4ms
Initial Duration	20.7ms	25.6ms
Final Duration	42.9ms	45.7ms
Energy Level	3.7dB	4.3dB

3. 聲音波型資料庫

由於客語和漢語一樣是以音節為基本的發音單元，音節的數目並不多，我們因此使用不帶聲調的671 個基本音節做為TTS 系統的基本合成單元，標示組成之聲母及韻母資訊，由一位女性語者以正常的速度讀671 個基本音節，再以人工處理先將基本音節波形切出，然後作聲韻母邊界切割以及標示pitchmarks，使用的工具是我們過去發展的軟體。

4. 語音合成器

語音合成採用被普遍使用的基週同步疊加(Pitch Synchronous Overlap-Add, PSOLA)演算法，它使用RNN產生之韻律參數來調整基本的音節波形的基週軌跡、能量位準、聲韻母長度及音節間停頓長度，其產生之合成聲音品質尚稱良好。在本計畫中，我們將引進HMM-based 的語音合成器，改善原本以PSOLA為基礎之語音合成器，內容詳述於第三章。

1.3 研究方向

本計畫擬進行客語語音處理之研究，包含三個研究主題：(1)高音質HTTS 系統之建立，(2)客語及國客雙語語音辨認系統之建立。(3)客語音韻學研究 以下為本計畫三大內容之概述：

(1)高音質HTTS之建立

改進先前研究建構之PSOLA-based HTTS 系統，完成一套客語文句轉語音系統。它由斷詞系統、停頓預估系統、文脈分析器以及基於隱藏式馬可夫模型之語音合成器所組成。文句輸入經由斷詞系統做分詞和詞性標記。由於客語語料的不足，本研究在此結合中文斷詞條件隨機域模型、客語外掛詞典、中文內部詞典及客語構詞規則形成客語斷詞系統。斷詞系統輸出斷詞資訊後，由靜音停頓預估系統預估詞後是否停頓，再由文脈分析器輸出和語音合成器相對應之合成單元及語言參數，依據文字標記檔及合成器訓練端所得到的模型進行音長、聲學、音高參數的預測，再得到合成語音輸出。最後我們也分別設計實驗對斷詞系統、靜音停頓預估系統、合成器以及HTTS系統做出客觀效能評估，HTTS系統在主觀評分上也得到了不錯的分數，顯示出我們已經有一套不錯的客語文字轉語音系統。以下為建立高音質HTTS之重點研究工作項目：

- (a) 單一語者大型speech corpus 之錄製及處理。
- (b) 客語詞典之擴增。
- (c) Text Analysis之改進。
- (d) 停頓預估。
- (e) HMM-based 語音合成器之建立。

(2)客語及國客雙語語音辨認系統之建立

根據既有國語語音辨認系統輔助建立客語語音辨認系統，並且由單語言的語音辨認系統，以不同定義的方式設計系統內的不同元件，建立雙語語音辨認系統。其中包括客語語言模型的建立、雙語語料分析、雙語音素集的設計、針對雙語特性設計語言模型、以及雙語系統等。由實驗證實，雙語語音辨認與客語語音辨認都朝向預期目標發展。其中客語語音辨認系統對於加入之國語語言參數資訊，確實都能影響其辨認效能，不過由於客語訓練語料數量的限制，客語語音辨認率明顯比國語語音辨認系統低，但是藉由分析及實驗結果，得到許多對於客語語音辨認系統的有效知識及經驗，相信對於日後的發展將有所助益。以下建立客語及國客雙語語音辨認系統之重點研究工作項目：

- (a) 訓練聲學模型(Acoustic Model, AM) 之文字語料收集及處理。
- (b) 多語者(約100人)大型speech corpus 之錄製及處理。
- (c) 訓練語言模型(Language Model, LM) 之文字語料收集及處理。
- (d) 國語三連音、客語雙連音、國客語混和三連音聲學模型訓練
- (e) LM 之建立。

(3)音韻學研究 – 客語「啊」時貌之間的關係探討

本研究的主要目的是解釋當客語的「啊」作為時貌標記時，其核心語意為表淺程度、短時之意，從這個語意出發經過核心語的隱喻轉換、和語意延伸，可貫穿「啊」所扮演的四個客語時貌標記（暫時貌、嘗試貌、反覆貌、以及緊密先前貌）之間的語意關連性。並將這些不同時貌對應之語音聲調表現關係建立起來，外來可用於語音合成之韻律產生規則使用。

1.4 具體成果概述

本計畫具體之成果可分為三大部分：(1)資料庫收集、(2)離型系統、以及(3)論文發表，分述如下：

(1) 資料庫收集

甲、 文字資料庫

i. 四縣客語詞典

1. 辨認用途(DictA)：

43,521 個客語詞條，詞典內容包含詞、四縣音標、音節碼、詞性、中文翻譯以及詞來源。其中部分客語詞條(26,524 個)有國語翻譯的對應，用於語音辨認的辨認詞典。

2. 合成用途(DictB)：

34,541 個客語詞條，詞典內容包含詞、四縣音標、音節碼、詞性、中文翻譯以及詞來源。用於斷詞器中的客語外掛詞典。

ii. 四縣客語文章(textA)：

此文章來源為網路上擷取而來，主要來源為全國語文競賽朗讀文章，文章內容大部份為故事散文集，共 218,034 字。經過文字正規化及客語詞典斷詞處理，總詞數有 143,341 個。用於建立客語 LM。

iii. 國客語平行文字語料(textB)：

此語料來源為龔萬灶老師撰寫的四縣客語文集，並且經由余秀敏老師人工標記詞類及國語翻譯資訊而成，其文章內容為客語故事。使用的詞類標記共有 49 類，共有 42 篇文章、63158 個字(包含標點符號)、36450 詞。用於語音合成系統及建立語音辨認之 LM。

乙、 語音資料庫

i. 語音合成用途

1. 陳老師四縣客語語料：

20kHz 的取樣頻率及 16-bit 之單聲道 pcm 格式錄音，文章來源主要由龔萬灶老師所主筆的「阿啾箭个故鄉」，包含 57 篇文章(故事、諺語)，包含 42,337 個音節、及 325 個語音檔，有人工斷詞修正、詞類標記、正確語音切割。

2. 龔老師四縣客語語料：

20kHz 的取樣頻率及 16-bit 之單聲道 pcm 格式錄音，文章來源由龔萬灶老師所主筆的「阿啾箭个故鄉」，包含 42 篇文章、63,158 個音節、639 個語音檔，有人工斷詞修正、詞類標記、對應中文翻譯、正確語音切割。

3. 羅老師海陸客語語料

20kHz 的取樣頻率及 16-bit 之單聲道 pcm 格式錄音，是為龔老師四縣客語語料的海陸腔平行版本，也包含 42 篇文章、639 個語音檔，錄製完成，但尚未做語音切割和發音標注。

ii. 語音辨認用途

1. 四縣腔客語語料：

屬於麥克風朗讀語音(16kHz 的取樣頻率及 16-bit 之單聲道 pcm 格式錄音)。文章內容主要為客語故事，主要由苗栗的兩位退休教師，一位是龔萬灶老師、另一位為陳碧娥老師所撰寫；並加入中華大學余秀敏老師所新增的客語文句。由 92 人(男 43 人、女 49 人)朗讀錄製而成，共 10,229 個檔案、153,911 個音節。

2. 國客夾雜語料

屬於麥克風朗讀語音(16kHz 的取樣頻率及 16-bit 之單聲道 pcm 格式錄音)。文句由中華大學余秀敏老師根據日常國客語夾雜使用情況所撰寫，文句內容分為客語為主文句及國語為主文句，各 100 句。由 20 人(男女各 10 人)錄製而成，每人 20 句國語為主文句、20 句客語為主文句，共 800 個檔案(國語為主、客語為主各 400 個)，國語為主之語料包含 4,956 個音節，而客語為主之語料包含 4,960 音節。

(2) 雛型系統

- 甲、 中文 Text Analysis
- 乙、 四縣客語 Text Analysis
- 丙、 四縣客語文字轉語音系統
- 丁、 四縣客語語音辨認系統
- 戊、 國客語混合語音辨認系統

(3) 論文發表

- 甲、 客語文字轉語音系統方面

Yi-Ling Tsai, Hsiu-Min Yu, Yih-Ru Wang, Chen-Yu Chiang, Lieh-Shih Lo, and Sin-Horng Chen, “An HMM-based Hakka Text-to-Speech System,” accepted by OCOCOSDA 2010.

- 乙、 客語、國客混合語音辨認系統方面

Tsai-Lu Tsai, Chen-Yu Chiang, Hsiu-Min Yu, Lieh-Shih Lo, Yih-Ru Wang and Sin-Horng Chen, “A Study on Hakka and Mixed Hakka-Mandarin Speech Recognition,” accepted by ISCSLP 2010.

- 丙、 客語音韻學研究方面

余秀敏,“客語啊時貌之間的關係探討”,第八屆台灣語言及其教學國際研討會
論文集, pp. 216~227.

丁、 碩士論文

- i. 蔡依玲,“基於隱藏式馬可夫模型之客語文句轉語音系統”, 國立交通大學電信研究所碩士論文, 民國 99 年
- ii. 蔡財祿,“國客雙語語音辨認”, 國立交通大學電信研究所碩士論文, 民國 99 年

(本篇碩士論文榮獲第十屆中華明民國計算機語言學學會碩博士論文獎佳作, 詳情請參閱http://www.aclclp.org.tw/grants_c.php)

二、客語音韻之介紹

客語又稱客家話或客話，是漢語方言學者及聲韻學家們公認的漢語七大方言群中的一支重要方言系統[3]，以客語為母語的人口約佔中國漢族總人口之百分之四，估計有四千五百萬人左右[3]，這些客語族群主要分佈在廣東東北、江西南部、福建西部及北部、廣西南部、湖南、四川、東南亞等的海外華僑以及台灣等地[3]。楊福錦1967《客方言的音韻成素》，依據聲、韻、調的語音特徵差異，廣義的將客語族群分為七大次方言區：梅縣區、興寧區、饒平區、海陸區、香港區、汀州區、以及永定區[3]，也許是廣東梅縣是純客語的最大縣城，加上提倡客語和客家族群研究的最早先驅羅香林先生也是梅縣人，學術界以及客家人本身一向以梅縣的次方言為客語的主要代表。至於台灣方面，客家族群分佈在台灣北、中、南及花東地區的縣市城鎮，除了較少人口的詔安、大埔、饒平、和永定等次方言外，最主要的台灣客語次方言分別是四縣話（梅縣區的次方言）和海陸話（海陸區的次方言）[3]，基本上仍以四縣話為主，而苗栗縣是台灣各客語縣市中，說四縣客語人數最多的一個縣市[1]，因此，以苗栗客語為研究對象，對於客語語音技術的研發來說，十分具有代表性。以下以余秀敏1994《苗栗客家話音韻研究》[5]、羅肇錦1990《台灣的客家話》[3]以及鍾榮富2004《台灣客家語音導論》[2]所提供的資訊為主，並以微幅修正的通用拼音為主要標音法，分別介紹苗栗客語的音節結構、聲母、韻母、及聲調等的語音特徵。

2.1 客語的音節結構

如同其他的漢語方言，客語也是聲調語言：一個音節搭配一個具有辨別語意功能的聲調，其聲韻調的音節結構如下圖所示[2]：

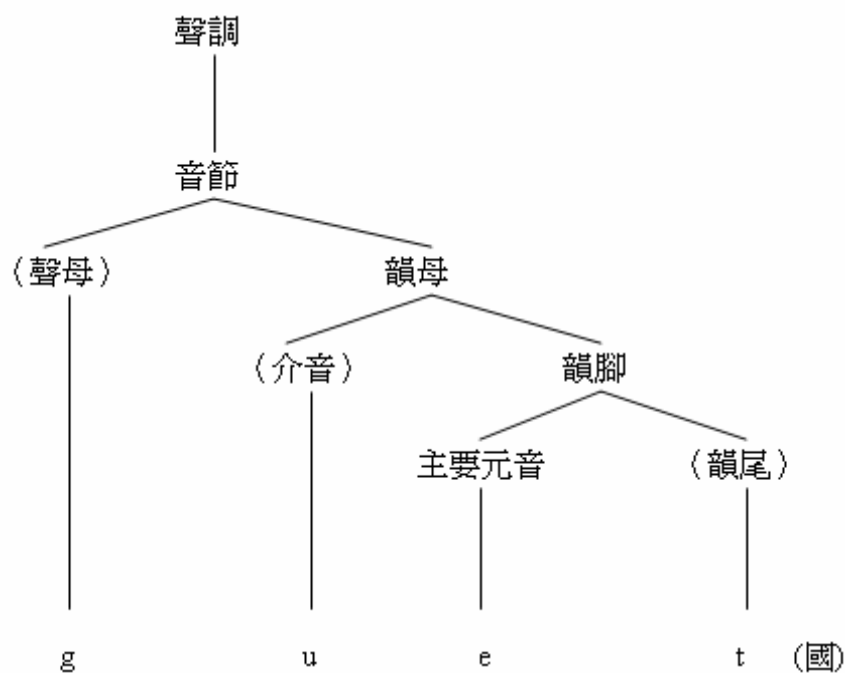


圖2.1：客語音節結構

由圖2.1 顯示的結構來看，最上層的兩個音節要素可以分成「聲母」與「韻母」，而「韻母」又可再細分為「介音」與「韻腳」，最後「韻腳」的部分則再次分為「主要元音」和「韻尾」，這些結構要素中凡是加上刮號的因子都不是構成音節的必要條件，因此，可以得知，一個音節不可或缺的唯一要素是「聲調」以及「主要元音」。在這個結構下，可填入6 個四縣聲調、17 個聲母、2 個介音、5個主要元音、以及8 個韻尾。除去聲調的考量，截至目前為止，共收錄了671 種不同的四縣音節。

2.2 客語的聲母

四縣客家話的聲母是位於音節開頭的輔音，空聲母是指音節開頭沒有輔音，四縣客家話共有 17 個聲母，其中空聲母也包含在內；四縣客家話的聲母只有 m、n、ng、l、v 為發音時聲帶有振動具有週期性的濁音，其餘聲母都屬於清音。我們依照發音部位以及發音方式用表 2.1 將四縣客家話的聲母分類，其中括號內為例字。

表 2.1: 四縣客家話聲母表

發音方式 發音部位	塞音		塞擦音		鼻音	邊音	擦音	
	清音		清音		濁音	濁音	清音	濁音
	不 送 氣	送 氣	不 送 氣	送 氣				
雙唇	b (飛)	p (片)			m (慢)			
唇齒							f (飯)	v (位)
舌尖	d (知)	t (梯)			n (泥)	l (來)		
舌尖前			z (正)	c (鄭)			s (師)	
舌根	g (高)	k (扣)			ng (咬)			
喉							h (後)	

聲母發音的變化中值得注意的是，舌尖前音z、c、s（同國語ㄗ、ㄘ、ㄙ的發音），若出現在高前原音i 的前面時，則基層音（basic phones）會顎化（palatalized）成像國語的ㄗ、ㄘ、ㄙ的變體音（allophones）[2]。

2.3 四縣客家話之韻母

韻母是音節中位於聲母後面的部分，韻母可以再細分三部分，韻頭、韻腹以及韻尾，其中韻腹是必須要有的，韻頭和韻尾則可有可無。韻頭就是介音，有 i 或 u；韻腹為主要元音，四縣客家話的主要元音有 a、e、i、o、u、ii 共 6 個；韻尾分為元音韻尾和輔音韻尾，元音韻尾包含 i、u 等，輔音韻尾因聲調的不同又分為兩種，一種是以鼻音節

尾的舒聲韻，有 m、n、ng，入聲韻則是由塞音節尾，共有 p、t、k。四縣客家話的韻母共有 71 個，如表 2.2

表 2.2: 四縣客家話韻母表

代碼	韻母	代碼	韻母	代碼	韻母
1	ii	25	em	49	iep
2	i	26	iem	50	ep
3	e	27	am	51	ap
4	ie	28	iam	52	iap
5	ue	29	iin	53	iit
6	a	30	in	54	it
7	ia	31	en	55	et
8	ua	32	ien	56	iet
9	o	33	uen	57	uet
10	io	34	an	58	at
11	u	35	uan	59	uat
12	iu	36	on	60	ot
13	iui	37	ion	61	ut
14	ai	38	un	62	iut
15	uai	39	iun	63	ak
16	oi	40	ang	64	iak
17	ioi	41	iang	65	ok
18	ui	42	ong	66	iok
19	eu	43	iong	67	uk
20	au	44	ung	68	iuk
21	ieu	45	iung	69	FNULL
22	iau	46	uang	70	iing
23	iim	47	iip	71	uak
24	im	48	ip		

2.4 四縣客家話之聲調

漢語當中國語的聲調有四個，閩南語有七個，四縣客家話的聲調和其他次方言相較之下較為簡單，共有六個，其調類、調號和例字如表 2.3 所示：

表 2.3: 四縣客家話聲調表

調類	陰平	陰上	陰去	陰入	陽平	陽入
調號	1	2	3	4	5	8
例字	千	錢	淺	賤	絕	切

四縣客家話和中文相同，也有變調的情形，變調規則會跨詞出現，但不會出現在句子與句子之間。其中變調規則共有三種：

1. 由兩個陰平(調號 1)字構成的字彙，讀時前字變調讀陽平(調號 5)
陰平(調號 1) + 陰平(調號 1) → 陽平(調號 5) + 陰平(調號 1)
例：「新衫」 sin1 sam1 → sin5 sam1
「買新衫」 mail sin1 sam1 → mai5 sin5 sam1
2. 陰平(調號 1)字與陰去(調號 3)字構成的詞彙，讀時前字變調讀陽平(調號 5)
陰平(調號 1) + 陰去(調號 3) → 陽平(調號 5) + 陰去(調號 3)
例：「針線」 ziim1 sien3 → ziim5 sien3
「拿針線」 na1 ziim1 sien3 → na5 ziim5 sien3
3. 陰平(調號 1)字與陽入(調號 8)字構成的詞彙，讀時前字變調讀陽平(調號 5)
陰平(調號 1) + 陽入(調號 8) → 陽平(調號 5) + 陽入(調號 8)
例：「音樂」 im1 ngok8 → im5 ngok8
「聽音樂」 tang1 im1 ngok8 → tang5 im5 ngok8

2.5 四縣客家話之音節

客語文章中紀錄的單元大多是音節，一個中文字代表一個音節。依據聲韻組合規則，四縣客家話可以衍生出 13 種不同的類型，這也是四縣客家以及其它漢語方言所有音節的縮影，表 2-4 列出這 13 種音節結構，由於聲調是每個音節所必備，在下表音節類型裡，聲調用數字表示。而實際上四縣客家話的音韻系統共有 671 個音節，詳細列表如附錄一。

表 2.4: 四縣客家話音節類型表

音節類型	例字
元音	o2 (襖)、i5 (于)
元音化的鼻音輔音	m5 (毋)、ng5 (魚)
元音 + 輔音韻尾	it4 (一)、ap4 (鴨)
元音 + 元音韻尾	ai2 (矮)、oi3 (要)
介音 + 元音	iu3 (柚)、ia5 (爺)
介音 + 元音 + 輔音韻尾	iok8 (浴)、iuk4 (育)
介音 + 元音 + 元音韻尾	ieu5 (謠)、iau1 (飫)
輔音 + 元音	ho5 (何)、pa1 (爸)
輔音 + 元音 + 輔音韻尾	nap8 (納)、ziip4 (汁)
輔音 + 元音 + 元音韻尾	cai2 (採)、tui3 (退)
輔音 + 介音 + 元音	mia1 (摸)、dui3 (對)
輔音 + 介音 + 元音 + 輔音韻尾	giuk (趨)、liok8 (略)
輔音 + 介音 + 元音 + 元音韻尾	tiau5 (跳)、diau3 (釣)

三、客語文字轉語音系統

語音合成的技術主要有兩大類，一為基於語料庫之合成法(corpus-based)，近期的單元選取合成系統(unit selection synthesis system)所合成出的語音品質已相當穩定，但在輸出的語者特性或情緒上難以調適，並且須保存大量語料庫；第二種是基於數據處理之參數合成法(parametric speech synthesis)，不同於直接從語料庫挑選語音片段單元合成，在訓練完模型參數後，只需保留訓練得到的模型，在可攜性(Portable)及適應性(Flexibility)上有很大的進步。近年來，基於隱藏式馬可夫模型的語音合成系統(HMM-based Speech Synthesis System, HTS)[29,30]是基於 parametric speech synthesis 處理之合成法中最廣為應用的合成技術，其應用上的彈性大，能合成出較自然的語音。

本研究將使用基於隱藏式馬可夫模型語音合成系統，完成客語文字轉語音之系統實作，系統基本架構如圖 3.1 所示：

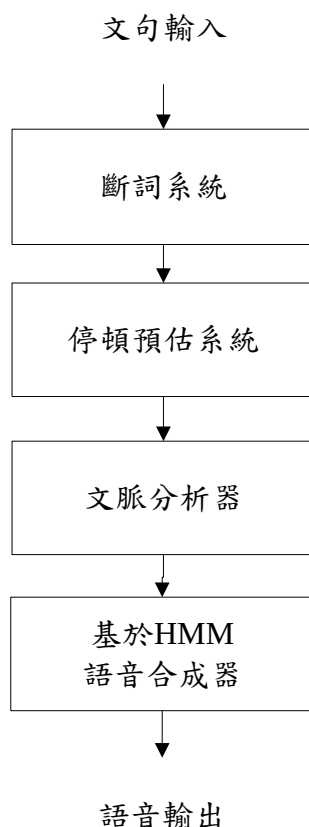


圖 3.1: 客語文字轉語音系統

本系統包含斷詞系統(Text Analysis)、停頓預估模型、文脈分析器以及基於 HMM 之語音合成器；文句輸入經由斷詞系統做分詞(Segmentation)及詞性標記(Part of Speech Tagging)，接著由停頓預估模型預估詞後是否有停頓，文脈分析器將前一階段的輸出轉成合成單元序列及其對應語言參數，接著 HTS 找到相對應之頻譜、音長和音高的 HMM

模型，再還原成語音訊號輸出。

3.1 斷詞系統概述

斷詞是文字轉語音系統中第一步驟，此步驟的輸入是一個完整的句子，輸出是已標記好的詞彙序列以及其相對應的詞性標籤。斷詞的好壞會直接影響合成聲音的品質與正確性，是文字轉語音當中很重要的單元。其系統架構如圖 3.2：

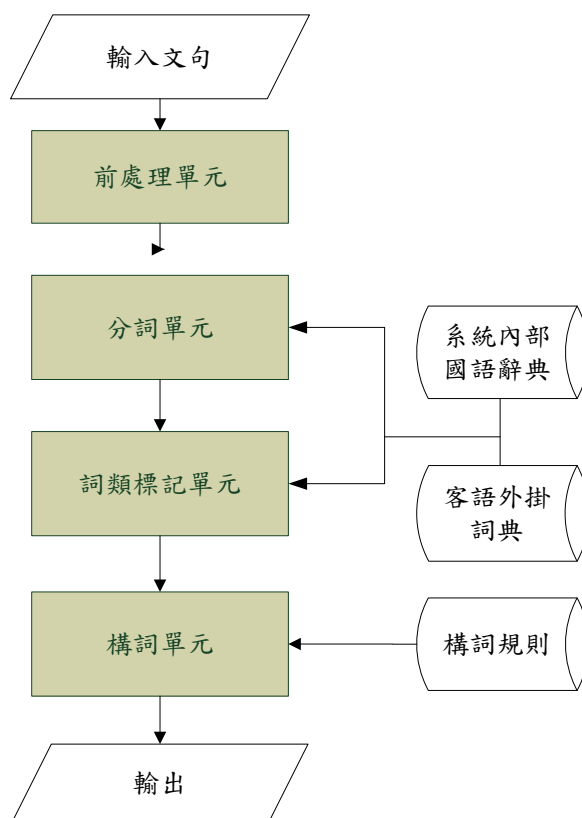


圖 3.2: 斷詞系統

其中各個單元之功能在以下做簡單介紹：

(1) 文字正規化單元

輸入的文句當中可能包含 ASCII code 或 Big-5 code，在進入斷詞單元之前，我們將所有的 ASCII code 轉成 BIG5 code 格式，使字串格式統一。例如「西元 1945 年」轉為「西元 1 9 4 5 年」；而中文的用字歧異和標點符號的統一也都在此一單元中解決，例如「裏」和「晒」轉為「裡」和「曬」；另外，標點符號中如逗號「，」皆統一轉成「，」；此外，客語文章和台語文章一樣，有許多造字情況，目前解決方式是利用 Big-5 code 中的保留區域對應其造字。

(2) 分詞單元

此單元目的是將輸入的文句做出適當分詞，也是整個系統當中最核心的部分。分詞單元是最前面的分詞階段，後來的「詞性標記單元」會利用此單元所產生的候選句標記詞性。

(3) 詞類標記單元

詞類標記單元利用機率統計模型，標記所有可能的詞性組合，並從候選句當中選出最合乎語法的句子。

(4) 構詞單元

四縣客家話中如「結結」（結實的樣子）同字疊字型態，或是名詞後加詞綴的「細人仔」（小孩子）等，都是無法全部收錄至詞典當中，因此利用構詞規則合併詞組以提供良好的斷詞結果。

3.2 中文分詞與詞性標記

中文分詞與詞性標記第一步是透過詞典及中文分詞條件隨機域模型得到可能的分詞候選句；接著經由查詢詞典得到候選句中每個詞彙的詞性標籤組合，最後利用詞性標記條件隨機域模型，從候選句中選出最佳結果，而此最佳結果為已標上詞性標籤的詞彙序列；本節將對中文分詞以及詞性標記做介紹。

當輸入某句文字序列序列 C 時，我們要找到其最佳對應的詞彙序列 W ，以及詞性標記序列 T ，可用以下公式表示：

$$P(W, T|C) = P(T|W, C) \times P(W|C) \quad (3-1)$$

可分為 $P(T|W, C)$ 和 $P(W|C)$ 兩機率值的相乘， $P(W|C)$ 表示出現某個文字序列時，出現某個詞彙組合的機率，即出現某個分詞組合之機率，我們由條件隨機域模型得到最高機率前幾名作為分詞單元輸出候選句； $P(T|W(S, B))$ 是在某個詞彙序列的情況下，出現某個詞性標記序列的機率，候選句配合詞典中每個詞彙可能出現的詞性標籤輸入至條件隨機域模型，模型會依照學習到的句法結構，給予候選句（從第一個到最後一個）每一個詞彙一個最高機率數值，即最可能出現詞性標籤，以及其對應的機率； $P(T|W(S, B))$ 和 $P(B|S)$ 兩機率相乘得到候選句的分數，最後選取分數最高組合的做為最佳分詞與詞性標記結果。

3.2.1 中文條件隨機域模型

條件隨機域 (Conditional Random Field) 常用在標註和分析序列資料，例如自然語言文字序列，是一種鑑別式的機率模型，在 2001 年由 Lafferty[31] 等人提出。條件隨機域是一種無向性的圖模型，在給定需要標記的觀察序列條件下，計算整個標記序列的聯合機率分佈；而不是在給定當前狀態條件下，定義下一個狀態的狀態分佈；也就是給定觀察序列 X ，使 $P(Y|X)$ 值為最大時的 Y 即為最佳答案，如以下式子所表示：

$$Y^* = \arg \max_Y P(Y|X) \quad (3-2)$$

X 是本研究中代表中文字元序列 C ，例如序列 X 「感動的事」， C_1 是「感」、 C_2 是「動」、 C_3 是「的」而 C_4 為「事」； Y 則代表詞位標籤序列，詞位標籤將於下節做詳細介紹，本例所對應的正確詞位標籤為「B₁ESS」。而 $P(Y|X)$ 如下式所表示：

$$P(Y|X) = \frac{1}{N(X)} \exp \left(\sum_{t=1}^T \sum_{i=1}^I \lambda_i f_i(Y_t = y, Y_{t-1}, X) \right) \quad (3-3)$$

$N(X)$ 為正規化數值，為了保證所有標記總和值為一， $\sum_y P(Y|X) = 1$ ， $f_i(Y_t = y, Y_{t-1}, X)$

表示序列 X 在位置 t 時的第 I 個特徵函數，此函數對應到 0 或 1。

$$f_i(Y, Y_{t-1}, X) = \begin{cases} 1, & \text{if } y = y_t \text{ is satisfied and } x = x_t \\ 0, & \text{otherwise} \end{cases} \quad (3-4)$$

例如，序列 X ：「感動的事」， Y 序列(標準答案)為「B₁ESS」($y_1 = B_1$ 、 $y_2 = B_2$ 、 $y_3 = S$ 、 $y_4 = S$)，當 $t=2$ 時 $f_i(Y_2 = E, Y_1 = B_1, X_2 = \text{動}) = 1$ 。而 $\lambda_1, \lambda_2, \dots, \lambda_k$ 為條件隨機域從訓練語料中學習的權重向量。

3.2.2 中文分詞

由於每個字在組成一個特定詞語時都佔據一個構詞位置，即字位，因此可以將分詞過程當作學習此字之字位訊息的機器學習過程。分詞是對字串中每一個字相對應做出一個在該處切分與否的決策過程；本研究採用 Zhao[32] 在 2006 年所提出的六字位標注集。六字位標注集之集合為 $T = \{B_1, B_2, B_3, M, E, S\}$ ，其中 B_1 、 B_2 和 B_3 分別表示該字位於一個詞的前三字位置， M 則代表該字位於詞中第四個字元之後但非詞尾的位置， E 表示該字位於一個詞的詞尾， S 代表單字詞。

我們使用線上開放的套裝軟體 CRF++[33] 來訓練中文斷詞模型，此軟體被廣泛應用在許多自然語言處理領域的研究。在訓練階段時，CRF++ 必須透過特徵樣版產生特徵參

數，特徵樣版表示我們擷取特徵的規則，本研究在訓練中文斷詞條件隨機域模型所選用的特徵如表 3.1 所表示。其中字元 C_n 、標點符號 PM_n 、數字字元 D_n 、英文字元 E_n 、和詞典詞相對應的 WL_n 及 WT_n 幾項特徵為 Zhao[32]所提出；而本研究加入每個字元的部首資訊 B_n 以及 DP_n 和 DT_n 兩項未收錄詞典詞之特徵， DP_n 及 DT_n 是關於在同一個句子中是否找到重複出現的詞，可以補足詞典中未收錄之詞；例如在文章中常出現但詞典並未收錄的專有名詞或是長詞，這兩項特徵可以提高此部分的分詞效能。

在訂定適當的特徵後，我們將每個特徵依據其相關性及其前後資訊分組，稱為特徵樣版，我們所採用的特徵樣版詳細列在表 3.2。

表 3.1: 中文分詞之條件隨機域訓練特徵

特徵	含意
C_n	當前字元
B_n	當前字元的部首
PM_n	當前字元是否為標點符號
WL_n	在詞典可找到包含當前字元的最長詞之詞長
WT_n	在詞典可找到包含當前字元的最長詞中，當前字元的詞位標注(B_1, B_2, B_3, M, E, S)
DP_n	在同一句子當中，可以找到包含當前字元的最長詞之詞長，而此詞在包含當前字元的詞出現前曾經出現過
DT_n	若當前字元可找到其包含當前字元的最長詞，而此詞在包含當前字元的詞出現前曾經出現過，當前字元在此詞當中的詞位標注
D_n	當前字元是否為數字
E_n	當前字元是否為英文字母

表 3.2 中文分詞之條件隨機域訓練特徵樣版

特徵樣版	含意
$C_{n-1}, C_n, C_{n+1}, (C_{n-1} C_n), (C_n C_{n+1}), (C_{n-1} C_{n+1}), PM_n$	字元與其臨近字元資訊
$(B_n B_{n-1}), (B_{n+1} B_n), (B_{n+2} B_{n+1})$	部首資訊
$(WL_n WT_n), (WT_{n-1} WT_n WT_{n+1}), (WL_n WT_n DL_n DT_n)$	詞典詞及重複詞資訊
$(C_n D_{n-1} D_{n+1})$	數字字元資訊
$(E_{n-1} E_n), (E_n E_{n+1}), (E_{n-1} E_n E_{n+1})$	英文字元資訊

3.2.3 中文詞性標記

詞性(Part of Speech, POS)可以代表一個詞在句子中的語法角色，而詞性標記單元的功能，就是在斷詞單元產生 N 個最佳候選句後(本系統中 $N=3$)，給予每個詞彙相對應的

所有詞性組合，再由條件隨機域模型算出有最高機率的詞彙組合及其詞性標籤。本研究在訓練中文詞性條件隨機域模型所使用的特徵及特徵樣版如表 3.3 及表 3.4 所表示：

表 3.3 中文詞性標記之條件隨機域訓練特徵樣版

特徵	含意
PM_n	標點符號類別 0：沒有標點符號 1：「、」 3：「。」、「；」、「：」、「？」、「！」、「—」 2：其他標點符號
WL_n	當前詞長
$LPOS_n$	若詞典有該詞，則此詞收錄在詞典中的所有詞性組合 例如：「一」的 $LPOS_n$ 為「Cbb_Di_D_Neu」
FC_n	若該詞未收錄於詞典，則該詞得第一個字元 例如：「約克夏」的 FC_n 為「約」
LC_n	若該詞未收錄於詞典，則該詞得最後一個字元 例如：「約克夏」的 FC_n 為「夏」

表 3.4 中文詞性標記之條件隨機域訓練特徵樣版

特徵樣版	含意
$PM_{n-1}PM_nPM_{n+1}$	標點符號資訊
$LPOS_{n-1}, LPOS_n, LPOS_{n+1}, (LPOS_{n-2}LPOS_{n-1}), (LPOS_{n-1}LPOS_n), (LPOS_nLPOS_{n+1}), (LPOS_{n+1}LPOS_{n+2})$	POS 資訊
WL_n, FC_n, LC_n	未收錄詞典詞資訊

3.3 客語分詞與詞性標記

由於我們在客語語料上的不足，無法以條件隨機域方法獨自訓練出客語專用的斷詞系統，因此本研究採用客語外掛詞典、國語詞典、客語構詞規則以及中文條件隨機域模型組合成客語分詞與詞性標記系統。

在分詞階段，文句輸入經由優先搜尋客語外掛詞典，若是在詞典搜尋到該詞則將此詞切分開，其餘文字經由中文條件隨機域模型及國語內部詞典挑選出分詞候選句，接著查詢外掛詞典及內部詞典得到每個詞的詞性標記組合，最後由中文條件隨機域模型得到

最佳斷詞與詞性標記序列。

3.3.1 客語詞典

詞典在 TTS 系統中有非常重要的功能，在分詞和標記詞性階段能提供詞和其詞性資訊，在合成語音前的文字分析器當中必須查詢詞典得到對應的音標和聲調。我們主要的收集來源為「完整版中級詞彙彙整檔案」、「台北市客委會-現代客語詞彙彙編」、「教育部客家語常用詞辭典」等；詞典內容包含詞、四縣音標、音節碼、詞性、中文翻譯以及詞來源。目前所收錄的客語詞之數量如表 3.5 所示：

表 3.5 客語詞典統計表

詞長	1	2	3	4	5	6	7	8	total
詞條數	6747	18078	5095	4217	250	80	60	14	34541

客語外掛詞典是客語分詞與詞性標記系統中很關鍵的資訊，我們先將客語詞典中與中文內部詞典相同的共用詞彙刪除，其所有外掛詞典之資料量在表 3.6 詳細列出。

表 3.6 客語外掛詞典統計表

詞長	1	2	3	4	5	6	7	8	total
詞條數	91	5296	5521	3401	293	75	80	12	14769

3.3.2 構詞規則

分詞會將輸入的文句和詞典做比對，我們無法將所有可能的詞條列於詞典當中，這些無法收錄於詞典之詞，有些是有規律的，可由「詞類標記單元」得到詞性標籤後再經由「構詞單元」結合出來。

客家話的構詞規則和國語非常相似，大致上分為四種，分別是重複，附加，附合及合併。所謂重複是指單一詞彙之重複，如「洗湯」重複成「洗洗湯湯」（表廚房內擦擦洗洗之事）。附加則是在詞的前後附加綴詞，如在「婆」之前加前綴詞「阿」形成「阿婆」一詞，後綴詞如在「刀」之後加「仔」形成「刀仔」。附合詞是指兩個獨立的詞結合而成另一個新詞，例如「青菜」是由「青」和「菜」組成，但「青菜」不一定是指「青色的菜」，而是所有蔬菜的泛稱，附合詞兩個詞之間不能插加任何詞語，由此可見附合詞並不單純是兩個詞的併列，而是有其密不可分的構詞特性。第四種構詞方式稱為合併，這是中文構詞規則所沒有的，如 cin1（親）gal（家）念快時會變成 cial（親家），這種合併結果是單音節多詞素的構詞。我們解決附合詞的方式為新增詞條至詞典，而合併詞在四縣客家話當中並不常見，因此本系統所採用的構詞規則主要以重複和附加為

主。

我們採用中央研究院詞庫小組所制訂的「中文分詞處理原則」，由其處理原則中挑選出符合客語文章之構詞規則加入本系統，詳細構詞規則如表 3.7、表 3.8 所示。表 3.7 列出了客語和國語完全相同的構詞規則，表 3.8 則為客語和國語平行的構詞規則，其構詞結構一樣，但關鍵字由中文詞替換為客語詞。

表 3.7 國客語共用構詞規則

國語構詞規則	國語範例	客語範例
XXY(Z)	說說看	日日大
X-XY-Y	乾乾淨淨	洗洗湯湯
V+了/著/過/看/看看+(V)	想過	歇過
名詞後詞綴：氏/度/性/家/長/師/員/們/子/倆	女性朋友們	細妹伴仔
排行:數字+姐/妹/哥/弟/伯/叔/姑	二伯	二伯
VG+為/成/作/不了/得了	身為	身為

表 3.8 國客語平行構詞規則

國語構詞規則	國語範例	客語構詞規則	客語範例
V+不+到/起/出	看不起	V+毋+到/起/出	看毋起
越+來+越	越來越少	緊+來+緊	緊來緊少
越+ X+越+Y	越穿越多	緊+ X+緊+Y	緊著緊多
VB/VH+不休/兮兮/不絕/不下/不得/開來	蜿蜒不絕	V+毋得	捨毋得
老/小/阿+Na	小妹妹	細+Na	細妹仔

3.3.3 客語借字問題

在客語的語料紀錄當中，遇到有音無字的情形時，解決方法除了造字以外，經常使用借中文字表示客語字，例如「每擺」（每次）中，「擺」音同「次」的客語發音，因此客語文章中使用「擺」代表中文字「次」；同時，「擺脫」（擺脫）是客語文章中的國客共用詞，「擺」在此的意思又與中文原意相同。

借字情形出現在二字詞以上時，我們新增此類情況的詞彙至客語外掛詞典，大部分可正確進行分詞與標記詞性；但在一字詞的情況下，如前例中客語文章中的「擺」代表了中文字「擺」和「次」，中文字「擺」有可能出現的詞性組合為「VAC_VC」，中文字「次」的詞性組合則為「Nes_Nf」，因此「擺」在客語文章中所有可能的詞性組合為「Nes_Nf_VAC_VC」，但此種組合並未在訓練時加入中文詞性條件隨機域模型當中，因此由本研究的斷詞器無法標記出正確的斷詞及詞性結果，並且容易造成搶詞情形。

為了處理客語字之詞性組合和中文模型無法相對應的問題，我們從中文模型可允許的詞性組合中找出最相近的，上例中為「Na_Nf_VC」，並將其加入外掛詞典，如下例所表示：

客語字	語料庫中出現對應中文字/POS	對應中文字/POS	加入客語外部詞典 POS
擺	次/Nf	次/Nes_Nf	Na_Nf_VC
	擺/VC	擺/VAC_VC	

本實驗從客語語料中統計出現次數最多的 40 個一字詞，將有此問題之單元整理其對應詞性，並加入外掛詞典中，所加入之一字詞如附錄二所表示。

3.4 HMM-based 語音合成器

基於隱藏式馬可夫模型的語音合成系統(HMM-based Speech Synthesis System, HTS)[29,30]是目前最廣為應用的合成法，與傳統的串接式合成系統比較，基於模型的語音合成方法具有較高的可攜性以及適應性，儲存模型所需要的空間，遠小於串接式合成系統所需儲存大量的音檔；另一方面，模型的可訓練性，能快速使模型參數進行調式或轉換。本節將詳細介紹 HTS 工作原理、我們所採用的決策樹問題集、文本相關資訊，以及對本研究在語音合成前所做的停頓預估系統加以討論。

3.4.1 HMM-based 語音合成系統

基於 HMM 之語音合成系統架構如圖 3.3 所示，主要分為訓練以及合成兩個階段：

(一) 訓練階段：

我們將得到切割資訊後的聲音語料，估算其特徵向量作為參數，包括頻譜部分(spectrum,梅爾倒頻譜)以及其動態特徵向量、激發訊號部分(excitation,logF0) 以及其動態特徵向量；聲音語料所對應的文字則由斷詞系統和文本分析器分析出對應的文脈相關序列，使系統在訓練階段時可以透過合併分裂樹學習合成單元間的相連狀況；包含切割資訊的語音訊號以及文本相關序列先經由 HMM 訓練，產生各合成單元的音高特徵模型、頻譜特徵模型和音長特徵模型，再透過狀態合併決策樹問題集對模型進行分群，得到和文本相關的 HMM 模型。

(二) 合成階段：

使用者輸入文字，由斷詞和詞性標記系統得到斷詞資訊後，再經由停頓預估器預估出停頓位置（停頓預估系統將在 4.2 節詳細介紹），送入和訓練部分相同的文本分析器得到合成單元及其所對應的語言參數序列；透過分類及回歸樹挑選出相對應之 HMM 模型序列，根據參數產生演算法產生特徵參數，最後通過梅爾對數頻譜近似濾波器(Mel Log Spectrum Approximation filter, MLSA filter)合成語音訊號輸出。

基於 HMM 之語音合成器主要核心技術有三個；(1) 包含動態特性的參數生成演算法，此演算法將在下節做詳細介紹。(2) 多空間機率分佈隱藏式馬可夫模型(Multi-space Probability Distribution HMM, MSD-HMM)，基頻在濁音處是有值的，但在清音處為零，為了解決基頻在濁音處和清音處轉換上的不連續，MSD-HMM 在此定義濁音處的參數維

度是一，而清音處參數維度是零。(3)基於梅爾倒頻譜的聲音合成技術，其中包含訓練階段梅爾倒頻譜的抽取，以及合成階段 MLSA filter 將梅爾倒頻譜系數還原為語音訊號。

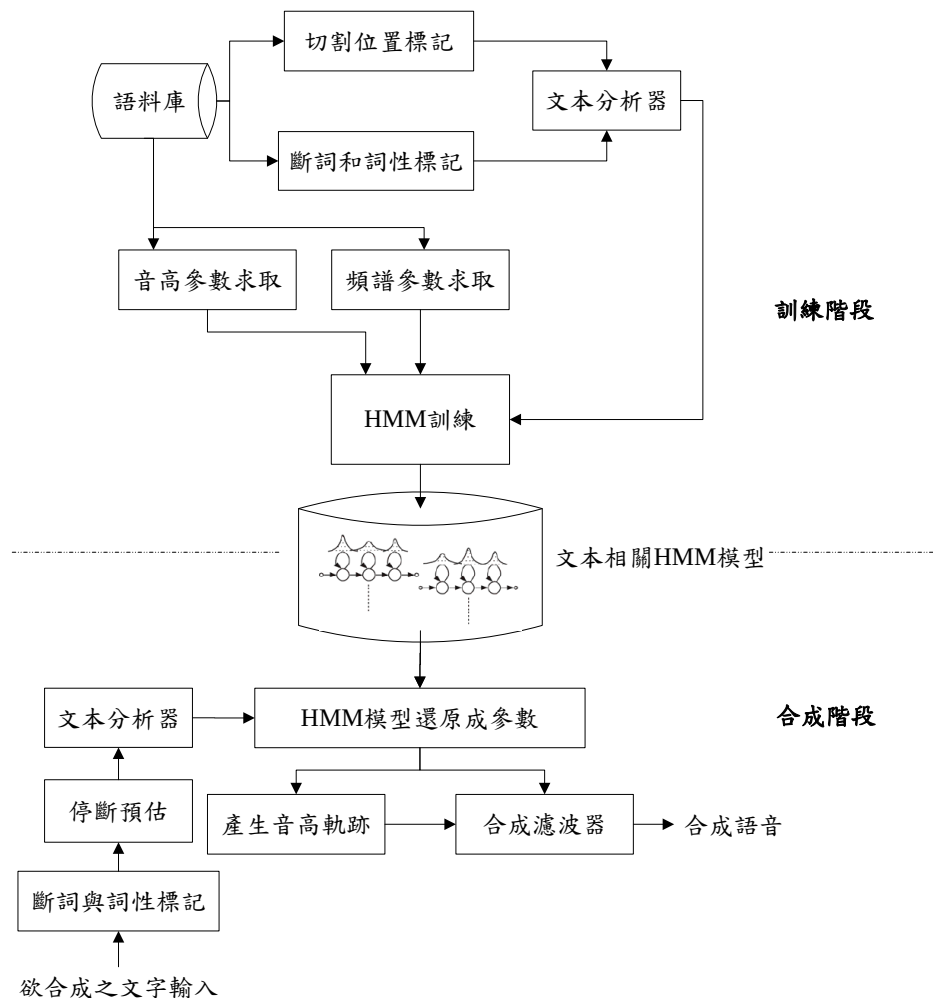


圖 3.3: 基於隱藏式馬可夫模型之語音合成系統架構

3.4.2 考慮動態特性的參數生成演算法

考慮動態特性的參數生成演算法是 HTS 能合成出流暢自然語音的關鍵，本節將對動態參數以及其參數生成演算法加以詳細介紹。

3.4.2.1 動態參數

系統所抽取出的梅爾倒頻譜數值以及 $\log F_0$ 數值稱為靜態參數(Static feature)，代表了音高走勢以及頻譜特徵的變化，而此種變化會受到前後合成單元的影響，也就是特徵參數的改變與前後時間的關聯性非常大，因此有學者提出動態參數(Dynamic feature)的觀念在語音辨識的 HMM 模型參數上，後來也在基於 HMM 的語音合成器中被使用。動態參數的考慮能讓訓練出的模型單元和單元間的連接更加流暢。基於 HMM 的語音合成器中的動態參數為計算靜態參數在時間軸上的前後關連性而得到，最後的觀察參數為原本

的靜態參數以及其動態參數，不僅擴大模型的參數維度，也讓模型在前後時間關係有更好的描述。

令 $c_t = [c(1), c(2), \dots, c(M)]^T$ 為在時間點 t 的靜態特徵向量（例如頻譜參數），而相

對於 c_t 的一階動態參數 Δc_t ，以及二階動態參數 $\Delta^2 c_t$ 計算式如下：

$$\Delta^n c_t = \sum_{i=-L^{(n)}}^{L^{(n)}} \omega^{(n)}(i) c_{t+i}, \quad n=0, 1, 2 \quad (3-5)$$

$\omega^{(n)}(i)$ 表示第 n 階動態參數對於時間點 $t+i$ 的靜態參數所附予的權重係數， i 表示作用範圍，也就是時間點 t 往前後各 $L^{(n)}$ 個時間單位。

$$\omega^n = \left[0, \dots, \omega_{-L^{(n)}}^n, \dots, \omega_0^n, \dots, \omega_{L^{(n)}}^n, \dots, 0 \right]^T \quad (3-6)$$

其中 $\omega^{(0)}(0)=1$ 以及 $L^{(0)}=0$ 。

動態參數表現在時間上的物理意義會因為各階動態參數所對應設計的 ω^n 而有所不同，本研究的動態參數計算如下：

$$\Delta c_t = \frac{1}{2} c_{t+1} - \frac{1}{2} c_{t-1} \quad (3-7)$$

$$\Delta^2 c_t = (c_{t-1} - c_t) + (c_{t+1} - c_t) \quad (3-8)$$

我們在圖 4-2 表示出一階及二階動態參數的關係：

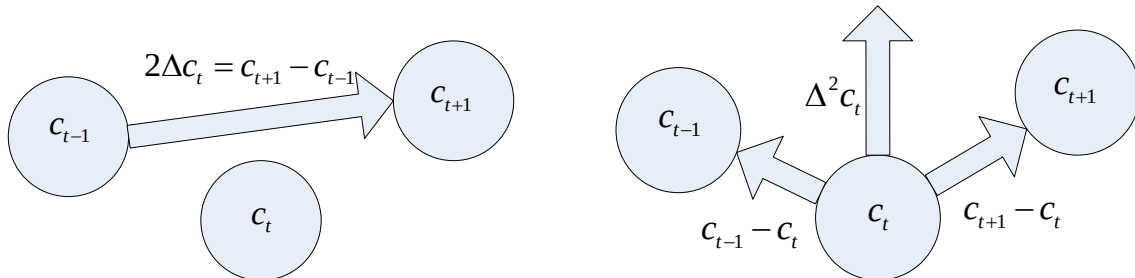


圖 3.4: 一階及二階動態參數關係圖

由圖可以看出，一階動態參數和自己本身的靜態參數沒有關係，是下一個單元以及

前一單元的差距；而二階動態參數則是由前一單元與自己的差距和後一單元與自己的差距相加。我們將靜態參數以及一階、二階動態參數作為模型的參數向量，則在時間點 t 的參數向量如下表示： $o_t = [c_t, \Delta c_t, \Delta^2 c_t]$ ，其中 $c_t = [c_t(1), c_t(2), \dots, c_t(M)]$ 。

而整個句子的參數向量是以 $O = [o_1', o_2', \dots, o_T']$ 表示， O 和 c 之間的對應關係矩陣 W 如圖 3.5 所示：

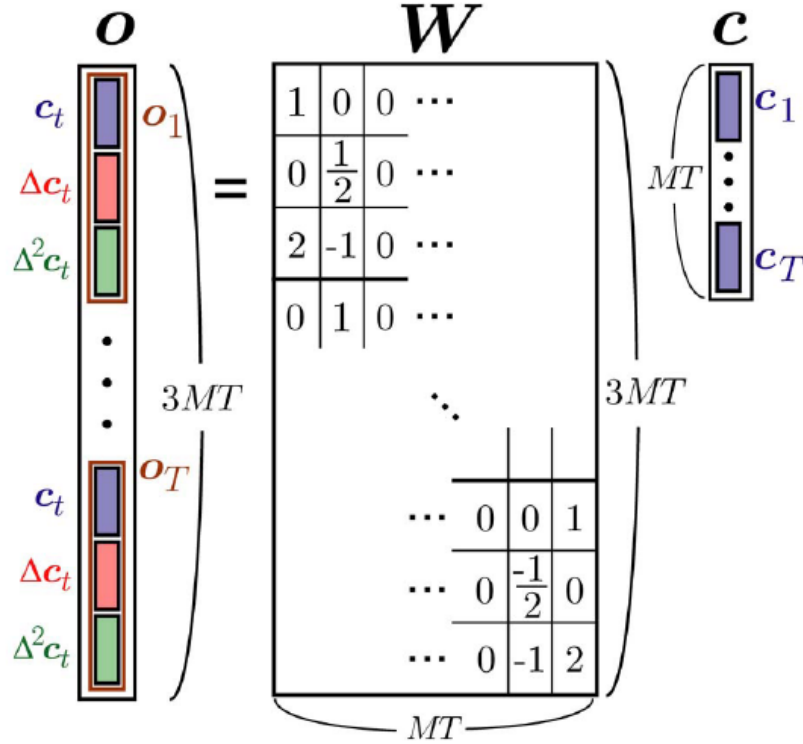


圖 3.5: 轉換矩陣表示示意圖。

W 矩陣的維度是 $3MT \times MT$ ，將原本的靜態參數向量 c 轉換到參數向量 O ，參數維度也從 $MT \times 1$ 擴大到 $3MT \times 1$ 。

3.4.2.2 參數生成演算法

在合成階段，文字經由和訓練階段相同的文本分析器後得，到相對應的前後文相關語言參數序列 $S_1, S_2, \dots, S_N$ ，經過 HMM 模型預測後得到模型序列 $\lambda_1, \lambda_2, \dots, \lambda_N$ ，以及模型參數向量 $O_1, O_2, \dots, O_N$ ， O_n 是由平均值向量 μ_n (Mean vector) 及共變異數向量 U_n (Covariance matrix) 所組合而成， μ_n 和 U_n 的維度包含靜態參數、一階動態參數以及二階動態參數的維度，是訓練階段所抽取出的參數維度之三倍。在合成語音前，必須將 O_n 還原成對應的 C_n ，才能得到頻譜或音高軌跡，此轉換方式稱為參數生成演算法，如圖

3.6 表示。

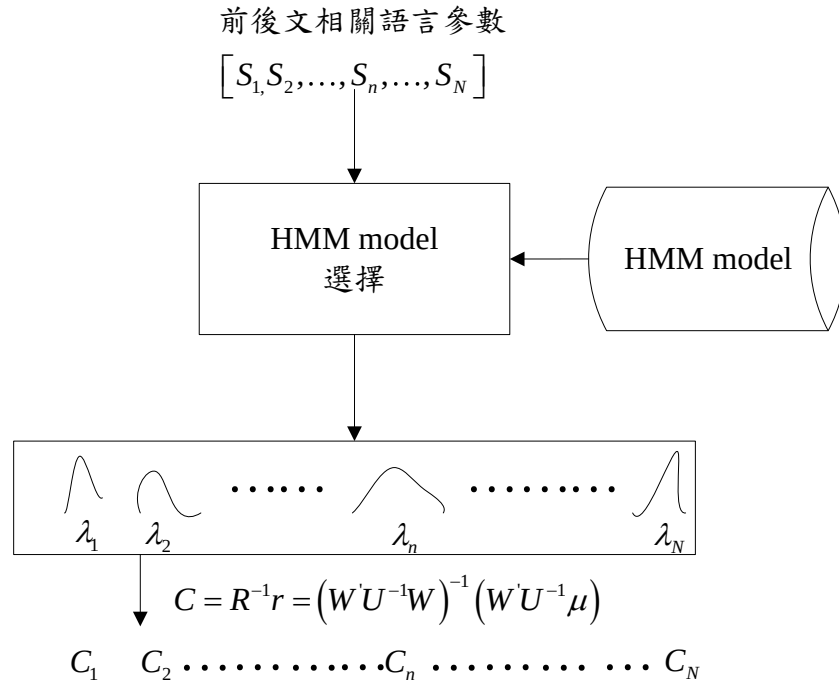


圖 3.6: 參數生成示意圖

合成階段將合成單元參數串接成 O ， $O = \{O_1, O_2, \dots, O_N\}$ ，接著產生整個輸入字串的 HMM 模型 $\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_N\}$ ，模型的平均值向量為 $\Lambda_\mu [\mu'_1, \mu'_2, \dots, \mu'_N]'$ ，共變異數矩陣 $\Lambda_U = \text{diag}[U_1, U_2, \dots, U_N]$ ，我們需要找出使得已知模型 Λ 產生 O 中機率最大的 C ，也就是找到最佳的 C 使 $P(O|\Lambda)$ 有最大機率值。其中 $O=WC$ ，由 $P(O|\Lambda)$ 取對數值得到：

$$\log P(O|\Lambda) = -\frac{1}{2}(O - \Lambda_U)' \Lambda_U^{-1} (O - \Lambda_U) - K \quad (3-9)$$

其中 $K = \frac{1}{2} \log |\Lambda_U| - \frac{3}{2} MN \log 2\pi$ ，和 O 不相關。

將式子對 C 偏微分，在偏微分為零的地方為極大值：

$$\frac{\partial \log P(WC|\Lambda)}{\partial C} = 0 \quad (3-10)$$

可以得到：

$$W'U^{-1}WC = W'U^{-1}\mu \quad (3-11)$$

最後我們得到： $C = R^{-1}r$

其中 $R = W'U^{-1}W$ 而 $r = W'U^{-1}\mu$ 。

3.4.3 切割位置

進行語音合成的訓練之前，必須先將語音標示出各合成單元的邊界資訊，才能正確抽取出各合成單元的頻譜、音高及音長之特徵值。本實驗所採用的合成單元為聲母與韻母。我們由 HTK (Hidden Markov Toolkit) 訓練出聲母及韻母的 HMM 模型，對所有語料進行強制切割 (Forced Alignment)。參數設定為 38 維，其中包含 12 階的梅爾倒頻譜參數 (Mel-frequency Cepstral Coefficients, MFCCs) 與能量對數值 (log energy)，及其一階微分與二階微分，扣除原本的能量對數值後共 38 維，音框大小 (frame size) 設為 32ms，音框位移 (frame rate) 設為 5ms。

由 HTK 得到初步切割位置後，我們使用本實驗室之高解析度自動語音切割系統，其架構是根據聲/韻母的層級來分類，並依照各個分類彼此之間可能的轉換狀態訂定目標函數，再使用類神經網路之多層感知器架構，以半監督式之模型訓練方法建立起端點偵測器模型，將 HMM 的自動切割結果以端點偵測器所產生的概似度經正規化後作為分數，進行維特比搜尋並限制搜尋範圍在初始切割位置前後 100 毫秒之內，最後得到本研究之聲/韻母層級自動語音切割位置。

3.4.4 文本相關資訊

在將語料及其對應文字送入訓練以及將文字輸入合成語音前，我們要將文字檔轉為合成單元及其相對應之文本相關資訊；文本相關資訊包含聲學模型之合成單元以及前後合成單元的特徵資訊，在 HTS 訓練階段時根據文本相關資訊和決策樹問題集進行比對分群，合成部分則由決策樹取得參數而合成語音。本研究所採用的合成單元為聲母/韻母/停頓 (short pause, sp) 資訊，表 3.9 為文本相關資訊和說明：

表 3.9: 文本相關資訊格式

	聲母/韻母/停頓(initial/final/sp)
p1	前一個聲母/韻母/停頓
p2	當前聲母/韻母/停頓
p3	下一個聲母/韻母/停頓
	音節(Syllable)
t1	前一個音節聲調
t2	當前音節聲調
t3	下一個音節聲調
w1	當前音節在詞中的位置(從前面數來)
w2	當前音節在詞中的位置(從後面數來)
w6	音節位置屬於詞間或詞中
s1	當前音節在句中的位置(從前面數來)
s2	當前音節在句中的位置(從後面數來)
PM	音節後是否有標點符號 1：「，」 2：「。」 3：「、」 4：其他標點符號 5：無標點符號
	詞(Word)
w3	前一個詞由幾個字組成
w4	當前詞由幾個字組成
w5	下一個詞由幾個字組成
POS1	前一個詞的詞性
POS2	當前詞的詞性
POS3	下一個詞的詞性
	句(Sentence)
s3	前一個句子由幾個字組成
s4	當前句子由幾個字組成
s5	下一個句子由幾個字組成

3.4.5 狀態合併決策樹問題集

決策樹問題集紀錄資料分群規則，HTS 建立決策樹時會依據問題集進行分群。為了達到最佳合併分裂結果，我們考慮了合成單元相關、字相關、詞相關、句相關以及聲調相關五類問題。

- (一) 合成單元相關：考慮當前合成單元及其前後各一個合成單元
 - 聲母發音方法：塞音、非塞音、摩擦音、邊音、清音、濁音...。
 - 聲母發音部位：唇間、舌尖、牙喉...。
 - 韻母發音類別：韻母是以 a、e、i、o、u、ii 為基礎
- (二) 字相關：考慮當前字
 - 字在詞中的順序（正序、倒序）
 - 字在句中的順序（正序、倒序）
- (三) 詞相關：考慮當前詞以及其後詞
 - 詞中字元個數
- (四) 句相關：考慮當前句以及其前後句
 - 句中字元個數
- (五) 聲調相關：考慮當前合成單元及其前後各一個聲調
 - 聲調 1、2、3、4、5、8

3.4.6 停頓預估

我們有了斷詞系統和 HTS 所訓練出的 HMM 模型後，在 TTS 系統實作時，仍然需要停頓資訊的預估，才能合成出更加自然流暢的語音。我們分析語料庫中音檔和其對應的文字檔發現在文字轉語音的過程中，可以由文字直接預估語音的停頓位置；例如在有標點符號的句子和句子之間，通常會有停頓的發生。有良好效能的停頓預估能讓合成語音更加流利，本研究使用條件隨機域模型訓練停頓預估模型，以下將介紹模型特徵樣版的選擇，以及系統效能之評估。

一般人在唸一段語音時幾乎不會在詞內做一個明顯的停頓，因此本研究在停頓預估模型的訓練單元是以詞為單位。而我們從語料庫中發現，詞間是否有停頓和詞性標籤有很大的關係，例如在關係連接詞或是代名詞的前後通常都會有韻律停頓，我們舉例如下：

例如：故所(Cbb)「停頓」對(P)伯公(Nb)個(DE)稱呼(Na)
一定(D)愛(D)用(P)佢(Nh)「停頓」本身(Nh)個(DE)語言(Na)來(D)表現(VC)

而詞長也是影響詞間是否有停頓很大的因素之一，通常長詞之後的停頓機率較短詞高；最靠近詞間的聲母、韻母類別也是我們在此所考慮的重點。我們利用條件隨機域訓練預估詞後是否有停頓的模型，本研究的特徵及特徵樣版如表 3.10 和表 3.11 所表示：

表 3.10: 停頓預估模型之特徵

特徵	含意
PM_n	詞後是否有標點符號
WL_n	詞長
POS_n	詞性
Wi_n	詞從前面數來第一個字的聲母發音類別 1 : b、d、g 2 : p、t、k 3 : m、n、ng 4 : z 5 : c 6 : f、s、h
Wf_n	詞從前面數來最後一個字的韻母發音類別 1 : a、ai、au、am、an、ang、ap、at、ak 2 : e、eu、ieu、iau、im、em、en、ep、et 3 : i、ie、ia、io、iu、iui、ioi、iem、iam、in、ien、ion、iun、iang、iong、iung、ip、iep、iap、it、iet、iut、iak、iok、iuk、iing、iui 4 : o、oi、on、ong、ot、ok 5 : ue、ua、u、uai、ui、uen、uan、un、ung、uang、uet、uat、ut、uk、uak 6 : ii、iim、iin、iip、iit 7 : FNULL

表 3.11: 停頓預估模型之特徵樣版

特徵樣板	含意
$PM_{n-1}, PM_n, PM_{n+1}, (PM_{n-1}PM_nPM_{n+1})(PM_{n-2}PM_{n-1}PM_n)$	詞後是否有 PM 資訊
$WL_{n-2}, WL_{n-1}, WL_n, WL_{n+1}, WL_n, (WL_{n-2}WL_{n-1}WL_nWL_{n+1}WL_n)$	詞長資訊
$POS_{n-1}, POS_n, POS_{n+1}$	POS 資訊
Wi_{n+1}	下一個詞第一字聲母發音類別資訊
Wf_n	當前詞最後一字韻母發音類別資訊
$(Wi_{n+1}Wf_n), (POS_nWi_{n+1}Wf_n)$	組合資訊

3.5 實驗結果

TTS 系統最主要的兩大部分為斷詞單元以及基於 HMM 之語音合成器，本節將針對這兩部分做實驗設計以評估其效能。由於客語斷詞系統是結合中文分詞與詞性標記模型、客語外掛詞典以及客語構詞規則，因此，本節將介紹訓練中文斷詞模型之語料，並先對中文斷詞做效能分析；再結合客語外掛詞典和構詞規則後對客語語料做測試以評估

其系統效能。

3.5.1 中文斷詞系統之效能分析

本實驗所採用的中文斷詞訓練及測試語料庫為 SIGHAN 在中文分詞競賽中所提供標記正確分詞結果的語料。SIGHAN (Special Internet Group of the Association for Computational Linguistics)舉辦的會議是國際間關於中文分詞最有名的研討會，其中訓練語料共有 8350369 個單元，測試語料共有 196867 個單元。而由訓練語料以及特徵樣版所產生的總特徵數共為 3882102 個。而詞性標記系統所使用的訓練及測試語料庫為中央研究院平衡語料庫 3.0 版，此語料庫共有 9286 篇文章，總詞數為 5841942 個。我們從中取出 4426886 個詞做為訓練語料，剩下 484070 詞為測試語料。而詞性標籤共有 46 種，其 46 種詞性關係及詞性意義的對應關係在附錄三表示，本研究在詞性標記訓練部分共產生 5176104 個特徵。我們在此系統中所採用的內部詞典是向中研院中文處理小組所發行的中央研究院平衡語料庫詞典，由中研院資訊所及語言所對語言進行分析所設計的。詞典內容包含詞以及詞性組合標籤，詞典的詞數統計結果如表 3.12。

表 3.12: 中文內部詞典統計表

	詞數
一字詞	7153
二字詞	55662
三字詞	13696
四字詞	11237
五字詞	252
六字詞	114
七字詞	23
八字詞	6
總計	88142

(a) 斷詞效能評估

我們使用精確率(Precision)、召回率(Recall)及 F 分數(F-score)來作為評估中文斷詞實驗系統的效能，此三種方法定義如下式表示：

$$\text{精確率} = \frac{\text{斷詞器正確斷出的詞數}}{\text{斷詞器斷出的總詞數}} \quad (3-12)$$

$$\text{召回率} = \frac{\text{斷詞器正確斷出的詞數}}{\text{參考答案的總詞數}} \quad (3-13)$$

$$F \text{ 分數} = \frac{2 \times \text{精確率} \times \text{召回率}}{\text{精確率} + \text{召回率}} \quad (3-14)$$

本研究的中文分詞系統之精確率、召回率以及 F 分數如表 3.13 所示，由表中可以看出我們在中文分詞系統對中文分詞的效能上已有很好的結果：

表 3.13: 中文分詞實驗結果

	精確率	召回率	F 分數
外部測試語料	95.95%	96.79%	96.37
內部測試語料	98.94%	98.82%	98.88

(b) 詞性標記效能評估

而在詞性標記的實驗中，我們將已有標準分詞結果之中文詞彙序列輸入詞性標記系統，系統將每個詞彙標記出最佳對應詞性標籤輸出；我們使用正確率來作為評估詞性分析系統的效能，正確率之定義如下：

$$\text{正確率} = \frac{\text{系統標記詞性和參考答案相同的詞數}}{\text{總詞數}} \quad (3-15)$$

我們在詞性標記系統測試中，所得到的內部測試正確率 96.01%，而外部測試結果為 94.73%，由實驗數據顯示，我們在中文詞性標記系統已有不錯的效能。

3.5.2 客語分詞和詞性標記

在評估了中文分詞與詞性標記系統的效能後，我們接著加入客語外掛詞典及客語構詞規則，對客語文章進行分詞及詞性標記；本節將介紹客語測試語料以及設計實驗以評估客語斷詞系統之效能。

(a) 測試語料

客語測試語料庫為龔萬灶老師所主筆的「阿啾箭个故鄉」一書，其中包含 42 篇文章，共有 63158 個單元(包含標點符號)；由余秀敏老師將此客語文語料翻譯成相對應之中文平行語料，先從中研院斷詞器得到此中文平行語料之分詞和詞性結果標記，再以人工校正的方法，將此分詞和詞性標籤資訊對應至客語語料，最後得到標記好詞性及分詞結果之客語文章，我們把這份語料當作本實驗的參考標準答案。而此平行語料中，國客語使用詞彙完全一樣的詞共占了總詞數的 48.57%。

在詞性標記測試中我們使用和分詞測試相同的語料，輸入為參考答案的分詞結果，由客語詞性標記系統標記出相對應之詞性，再與參考答案的詞性標籤相比較以分析其效能，此測試語料共有 36450 詞。

(b) 客語分詞實驗

為了測試結合中文條件隨機域、中文詞典以及客語詞典在客語分詞的表現，本研究設計了二個實驗進行測試，分別是(1)實驗 A：僅使用中文斷詞系統對客語文章進行斷詞。(2)實驗 B：加入客語詞典和構詞規則，並採用客語外部詞典優先於中文詞典

(c) 客語詞性標記實驗

我們將參考答案的正確分詞結果輸入至系統標記出每個詞的詞性，為了測試客語外部詞典對於處理客語詞性的效能，在此設計兩個實驗，(a)實驗 A：僅使用中文詞性標記系統標記客語詞性。(b)實驗 B：加入客語詞典至中文詞性標記系統標記客語文章詞性。

(d) 客語分詞實驗結果

實驗 A 及實驗 B 的結果列於表 3.14，我們發現實驗 A，只由中文斷詞系統對客語文章進行斷詞的 F 分數有 74.03，而語料庫中國客語用詞完全相同的比例卻只有 48.57%，這不僅顯示出在客語文章中使用國語詞有很高的比例，更可以發現客語文章的文法結構與中文相似。而加入了客語外掛詞典後，很明顯可以看到實驗 B 的 F 分數較實驗 A 提升了 8.84，可以看出在加入客語外掛詞典後對斷詞效能的提升。

表 3.14: 客語分詞實驗結果

	精確率	召回率	F 分數
實驗 A	71.09%	77.22%	74.03
實驗 B	83.96%	81.80%	82.87

我們詳細比較實驗 A 和實驗 B 的斷詞輸出結果，發現到許多實驗 A 無法解決的搶詞情形在實驗 B 已獲得改善，以下舉出兩個例子說明：

例子 1：實驗 B 在外掛詞典中收錄了客語詞「朝晨」及「打早」

實驗 A 斷詞結果	客家人 在 年初 一朝 晨打 早
實驗 B 斷詞結果	客家人 在 年初一 朝晨 打早
參考斷詞結果	客家人 在 年 初一 朝晨 打早
中文平行語料斷詞結果	客家人 在 大年 初一 清晨 一大早

我們列出了實驗 A、實驗 B、參考答案及中文平行語料之分詞結果，實驗 A 中由於未加入客語詞「朝晨」及「打早」，造成中文詞搶客語詞的情況，此兩詞在實驗 B 的分詞輸出則已可正確斷出。

例子 2：實驗 B 在外掛詞典中收錄了客語詞「祖公祖婆」及「共下」

實驗 A 斷詞結果	請 祖公 祖婆 共 下來 過年
實驗 B 斷詞結果	請 祖公祖婆 共下 來 過年
參考斷詞結果	請 祖公祖婆 共下 來 過年
中文平行語料斷詞結果	請 祖先 一起 來 過年

實驗 A 的客語字串「共下來」被斷成中文詞「共」和「下來」，再將「共下」加入外掛詞典後，可正確斷出「共下」一詞。

(e) 客語詞性標記實驗結果

實驗 A 及實驗 B 對已給定的客語文章分詞結果標記詞性，正確率如表 3.15 所表示；由中文詞性標記系統直接標記客語文章詞性的正確率為 59.41%，而加入外掛詞典後正確率上升了 17.73%，由於在詞性標記階段，條件隨機域會依據詞典內的詞性組合給予每段文句最佳的配對詞性標籤，而未加入客語詞典時，系統無法找到其對應之詞性標籤，因此在實驗 B 的正確率較實驗 A 上升了許多。

表 3.15: 客語詞性標記實驗結果

	正確率
實驗 A	59.41%
實驗 B	77.14%

3.5.3 停頓預估模型之效能分析

在本研究中男性語料庫經由標記及修正切割位置後，得到各音節聲母、韻母及音節後停頓資訊，我們將切割位置中 25ms 以上的停頓保留，而 25ms 秒以下皆視為沒有停頓。其中訓練語料共有 36991 個詞，測試語料包含 4095 詞。我們在此設計了兩個實驗，以評估由條件隨機域訓練停頓預估模型的效能，分別為(1)實驗 A：詞後有標點符號則預估為有停頓。(2)實驗 B：由條件隨機域訓練模型預估詞後是否有停頓。

實驗 A：詞後有標點符號則預估為有停頓

實驗 B：由條件隨機域訓練模型預估出詞後是否有停頓

我們使用正確率和以下四種機率值來作為評估詞性分析系統的效能，其定義如下：

$$\text{正確率} = \frac{\text{系統標記是否有停頓和參考答案相同的詞數}}{\text{總詞數}} \quad (3-16)$$

$$P(1|1) = \frac{\text{預估結果有停頓}}{\text{正確答案有停頓}} \quad (3-17)$$

$$P(1|0) = \frac{\text{預估結果有停頓}}{\text{正確答案沒有停頓}} \quad (3-18)$$

$$P(0|0) = \frac{\text{預估結果沒有停頓}}{\text{正確答案沒有停頓}} \quad (3-19)$$

$$P(0|1) = \frac{\text{預估結果有停頓}}{\text{正確答案有停頓}} \quad (3-20)$$

實驗 A、實驗 B 內部及外部測試語料的實驗正確率如表 3.16 表示，而其四種機率結果列在 3.17。實驗 A、B 的 $P(1|0)$ 明顯比 $P(0|1)$ 低，這樣的結果對於文字轉語音中的停頓預估單元來說是合理的，在沒有停頓的情況下預估為有停頓的錯誤會造成合成語音的不自然；而實驗 B 的 $P(1|1)$ 較實驗 A 明顯上升許多；而在整體正確率方面，實驗 B 的內部測試及外部測試均較實驗 A 的正確率高，顯示出加入條件隨機域模型的停頓預估比只在標點符號處預估為停頓的效能提高。

表 3.17: 實驗正確率

	正確率
實驗 A	85.71%
實驗 B 內部測試語料	87.04%
實驗 B 外部測試語料	88.04%

表 3.18: 實驗之四種機率值

	實驗 A	實驗 B 內部測試語料	實驗 B 外部測試語料
$P(0 0)$	98.51% $\left(\frac{22970\text{個}}{23317\text{個}}\right)$	94.83% $\left(\frac{22111\text{個}}{23317\text{個}}\right)$	94.14% $\left(\frac{2409\text{個}}{2559\text{個}}\right)$
$P(1 0)$	1.49% $\left(\frac{347\text{個}}{23317\text{個}}\right)$	5.17% $\left(\frac{1206\text{個}}{23317\text{個}}\right)$	5.86% $\left(\frac{150\text{個}}{2559\text{個}}\right)$
$P(1 1)$	61.47% $\left(\frac{7569\text{個}}{12314\text{個}}\right)$	75.18% $\left(\frac{9258\text{個}}{12314\text{個}}\right)$	74.03% $\left(\frac{1035\text{個}}{1398\text{個}}\right)$
$P(0 1)$	38.53% $\left(\frac{4745\text{個}}{12314\text{個}}\right)$	24.82% $\left(\frac{3056\text{個}}{12314\text{個}}\right)$	25.97% $\left(\frac{363\text{個}}{1398\text{個}}\right)$

3.5.4 客語語音合成實驗結果與分析

(a) 實驗與測試語料

實驗所採用的語料為男性語料庫語，音檔有 458 個，共有 63158 個音節，平均發音速度為 0.24syllable/second，我們從中取出 407 個音檔做為訓練語料，其餘 51 個當作測試語料。女性語料庫中共有 292 個語音檔，總音節數為 42337 個，平均發音速度為 0.22syllable/second，其中 262 個音檔為訓練語料，測試語料共有 30 個音檔。

(b) 客觀評估

我們採用客觀評估對模型的效能進行評估，評估的語料以及標準如表 3.19 所示：

表 3.19: 客觀評估

	語料/受測者	評估標準
客觀評估	內部測試語料	RMSE
	外部測試語料	

我們採用均方根誤差(Root Mean Square Error, RMSE)作為本實驗的評估標準，計算目標和其預測的音高軌跡、音節長度以及頻譜能量之間的誤差，例如目標之音高軌跡為

$S = (s_1, s_2, \dots, s_N)$ ，而預測音高軌跡是 $T = (t_1, t_2, \dots, t_N)$ ，則 RMSE 之計算方式如下：

$$RMSE(S, T) = \sqrt{\frac{\sum_{i=1}^N (s_i - t_i)^2}{N}} \quad (3-21)$$

本實驗分別對男性及女性語料庫，計算在內部測試語料和外部測試語料的音高軌跡和音長之 RMSE 值，如表 3.20 及表 3.21 所示。我們由此表看出男性語料模型在預測音高軌跡有不錯的表現，比女性語料模型佳，這是由於女性在講話時，音高軌跡範圍較廣，而且本實驗的女性發音人聲音特性較沙啞，在原始語料音高參數求取上較不準確，因此在基頻的 RMSE 值較男性語料高。

表 3.20: 基頻之 RMSE 值

	內部測試語料	外部測試語料
女性語料庫	37.35(Hz)	45.64(Hz)
男性語料庫	16.08(Hz)	19.4(Hz)

表 3.21: 音長之 RMSE 值

	內部測試語料	外部測試語料
女性語料庫	37.31(ms)	33.13(ms)
男性語料庫	45.13(ms)	47.57(ms)

(c) 主觀評分

在主觀評估部分，使用平均主觀分數（Mean Opinion Score, MOS）當做評估標準；評估的語料以及標準如表 3.22 所示：

表 3.22: 主觀評估

	語料	受測者	評估標準
主觀評估	男性語料模型 實驗 A：10 句 實驗 B：10 句 女性語料模型 實驗 A：10 句 實驗 B：10 句	男性：5 位 女性：5 位	MOS

在男、女性語料庫實驗 A 及實驗 B 合成各 10 句語音，此文句不為內部訓練語料，由 5 位男性和 5 位女性聽得懂客語的受測者進行評估實驗，實驗時四組語音以先後次序播放，並要求受測者根據自己的觀感給予合成語音在自然度上的分數，評分標準如表 3.23：

表 3.23: MOS 評分標準

評等	分數	說明
優	5	合成語音非常自然
良	4	合成語音自然
可	3	合成語音表現普通
差	2	合成語音不自然
劣	1	合成語音非常不自然

我們在 MOS 所得到男性合成語音的分數如圖 3.7 所表示：

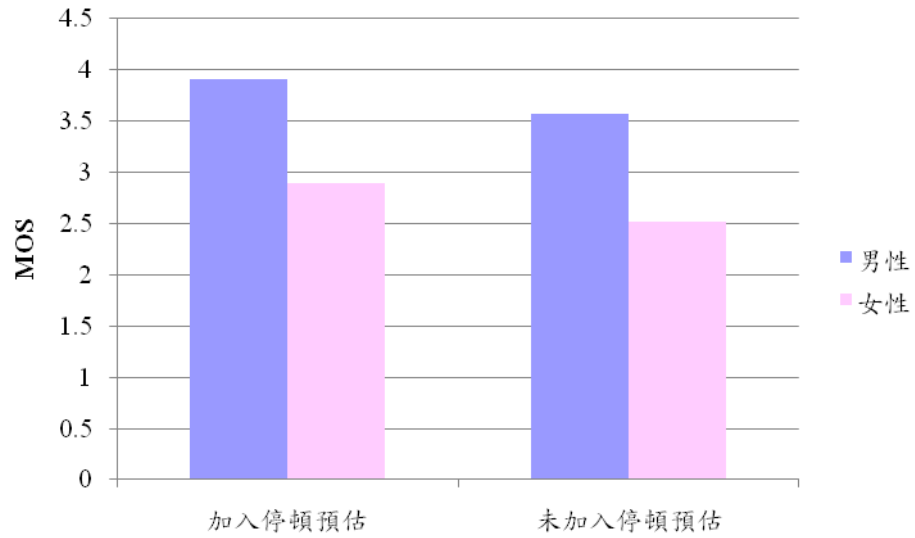


圖 3.7: 實驗之 MOS 比較

我們在男性模型的 MOS 表現都較女性 MOS 表現良好，這與客觀評估女性模型在基頻的 RMSE 表現較差相對應，而加入停頓預估系統的實驗，不管是男性模型或女性模型都比沒有加入停頓預估的合成語音分數高，顯示出我們在加入了停頓預估單元後，能合成出較自然的語音。

四、 客語及國客雙語語音辨認系統之建立

本研究是著重在客語及國客語混合之雙語言語音辨認之研究，一般雙語的語音辨認是針對兩個健全的單一語言辨認系統，其可以使用的資源較為豐富。而本研究因為客語為台灣方言之一，所以語料量並不足以訓練健全的單語言的辨認系統，但是客語和國語同為漢語語系語言，所以可借助國語幫助客語完成其辨認系統。目前政府也在推行客家文化，也有客語語文能力的相關認證，所以相信客語的重要性亦是與日俱增的，相信國、客語連續語音辨認是一件值得研究的課題。圖 4.1 為：基本語音辨認系統架構圖。

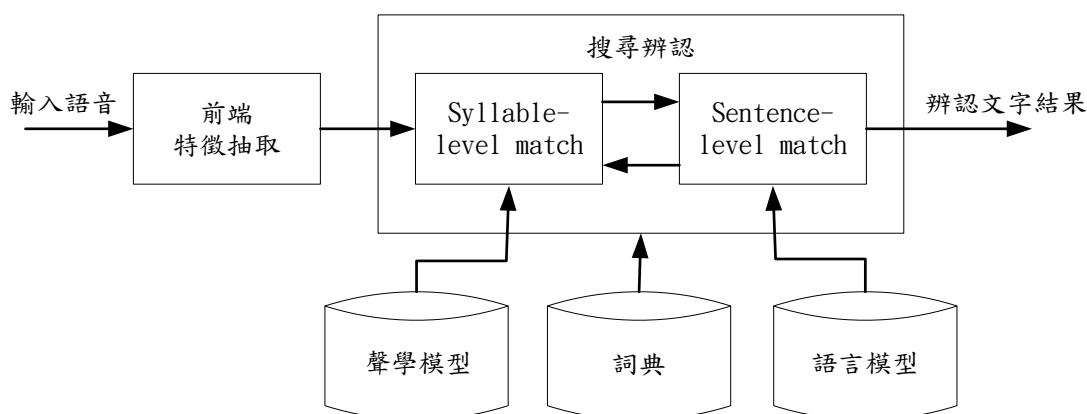


圖 4.1：基本語音辨認系統架構圖

由圖 4.1 可看出，發話者的語音需要經過特徵參數抽取、聲學模型的搜尋辨認及語言模型的搜尋辨認，最後才能得到文字辨認結果。簡單來說，一個基本的語音辨認系統主要是在於聲學模型與語言模型的訓練以及詞典的挑選。而本專題研究的目的是在於客語及國客雙語的語音辨認，其系統架構亦是建立於基本語音辨系統架構上，主要是針對聲學模型、語言模型以及詞典做改善及調整的動作，以期語音辨認系統能夠達到不錯的辨認率。

4.1 實驗語料介紹

本研究實驗語料主要分為文字語料和語音語料兩個部份，其中文字語料做為語言模型訓練使用，而語音語料則做為聲學模型訓練及測試語音辨認系統之效能使用。

4.1.1 語音語料庫介紹

實驗所使用的語音語料庫分為國語語音語料庫、客語語音語料庫以及國客雙語語音語料庫三個。其中國語語音語料庫及客語語音語料庫做為聲學模型的訓練及測試使用，國客雙語語音語料庫則做為測試使用。接著，將對這三個語料庫做更為詳細的介紹。

➤ 國語語音語料庫

實驗所使用的國語語音語料庫為 TCC300 語料庫，語料庫是由台灣大學、成功大學、交通大學各自擁有之語料庫集合而成，各校錄製之目的是為語音辨認研究，屬於麥克風朗讀語音。其中台大語料庫主要包含詞及短句，文章經過仔細設計，考慮了音節及其相連出現機率，由 100 人錄製而成；成大及交大語料庫主要包含長文語料，文章由中研院提供之 500 萬詞詞類標示語料庫中選取，每篇文章包含數百字，再切割成 3 至 4 段，每段含至多 231 字，由 200 人朗讀錄製，每人所讀文章皆不相同。詳細統計資訊如表 4.1 所示。

表 4.1：TCC300 語料庫資訊統計表

學校名稱	文章屬性	語者總數		總音節數		總檔案數	
台灣大學	短文 (平衡句)	男	50	男	27,541	男	3,425
		女	50	女	24,677	女	3,084
		總計	100	總計	52,218	總計	6,509
交通大學	長文	男	50	男	75,059	男	622
		女	50	女	73,555	女	616
		總計	100	總計	148,614	總計	1,238
成功大學	長文	男	50	男	63,127	男	588
		女	50	女	68,749	女	582
		總計	100	總計	131,876	總計	1,170

➤ 客語語音語料庫

台灣客家話有許多次方言，其中以四縣腔為最通行的腔調，本研究考量到客語的代表性和實用性，因此選用客語人口最多的苗栗四縣客家話為發展國、客語語音辨認技術的語料來源。以下所稱之客語皆為苗栗四縣客家話。

實驗所使用的客語語音語料庫，為交通大學語音處理實驗室分四個時期錄製而成，目的是為語音辨認研究，屬於麥克風朗讀語音。文章內容主要為客語故事[34]，主要由苗栗的兩位退休教師，一位是龔萬灶老師、另一位為陳碧娥老師所撰寫；並加入中華大學余秀敏老師所新增的客語文句。由 92 人朗讀錄製而成，詳細統計資料如表 4.2 所示。

表 4.2：客語語音語料庫資訊統計表

語者總數		總音節數		總檔案數	
男	43	男	70,193	男	4,627
女	49	女	83,718	女	5,602
總計	92	總計	153,911	總計	10,229

➤ 國客雙語語音語料庫

實驗所使用的國客雙語語音語料庫，為交通大學語音處理實驗室錄製而成，目的是為語音辨認研究，屬於麥克風朗讀語音。文句由中華大學余秀敏老師根據日常國客語夾

雜使用情況所撰寫，文句內容分為客語為主文句及國語為主文句，各 100 句。由 20 人錄製而成，每人 20 句國語為主文句、20 句客語為主文句。詳細統計如表 4.3 所示。

表 4.3：國客雙語語音語料庫資訊統計表

文句屬性	語者總數		總音節數		總檔案數	
客語為主	男	10	男	2,480	男	200
	女	10	女	2,480	女	200
	總計	20	總計	4,960	總計	400
國語為主	男	10	男	2,478	男	200
	女	10	女	2,478	女	200
	總計	20	總計	4,956	總計	400

4.1.2 文字語料庫介紹

實驗所使用的文字語料庫依照語言分為國語文字語料庫、客語文字語料庫兩個。主要是做為語言模型的訓練使用。

➤ 國語文字語料庫

實驗所使用的國語文字語料庫，來源有兩個：光華雜誌及 NTCIR，光華雜誌為一般雜誌內容文章，蒐集的範圍為 1976 年至 2000 年；NTCIR 為各個不同學科領域的文章構成，是建立資訊檢索系統的標竿測試集。詳細統計資訊如表 4.4 所示：

表 4.4：國語文字語料庫資訊統計表

來源	詞數(Word)	字數(Character)
光華雜誌	8,945,660	15,691,089
NTCIR	108,217,471	201,075,931
合計	117,163,131	216,767,020

➤ 客語文字語料庫

因為客語文字語料庫其文字量明顯較國語稀疏，所以此節將特別對與客語文字相關的語料加以介紹。實驗所使用的客語文字語料庫，主要分為：無標記的客語文章、人工標記的客語文章及詞典三個部份，以下將分別對這三個部份做詳細介紹：

(a) 無任何標記的客語文章(以下稱為 *textA*)

此文章來源為網路上擷取而來，主要來源為全國語文競賽朗讀文章，文章內容大部份為故事散文集。經過文字正規化及使用客語詞典斷詞後，詳細統計資訊如表 4.5 所示，總詞條數有 143,341 個。

表 4.5：無標記的客語文章資訊統計表

	一字詞	二字詞	三字詞	四字詞	五字詞	六字詞	七字詞	八字詞	總計
詞條數	2,797	8,722	2,485	946	89	17	15	42	15,115
詞頻	81,672	51,027	8,591	1,810	126	23	17	51	143,341

(b) 人工標記詞性、國語翻譯的客語文章(以下稱為 *textB*)

此文章來源為龔萬灶老師撰寫的客語文集，並且經由余秀敏老師人工標記詞類及國語翻譯資訊而成，其文章內容為客語故事。使用的詞類標記共有 49 類，詳細詞類標記資訊如附錄三；此文章平均一個客語詞條有 1.14 個詞類標記、1.28 個國語翻譯。詳細統計資訊如表 4.6 所示。

表 4.6：人工標記客語文章資訊統計表

	一字詞	二字詞	三字詞	四字詞	五字詞	六字詞	七字詞	八字詞	總計
詞條數	1,179	4,825	1,466	534	47	9	1	5	8,068
詞頻	25,729	16,610	3,393	758	51	10	1	5	46,581
平均翻譯數	1.90	1.19	1.15	1.07	1.00	1.00	1.00	1.00	1.28
平均詞類數	1.53	1.09	1.05	1.02	1.00	1.00	1.00	1.00	1.14

(c) 客語詞典(以下稱為 *dict*)

此客語詞典收錄來源有三個：台北市客委會、行政院教育部及龔老師文集。此詞典紀錄每個客語詞條的發音及詞類，平均一個客語詞條有 1.33 個詞類標記。詳細統計資訊如表 4.7 所示。

表 4.7：客語詞典資訊統計表

	一	二	三	四	五	六	七	八	總計
詞條數	6,643	23,690	7,454	5,224	323	84	85	18	43,521
平均 pos 數	2.42	1.16	1.08	1.06	1.04	1.01	1.00	1.00	1.33

另外，此客語詞典收錄的部分客語詞條(26,524 個)有國語翻譯的對應，這部份的客語詞條平均有 1.12 個國語翻譯。此部份翻譯標記的收錄來源有台北市客委會及余老師人工標記的客語文章。詳細統計資訊如表 4.8 所示。

表 4.8：國客雙語對照詞典資訊統計表

	一	二	三	四	五	六	七	八	總計
詞條數	1,225	16,762	4,924	3,295	205	59	44	10	26,524
平均翻譯數	1.94	1.09	1.08	1.03	1.00	1.00	1.00	1.00	1.12

4.2 客語與國語之間的語文特性

客語和國語都是漢語語系的語言，其音節結構和國語一樣，可以劃分為兩大部分：聲母和韻母，詳細結構請參閱第二章對客語音韻的簡介。

4.2.1 客語與國語的平行語料分析

此部份是由 *TextB* 分析人工標記的客語文章中客語文句與國語文句之間的關係。發現不管在文字撰寫或文法結構上，國語與客語都極為相似，

例如：

客語文句：	拜	伯公	係	客家人	異	平常	也	係	異	要緊	事情
國語文句：	拜	土地公	是	客家人	十分	平常	也	是	十分	重要	的事情
詞類標記：	VC	Nb	SHI	Na	Dfa	VH	D	SHI	Dfa	VH	DE Na

客語文句：	原本	嚴肅		面容	露出	微微仔	笑意	
國語文句：	原本	嚴肅	的	面容	露出	微微	的	笑意
詞類標記：	D	VH	DE	Na	VC	D	DE	Na

客語文句：	從	後生	嫁	來	這	隻	莊仔	到	今晡日	歸仙
國語文句：	從	年輕	嫁	來	這	個	莊子	到	今天	歸仙
詞類標記：	P	VH	VC	D	Nep	Nf	Nb	P	Nd	VA

上述所舉的例子，可以明顯看出國語與客語文句有相同的文法結構，接著由客語詞條與翻譯的對照，我們發現國語與客語在漢字的使用上，可分為三種：

➤ 客語詞條和國語翻譯『相同』

此部份是指客語詞條和國語翻譯的漢字使用上是一樣的，

例如：

客語詞條：二十一世紀、工廠、山盟海誓

國語翻譯：二十一世紀、工廠、山盟海誓

➤ 客語詞條和國語翻譯『相似』

此部份是指客語詞條和國語翻譯的漢字使用上是相似的，通常是語贅詞的使用不同及客語有特有的語助詞，例如：

客語詞條：七十零 男仔人 年仔 手仔

國語翻譯：七十多 男人 年 手

➤ 客語詞條和國語翻譯『不同』

此部份是指客語詞條和國語翻譯的漢字使用上是不一樣，通常是客語特有用法或是一字詞，

例如：

客語詞條： 三十暗晡 八月半

國語翻譯： 除夕夜 中秋節 的

表 4.9: TextB 之中國客翻譯之型態分析

分類	數量(比例)
相同	4,482(47.1%)
相似	2,597(27.3%)
不同	2,437(25.6%)
Total	9,516

4.2.2 客語與國語的詞類分析

此部份是對上述人工標記的客語文章與國語文章(NCTIR 及光華雜誌)觀察詞類標記的分布情況，客語與國語詞類標記分布情形如圖 4.2 至圖 4.4 所示，我們可以看出客語詞條和國語詞條在文章中，其詞類標記的所佔比例是非常近似的。因此本研究假設客語文句的詞類標記和國語文句是相同的。

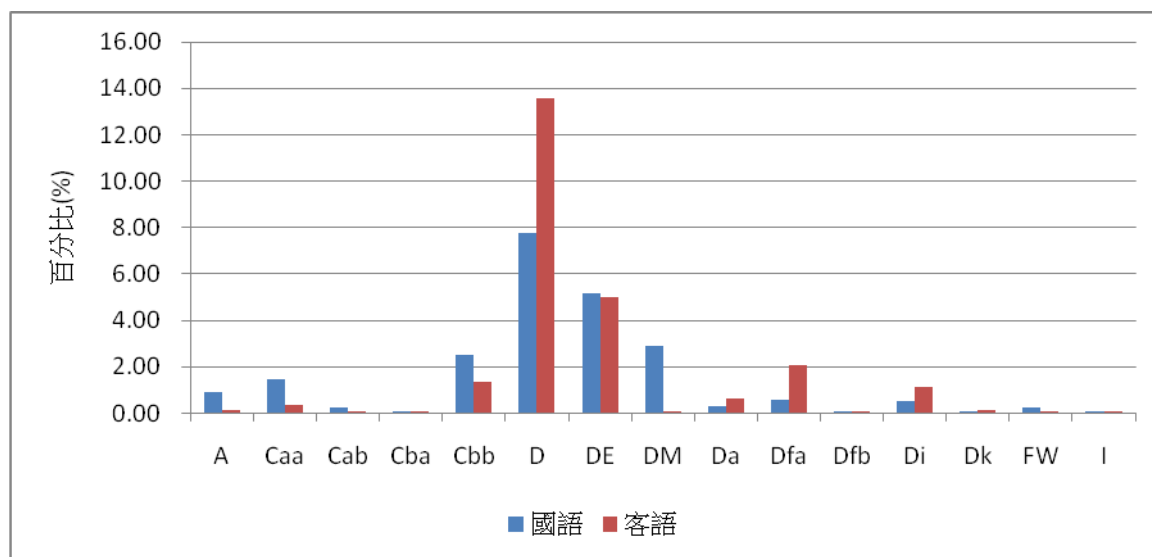


圖 4.2：客語詞類與國語詞類標記分佈比較(1)

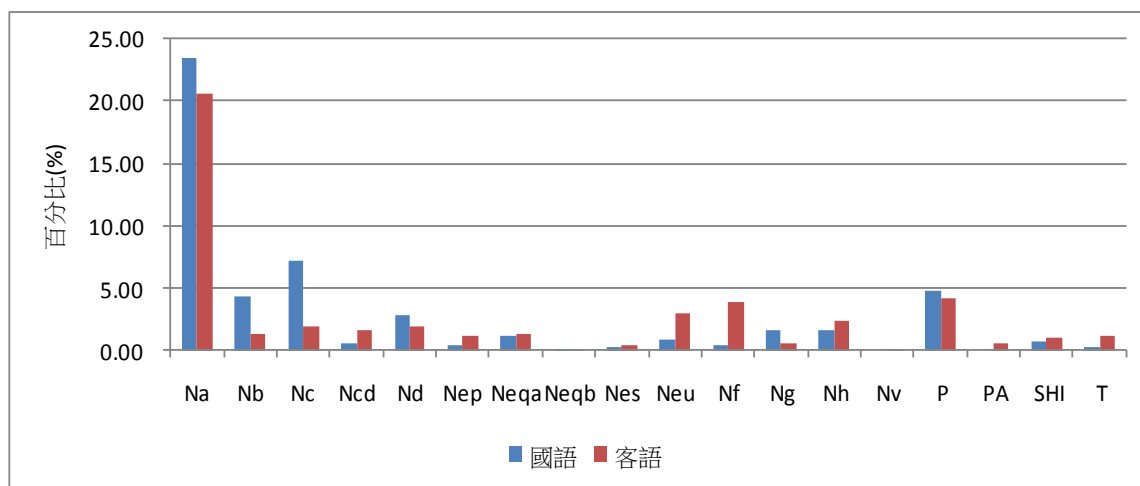


圖 4.3：客語詞類與國語詞類標記分佈比較(2)

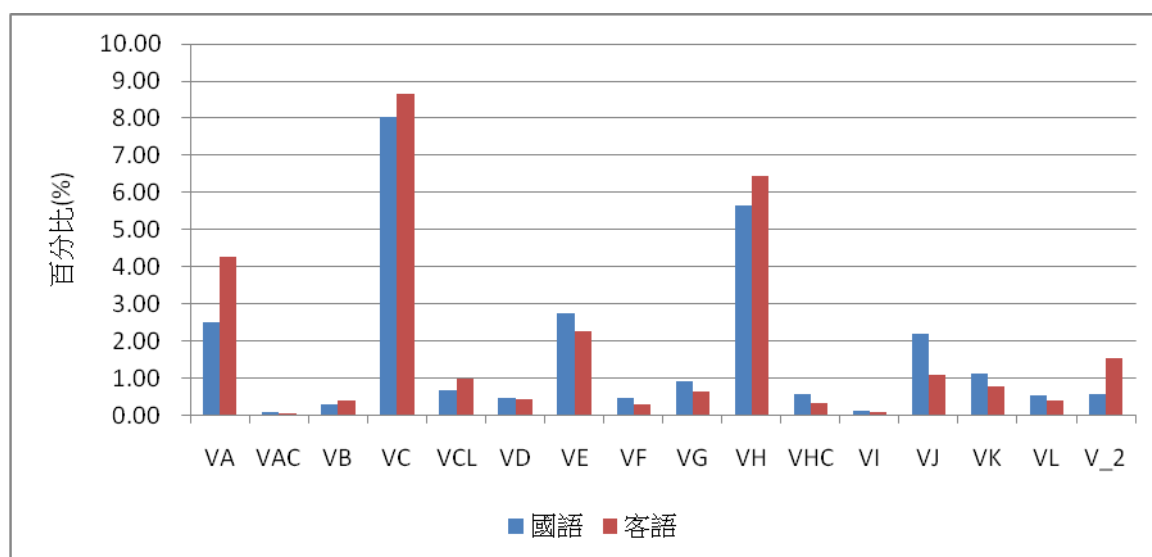


圖 4.4：客語詞類與國語詞類標記分佈比較(3)

4.3 聲學模型

4.3.1 訓練語料及測試語料

本研究所使用的語音語料分為國語語音語料、客語語音語料及國客語夾雜語音語料。其中國語語音語料及客語語音語料做為訓練及測試聲學模型用，詳細統計資料如表 4.10 及表 4.11 所示；而國客語夾雜語音語料做為測試聲學模型使用，詳細統計資料如表 4.12。值得注意的是，客語語音語料庫因為錄製的時期不同，造成不同時期語者唸的提示卡語料數目不同，造成語料分配較為不均勻的情況發生，主要有三個語者(2 女 1 男)各包含了 1 萬多個音節的情況發生，將這三位語者皆列為訓練語料。

表 4.10：客語語音語料庫統計

	訓練語料	測試語料
673 syllable	134,612	19,388
總段落數	8,658	1,571

表 4.11：國語語音語料庫統計

	訓練語料	測試語料
411 syllable	300,728	31,252
總段落數	8,036	843

表 4.12：國客語夾雜語音語料庫統計

	客語為主	國語為主	合計
1084 syllable	4,710	4,707	9,417
總段落數	380	380	760

4.3.2 聲學模型之建立

(a) 特徵參數抽取

在訓練模型前，首先必須獲得足以充分描述語音特性，且參數量較原語音信號小之特徵參數，而在語音處理當中，最廣泛為人使用之特徵參數為梅爾頻率倒頻譜係數 (Mel-Frequency Cepstrum Coefficient, MFCC)，本研究也將使用此特徵參數，以 32 毫秒之漢明窗 (Hamming window) 且每位移 10 毫秒為一筆資料，求取 12 維 MFCC 加上 1 維能量係數，以及這 13 維係數之一階與二階變量 (delta and delta-delta) 為特徵參數，但單純的能量在參數中較為缺乏鑑別性，因此除去能量係數，得到 38 維向量做為本研究語音資料之聲學特徵參數。在本研究也將利用倒頻譜平均值正規劃法 (Cepstrum Mean Normalization, CMN) 藉此消除不同語音信號之通道效應。

(b) 單語聲學模型之建立

由於使用 Flat Start 訓練聲學模型，此種做法必須花費較久的時間才能得到正確的模型，且語句較長時也容易發生模型位置錯誤之情況。因此本研究利用音節內右相關聲/韻母模型 (Right-context-dependent Initial / Final Model, RCD) 對語音語料做音節的切割，再根據音節的切割位置訓練單語聲學模型，其建立流程如圖 4.5 所示。

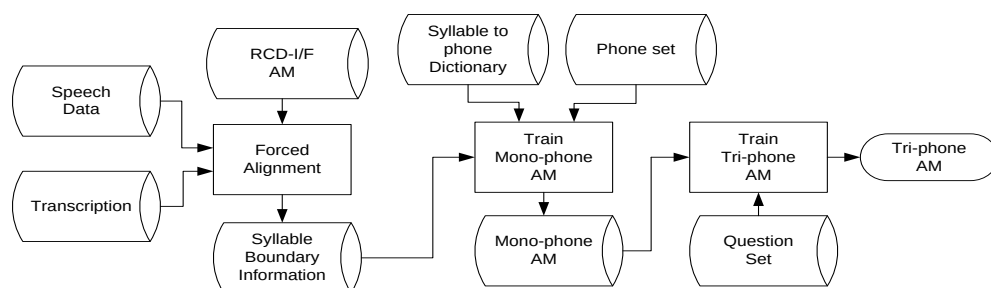


圖 4.5：單語聲學模型之建立流程

本研究之單語聲學模型以音素(phone)為單位，採用跨詞三連音素模型(Cross-word Tri-phone Model, Tri-phone)，每一個音素之 HMM 模型採用 3 個由左至右(left-to-right)的狀態(state)表示。使用跨音節三連音素模型時，為表示完整的聲學模型須包含所有的音素組合，也就是音素個數的三次方，其中必然包含在目標辨認語料中卻無訓練語料之組合，因此對所有 HMM 模型的狀態採用分類回歸決策樹(Classification and Regression Trees, CART)做為參數分享的方法，使用語言學的問題來做為決策問題，詳細決策問題如附錄四。

(c) 雙語聲學模型之建立

由於沒有國、客語夾雜的訓練語料，因此，本研究首先對已訓練出的國、客語音素模型(Mono-phone Model)做 HMM 模型的距離計算，接著根據距離及所定義的合併規則做音素模型的合併(距離計算方式及合併規則將在 4.3.3 節介紹)，最後完成雙語的跨詞三連音素模型(Bilingual Tri-phone Model)，其建立流程如圖 4.6 所示。

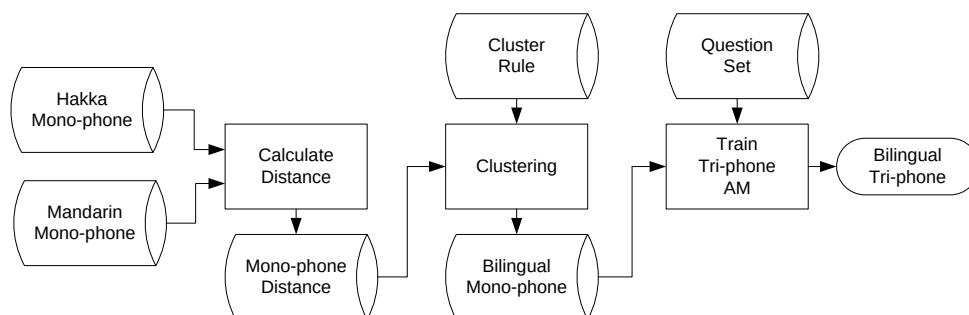


圖 4.6：雙語聲學模型建立流程

4.3.3 雙語音素集

由圖 4.6 雙語聲學模型建立流程得知，在雙語語音辨認中雙語音素集的定義是不可或缺的，因為聲學模型的訓練是根據基本發音單位去做訓練的，若定義的發音單位不合適則會直接影響最終的語音辨認效能，而雙語音素就是雙語的基本發音單位。目前已經有定義好的國語及客語單語音素集。所以本研究就是經由已定義的國語音素集與客語音素集建立一個良好的雙語音素集。並且希望所定義的雙語音素集不影響原先的單一語言音素集。

(a) 單語音素集

單語音素集是考量語言實際使用時的發音狀況，完整描述語言的特性定義而成。因為國語和客語皆為漢語語系，音節結構都為聲母加上韻母，所以音素的定義也將音節結構位置考慮進去，詳細如表 4.13 以及表 4.14。

表 4.13：國語音素集

聲母(21)			韻母(17)			
			介音 (3)	元音 (8)	韻尾 (4)	空韻母(2)
M_b	M_n	M_m	M_yi1	M_e	M_en	M_FNULL1
M_c	M_p		M_yu1	M_eh	M_ng	M_FNULL2
M_ch	M_q		M_wu1	M_o	M_yi3	
M_d	M_r			M_a	M_wu3	
M_f	M_sh			M_er		
M_g	M_s			M_yi2		
M_h	M_t			M_yu2		
M_j	M_x			M_wu2		
M_k	M_z					
M_l	M_zh					

表 4.14：客語音素集

聲母(21)			韻母(17)			
			介音 (2)	元音 (6)	韻尾 (8)	空韻母(1)
H_HH	H_m	H_l	H_yi1	H_a	H_ek	H_FNULL
H_NH	H_n		H_wu1	H_eh	H_em	
H_b	H_ng			H_ii	H_en	
H_c	H_p			H_o	H_eng	
H_d	H_q			H_wu2	H_ep	
H_f	H_s			H_yi2	H_et	
H_g	H_t				H_wu3	
H_h	H_v				H_yi3	
H_j	H_x					
H_k	H_z					

(b) 音素直接合併

此部份是將國語音素集與客語音素集直接合併產生雙語音素集，此音素集共有 76 個音素，詳細如表 4.15 所示。

表 4.15：直接合併之雙語音素集

聲母(42)			韻母(34)			
			介音 (5)	元音 (14)	韻尾 (12)	空韻母(3)
M_b	M_s	H_k	M_yi1	M_e	M_en	M_FNULL1
M_c	M_t	H_m	M_yu1	M_eh	M_ng	M_FNULL2
M_ch	M_x	H_n	M_wu1	M_o	M_yi3	H_FNULL
M_d	M_z	H_ng	H_yi1	M_a	M_wu3	
M_f	M_zh	H_p	H_wu1	M_er	H_ek	
M_g	M_m	H_q		M_yi2	H_em	
M_h	H_HH	H_s		M_yu2	H_en	
M_j	H_NH	H_t		M_wu2	H_eng	
M_k	H_b	H_v		H_a	H_ep	
M_l	H_c	H_x		H_eh	H_et	
M_n	H_d	H_z		H_ii	H_wu3	
M_p	H_f	H_l		H_o	H_yi3	
M_q	H_g			H_wu2		
M_r	H_h			H_yi2		
M_sh	H_j					

(c) log-likelihood measurement

Log-likelihood measurement [35,36] 是對相同音節結構分類下的音素做計算距離的動作，令 $O_{i,k}$ 為第 k 段標示為音素模型 i 的特徵參數序列，則兩個音素模型 M_i 及 M_j 之間的距離如式(3.1)所示：

$$D(M_i, M_j) = \frac{1}{K} \sum_{k=1}^K f(O_{i,k} | M_i) - \frac{1}{K} \sum_{k=1}^K f(O_{i,k} | M_j) \quad (4.1)$$

因式(3.1)算出的距離並非對稱，故對其取平均值而得到 log-likelihood measurement：

$$D_{i,j} = \frac{1}{2} (D(M_i, M_j) + D(M_j, M_i)) \quad (4.2)$$

其中， $f(O_{i,k} | M_j)$ 為 $O_{i,k}$ 對音素模型 M_j 的 Log-likelihood。其計算如式(4.3)所示，使用威特比演算法(Viterbi algorithm)找出最佳的路徑：

$$f(O_{i,k} | M_j) = \frac{1}{T} \log \left(\max_x \left\{ a_{x(0)x(1)} \prod_{t=1}^T b_{x(t)}(O_{i,k,t}) a_{x(t)x(t+1)} \right\} \right) \quad (4.3)$$

$O_{i,k,t}$ 表示為在 $O_{i,k}$ 中的第 t 個聲學特徵參數 sample， $x(t)$ 表示為第 t 個 sample 所屬的 state， $a_{x(t)x(t+1)}$ 為第 $x(t)$ 至 $x(t+1)$ 的轉移機率， $b_{x(t)}$ 為第 $x(t)$ 個 state 的 likelihood。

$$b_j(O_{i,k,t}) = \sum_{m=1}^M c_{jm} N(O_{i,k,t}; \mu_{jm}, \Sigma_{jm}) \quad (4.4)$$

m 表示為 b_j 這個 state 中第 m 個 mixture， c_{jm} 表示為 b_j 這個 state 中第 m 個 mixture 的 weight。最後，log-likelihood measurement 實作流程分為三個步驟，步驟如下：

- 步驟 1：各自訓練國語及客語的音素模型，並對原始語料做音素的切割而得音素的段落標記(forced alignment)。
- 步驟 2：對於每一個國語或客語音素，以式(3.2)計算對另一種語言(客語或國語)音素的 log-likelihood measurement，也就是 $D_{i,j}$ 。
- 步驟 3：根據計算出的 $D_{i,j}$ 及國客語一對一的原則做距離大小的排序。

表 4.16 為根據 log-likelihood measurement 計算兩語言之間音素模型的距離，根據距離及一對一做排序的動作。

表 4.16：log-likelihood distance

排名	距離	客語音素		國語音素	
		音素模型	音節結構分類	音素模型	音節結構分類
1	4.264	H_eng	韻尾	M_ng	韻尾
2	4.408	H_a	元音	M_a	元音
3	4.444	H_en	韻尾	M_en	韻尾
4	4.815	H_yi2	元音	M_yi2	元音
5	4.972	H_eh	元音	M_eh	元音
6	5.065	H_l	聲母	M_l	聲母
7	5.251	H_wu2	元音	M_wu2	元音
8	5.472	H_s	聲母	M_s	聲母
9	6.127	H_m	聲母	M_m	聲母
10	6.469	H_yi1	介音	M_yi1	介音
11	6.479	H_n	聲母	M_n	聲母
12	6.526	H_x	聲母	M_x	聲母
13	6.542	H_et	韻尾	M_yi3	韻尾
14	6.646	H_z	聲母	M_z	聲母
15	6.669	H_t	聲母	M_t	聲母
16	6.678	H_p	聲母	M_p	聲母
17	6.684	H_b	聲母	M_b	聲母
18	6.726	H_h	聲母	M_h	聲母
19	6.878	H_o	元音	M_o	元音
20	6.886	H_j	聲母	M_j	聲母
21	6.936	H_c	聲母	M_c	聲母
22	7.214	H_d	聲母	M_d	聲母
23	7.286	H_f	聲母	M_f	聲母
24	7.589	H_ii	元音	M_FNULL2	空韻母
25	8.070	H_q	聲母	M_q	聲母
26	8.277	H_ek	韻尾	M_wu3	韻尾
27	9.324	H_wu3	韻尾	M_e	元音
28	9.363	H_ng	聲母	M_r	聲母
29	9.437	H_g	聲母	M_g	聲母
30	10.306	H_yi3	韻尾	M_er	元音
31	10.860	H_FNULL	空韻母	M_yu2	元音
32	10.916	H_k	聲母	M_k	聲母
33	11.791	H_HH	聲母	M_sh	聲母
34	12.343	H_wu1	介音	M_wu1	介音
35	13.354	H_em	韻尾	M_FNULL1	空韻母
36	16.263	H_NH	聲母	M_yu1	介音
37	18.600	H_v	聲母	M_ch	聲母
38	22.677	H_ep	韻尾	M_zh	聲母

接著，根據相同漢語音標、相同音節結構位置以及距離大小後，決定合併 22 個音素，其中聲母有 14 個、介音 1 個、元音 5 個以及韻尾 2 個。詳細如表 4.17 以及 4.18 所示。

表 4.17：雙語音素集，C_表示合併後的音素

聲母(28)			韻母(26)			
			介音 (4)	元音 (9)	韻尾 (10)	空韻母(3)
C_b	C_s	H_v	C_yi1	C_a	C_eng	H_FNULL
C_c	C_t	M_ch	H_wu1	C_eh	C_en	M_FNULL1
C_d	C_x	M_g	M_yu1	C_o	H_ek	M_FNULL2
C_f	C_z	M_k	M_wu1	C_wu2	H_em	
C_h	H_HH	M_q		C_yi2	H_ep	
C_j	H_NH	M_r		H_ji	H_et	
C_l	H_g	M_sh		M_e	H_wu3	
C_m	H_k	M_zh		M_er	H_yi3	
C_n	H_ng			M_yu2	M_wu3	
C_p	H_q				M_yi3	

表 4.18：合併比率

單位	合併前			合併後	共用比率
	國語	客語	合計		
Phone	38	38	76	54	40.7%
Initial	21	21	42	28	50%
Final	40	71	111	100	11%
Syllable	411	673	1,084	1,008	7.5%

4.4 客語語言模型

4.4.1 基本客語語言模型

基本客語語言模型是針對客語文章(textA)，使用統計的方式建立而成：

$$P_{baseline}(w_i | w_{i-1}) = \begin{cases} \frac{C_{textA}(w_{i-1}, w_i)}{C_{textA}(w_{i-1})} & , C_{textA}(w_{i-1}, w_i) \geq k \\ \beta_{baseline}(w_{i-1}) P_{baseline}(w_i) & , otherwise \end{cases} \quad (4.5)$$

及

$$P_{baseline}(w_i) = \frac{C_{textA}(w_i)}{\sum_{w_i} C_{textA}(w_i)} \quad , C_{textA}(w_i) \geq k \quad (4.6)$$

$C_{textA}(\bullet)$ 表示詞串在客語文章(textA)出現次數。

4.4.2 強化客語語言模型

因為客語的文字資料量非常稀少，使用統計的方式建立的基本客語語言模型其參數量是非常稀疏的(只有 19,213 個 bi-gram)，所以希望能夠使用一些其他的資訊來訓練客語語言模型，使得語言模型更為強健。

根據現有的文字語料庫，發現除了客語文字本身外，可以用幫助建立語言模型的資訊還有客語詞類標記、國語詞類標記及國語語言模型等三種資訊。接下來將介紹如何使用這些資訊來幫助客語語言模型的建立。

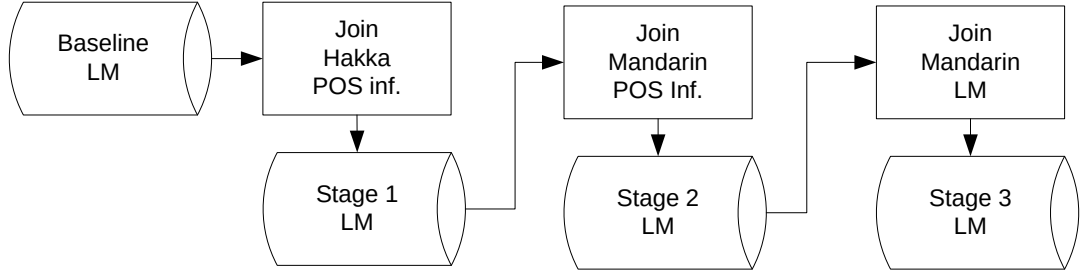


圖 4.7：強化客語語言模型流程圖

如圖 4.7 所示是採取漸進的方式逐步強化客語語言模型，依序對基本語言模型加入客語詞類資訊、國語詞類資訊及國語語言模型資訊。其架構如下式：

$$P(w_i | w_{i-1}) = \begin{cases} P_{baseline}(w_i | w_{i-1}) & , C_{textA}(w_{i-1}, w_i) \geq k \\ \alpha(w_{i-1})(\beta P_{pos}(w_i | w_{i-1}) + (1-\beta)P_{trans}(w_i | w_{i-1})) & , otherwise \end{cases} \quad (4.7)$$

和

$$P(w_i) = \begin{cases} P_{baseline}(w_i) & , C_{textA}(w_i) \geq k \\ \alpha(\beta P_{pos}(w_i) + (1-\beta)P_{trans}(w_i)) & , otherwise \end{cases} \quad (4.8)$$

其中 $\alpha(w_{i-1}) = 1 - \sum_{w_i \in C(w_{i-1}, w_i) \geq k} P_{baseline}(w_i | w_{i-1})$ ；而 $P_{pos}(w_i | w_{i-1})$ 表示經由詞類資訊所建立的

語言模型； $P_{trans}(w_i | w_{i-1})$ 為國語語言模型資訊建立的語言模型。接著，將對每個階層做詳細的介紹。

(a) 由客語詞類資訊建立語言模型(Stage1)

第一個階層將對基本客語語言模型加入客語詞類資訊，使用 class N-gram model 經由詞條和詞類的對應關係，將詞條的分數退化至詞類，以解決資料量過於稀少的問題。如式(4.9)及式(4.10)所示：

$$P_{pos}(w_i | w_{i-1}) = \sum_{G(w_i), G(w_{i-1})} P_H(w_i | G(w_i)) P_H(G(w_i) | G(w_{i-1})) P_H(G(w_{i-1}) | w_{i-1}) \quad (4.9)$$

和

$$P_{pos}(w_i) = \sum_{G(w_i)} P_H(w_i | G(w_i)) P_H(G(w_i)) \quad (4.10)$$

其中 $G(w_i)$ 表示客語詞條 w_i 的詞類標記； $P_H(G(w_i) | G(w_{i-1}))$ 和 $P_H(G(w_i))$ 表示客語詞類模型； $P_H(G(w_{i-1}) | w_{i-1})$ 和 $P_H(w_i | G(w_i))$ 表示客語詞條與詞類標記的對應模型。

接著，將詳細介紹詞類模型及客語詞條與詞類標記的對應模型。首先是詞類模型，詞類模型是根據人工標記的客語文章(textB)訓練而來，此模型記錄了詞類之間的轉移分數：

$$P_H(G(w_i) | G(w_{i-1})) = \frac{C_{textB}(G(w_{i-1}), G(w_i))}{C_{textB}(G(w_{i-1}))} \quad (4.11)$$

和

$$P_H(G(w_i)) = \frac{C_{textB}(G(w_i))}{\sum_{G(w_i)} C_{textB}(G(w_i))} \quad (4.12)$$

而客語詞條與詞類標記的對應模型則是記錄了客語詞條與詞類的相互對應分數，此分數有兩個來源，分別是由人工標記的客語文章以及詞典(dict)估算而來，其中人工標記的客語文章的客語詞條與詞類的對應較為可靠，但是並不包含所有的客語詞條；而詞典則相反，它包含了所有的客語詞條，但是其對應關係較不可靠，因為它僅記錄了客語詞條與詞類的對應而已。所以這兩個對應分數，將對上述兩個來源做結合：

$$P_H(G(w_i) | w_i) = \begin{cases} P_{textB}(G(w_i) | w_i), & C_{textB}(w_i, G(w_i)) \geq k \\ \alpha_H \frac{P_{dict}(G(w_i) | w_i) P_H(G(w_i))}{\sum_{G(w_i)} P_H(G(w_i))}, & otherwise \end{cases} \quad (4.13)$$

其中， $\alpha_H = 1 - \frac{\sum_{w_i \in C_{textB}(w_i, G(w_i)) \geq k} P_{textB}(G(w_i) | w_i)}{\sum_{w_i} P_{textB}(G(w_i) | w_i)}$ ； k 為一個常數值，當對應次數大於 k 時將

直接使用人工標記的客語文章所估算的對應分數，反之則使用詞典內的分數。而另外一個對應分數，因為客語的詞類標記僅分成 49 類，若直接估算 $P_H(w_i | G(w_i))$ ，則會發生資料量過於稀少的問題，所以此對應分數採用間接預估的方式計算：

$$P_H(w_i | G(w_i)) = \frac{P_H(G(w_i) | w_i)}{P_H(G(w_i))} P_{baseline}(w_i) \quad (4.14)$$

其中 $P_{baseline}(w_i)$ 表示基本客語語言模型的分數。

(b) 由國語詞類資訊建立語言模型(Stage2)

雖然詞類資訊在第一個階段就已經有考量，但是因為客語文字語料過於稀少，所以單純從客語估算出的詞類模型及對應模型其實並不可靠。所以第二個階段，將使用較為可靠的國語詞類資訊對第一階段的各個模型重新訓練。

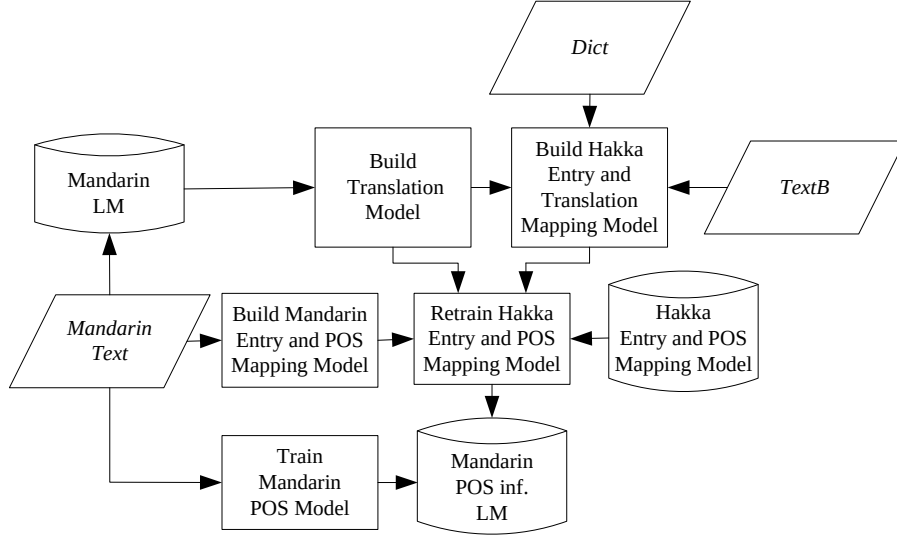


圖 4.8：國語詞類資訊語言模型的建立流程圖

圖 4.8 為國語詞類資訊語言模型的建立流程圖。主要是利用國語文章較為豐富的特性，藉由翻譯加強客語詞條對應至詞類的分數；並且將原本只由人工標記的客語文章訓練出的詞類模型取代為國語的詞類模型。以下將依序介紹各個模型的建構方式及特性：

$$P_{H,new}(G(w_i)|w_i) = \begin{cases} \sum_{M_i} P_M(G(M_i)|M_i) P_T(M(w_i)|w_i), & M(w_i) = M_i \\ P_H(G(w_i)|w_i) & , otherwise \end{cases} \quad (4.15)$$

$$P_{H,new}(w_i|G(w_i)) = \frac{P_{H,new}(G(w_i)|w_i)}{P_M(G(w_i))} P_{Trans}(w_i) \quad (4.16)$$

$$P_M(G(M_i)|G(M_{i-1})) = \frac{C_{textM}(G(M_{i-1}), G(M_i))}{C_{textM}(G(M_i))} \quad (4.17)$$

其中式(4.17)為由國語文章估建立的詞類模型，式(4.15)及式(4.16)為藉由翻譯將國語的詞類資訊對應至客語詞條上的對應模型，將這三個模型代入式(4.9)及式(4.10)中，即完成加入國語詞類資訊的動作。其中 M_i 表示國語詞條； $M(w_i)$ 表示客語詞條 w_i 的國語翻譯； $P_M(G(M_i)|M_i)$ 表示國語詞條與詞類的對應模型； $P_T(M(w_i)|w_i)$ 表示客語詞條與

翻譯的對應模型； $P_M(G(w_i))$ 表示國語詞類模型的 unigram； $P_{Trans}(w_1)$ 表示翻譯模型。

接著將對這四個模型做進一步的介紹。

首先，因為國語有豐富的文字語料，所以國語詞條與詞類的對應模型及國語詞類模型，都可以使用簡單的統計方式，就能建立不錯的模型：

$$P_M(G(M_i)) = \frac{C_{textM}(G(M_i))}{\sum_{G(M_i)} C_{textM}(G(M_i))} \quad (4.18)$$

及

$$P_M(G(M_i)|M_i) = \frac{C_{textM}(G(M_i), M_i)}{C_{textM}(M_i)} \quad (4.19)$$

接著，客語詞條與翻譯的對應模型，這個對應模型記錄了客語詞條與翻譯的對應分數，這個分數是根據人工標記的客語文章及詞典的翻譯對照估算而來的。其中人工標記的客語文章的對應和詞類標記一樣較為可靠，但是並不包含所有的客語詞條；而詞典內的詞條與翻譯對應的組合數較多，但是沒有對應次數的資訊。此分數是將這兩個來源做結合：

$$P_T(M(w_i)|w_i) = \begin{cases} P_{textB}(M(w_i)|w_i) & , C_{textB}(w_i, M(w_i)) \geq k \\ \alpha_T \frac{P_{textB}(M(w_i)|w_i) P_{Trans}(M(w_i))}{\sum_{M(w_i)} P_{Trans}(M(w_i))} & , otherwise \end{cases} \quad (4.20)$$

其中， $\alpha_T = 1 - \frac{\sum_{w_i \in C_{textB}(w_i, M(w_i)) \geq k} P_{textB}(M(w_i)|w_i)}{\sum_{w_i} P_{textB}(M(w_i)|w_i)}$ 。k 為一個常數值，當對應次數大於 k 時將

直接使用人工標記的客語文章所估算的對應分數，反之則使用詞典內的分數。

最後翻譯模型，此模型記錄了國語翻譯的轉移分數。目前根據詞典內收錄的國語翻譯與國語語言模型中詞頻最高的前 60,000 詞之對應關係，可分為三種情況，詳細統計資訊如表 4.19 所示：

表 4.19：國語翻譯與 60,000 詞分布情況

分類	詞條	
	數量	比率
Case 1	14,636	57.3%
Case 2	9,487	37.1%
Case 3	1,419	5.6%
合計	25,532	

➤ Case 1

詞典內許多客語詞條的翻譯是直接包含在國語語言模型內的，在國語文字語料中，它們是詞頻最高的前 60,000 詞。

例如：

改朝換代 更正 更衣室 求得

因為上述的這些翻譯能夠直接在國語語言模型中找到，因此它們在國語的分數也是最為可靠的，故對於這些翻譯直接複製國語語言模型的分數給它：

$$P_{Trans}(M(w_i)) = P_M(M(w_i)) \quad (4.21)$$

➤ Case 2

詞典內許多的翻譯，因為在國語文字語料中的詞頻較低，故未包含在前 60,000 詞中，所以無法直接在語言模型中找到。但是，通常經過斷詞以後，拆解後的詞條都能在前 60,000 詞內。此部分的翻譯通常為較長的國語詞條或諺語，

例如：

一人飽全家飽 一年之計在於春 人心節節高

或是因為客語詞條的詞意為多個國語詞條的詞意所組成，

例如：

客語詞條『其』的國語翻譯為『他的』是由國語詞『他』和『的』組成

上述兩種情況都會造成翻譯為由多個國語詞組成。此部分翻譯模型的分數做法為使用國語前 60,000 詞對翻譯做斷詞後，再給予翻譯模型分數：

$$\begin{aligned} P_{Trans}(M(w_i)) &= P_M(M(w_i) = M_1, M_2, \dots, M_n) \\ &= P_M(M_1) P_M(M_2 | M_1) \dots P_M(M_n | M_{n-1}) \end{aligned} \quad (4.22)$$

其中 n 表示國語翻譯由 n 個國語詞條組成。

➤ Case 3

還有一些詞典內的翻譯，即使拆解成一字詞後，依然無法由國語的前 60,000 詞組成，部分是因為這些翻譯含有造字，

例如：

『𩺰』魚

或是含有一些少見的國語一字詞，

例如：

『𩺰』腳貨 『𩺰』鼠 『鷓鴣』菜

上述兩種情況將會造成翻譯無法直接由國語前 60,000 詞估計，但是這些詞條在國語文字語料中亦是少見詞條，即使由國語文字中取得分數亦不夠可靠，所以目前不計算此類的翻譯模型分數。

(c) 由國語語言模型資訊建立語言模型(Stage3)

這個階層主要目的是藉由詞典翻譯直接使用國語語言模型中的詞條資訊。此階層將針對表 4.19 中前後客語詞條的翻譯皆為 Case 1 的情況，做複製國語詞條轉移分數的動作：

$$P_{Trans}(w_i | w_{i-1}) = \sum_{M(w_i), M(w_{i-1})} P_T(w_i | M(w_i)) P_M(M(w_i) | M(w_{i-1})) P_T(M(w_{i-1}) | w_{i-1}) \quad (4.23)$$

其中 $P_T(w_i | M(w_i))$ 及 $P_T(M(w_{i-1}) | w_{i-1})$ 為客語詞條與翻譯的對應模型；

$P_M(M(w_i) | M(w_{i-1}))$ 為國語語言模型，前後翻譯均為前 60,000 詞。

表 4.20：Stage 3 客語語言模型混淆度估測

語言模型權重參數 (pos, trans)	混淆度評估 (perplexity)	
	訓練語料	測試語料
(1, 0)	979	1,320
(0.95, 0.05)	1,229	1,725
(0.9, 0.1)	1,238	1,730
(0.8, 0.2)	1,230	1,756
(0.7, 0.3)	1,251	1,796

4.5 雙語語言模型

此節是針對日常生活中國客語夾雜使用情況而設計語言模型，目前僅考量以客語為背景語言的語音辨認。此類型的語者所使用到的國語詞彙通常是專有名詞或較為新穎的詞彙，穿插在客語文句之中。

例如：

頭擺人毋知麼 安到《網路遊戲》
《摩托車》要買較大台 正會較穩

其中《》內的文句為國語。

因為，目前並無大量此類型的雙語文字，所以本研究是針對實驗所用的雙語語音語料中國語語音和客語語音所佔的比例及出現的國語詞彙做語言模型的設計。

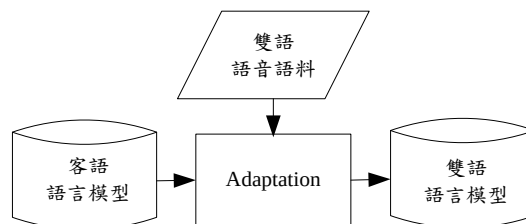


圖 4.9：雙語語言模型建置流程圖

圖 4.9 為雙語語言模型建置流程圖，用雙語語音語料中國客語所占的比例，給予語言轉換及國語詞條的分數。其中分數如式(4.24)及式(4.25)所示：

$$P(M_i) = \frac{k}{C_{bilingual}(M)} \quad , \quad M_i \in bilingual \quad (4.24)$$

其中 $C_{bilingual}(M)$ 表示在雙語語料中國語詞條數目。 k 為國語詞條在雙語語料中所占的比例。而考慮語言之間相互轉換時，詞條的分數為：

$$P(B_i | B_{i-1}) = \begin{cases} P(w_i | w_{i-1}) = (1-k)P(w_i | w_{i-1}) \\ P(M_i | w_{i-1}) = \frac{k}{C_{bilingual}(M)} \\ P(M_i | M_{i-1}) = \frac{k}{C_{bilingual}(M)} \\ P(w_i | M_{i-1}) = (1-k)P(w_i) \end{cases} \quad (4.25)$$

其中 B_i 表示雙語詞條，可能為國語詞條 M_i 或客語詞條 w_i ； $P(w_i | w_{i-1})$ 和 $P(w_i)$ 為 4.4 節所建立的客語語言模型。

將以上各節定義之語言模型，且包含參數調整的部分；各個語言模型對不同語料之混淆度比較，如表 4.21 所示。

表 4.21：語言模型混淆度估測

語言模型		Bigram 數	Perplexity	
			訓練語料	測試語料
客語語言模型	Baseline	19,213	1,149	1,559
	Stage 1	7,712,200	1,013	1,366
	Stage 2	12,003,727	979	1,320
	Stage 3	20,762,457	1,229	1,725
雙語語言模型	Baseline	16,312,103	-	885
	Stage 1	24,008,914		892
	Stage 2	28,300,441		882
	Stage 3	37,075,131		853

根據表 4.21 可觀察幾項資訊：

- (1) 客語語言模型：相對於 baseline 而言，Stage 1 及 Stage 2 對於客語語音辨認，都能降低語言模型對目標語料的混淆度。但是，Stage 3 卻提升了對目標語料的混淆度，是否會影響辨認率則有待觀察。

(2) 雙語語言模型：Stage1 至 Stage3 對於目標語料的混淆度，依照資訊的增量增加而降低。

(3) Stage2 與 Stage3 的差異：對客語的測試語料而言，其混淆度 Stage3 明顯高於 Stage2；但是，在雙語的測試語料，其混淆度卻是相反的。此部份辨認率有待觀察。

4.6 實驗結果及討論

此節將介紹本研究所有的實驗設定及結果，並加以分析及討論。

4.6.1 聲學模型實驗

一般來說聲學模型之效能是直接由辨認率來評估，故本研究將對所建立之雙語混合聲學模型(Mix)、國語聲學模型(M)、客語聲學模型(H)及國客語直接合併之雙語聲學模型(H+M)等四個聲學模型做辨認率的比較，表 4.22 為各個模型的 GMM 數量。

表 4.22：模型 mixture 數量

Model	Unit	Mixture
H	RCD-I/F	19,813
H	Tri-ph.	19,784
M	Tri-ph.	52,807
Mix	Tri-ph.	52,627
H+M	Tri-ph.	72,591

接著，依照測試語料的不同做以下三種實驗：分別是客語的音節辨認率比較、國語的音節辨認率比較及雙語之音節辨認率比較。

➤ 客語音節辨認率之比較

表 4.23：不同聲學模型、待辨認音節對於客語測試語料之音節辨認率比較

Model	Unit	Syllable	Text	Cor.	Acc.	Del.	Sub.	Ins.	Total
H	RCD-I/F	H	H	51.36	50.18	598	8,789	228	19,299
H	Tri-ph.	H	H	53.52	52.31	557	8,413	234	19,299
Mix	Tri-ph.	H	H	52.16	50.87	580	8,652	250	19,299
H+M	Tri-ph.	H+M	H	49.99	48.57	541	9,111	274	19,299
Mix	Tri-ph.	Mix	H	43.43	42.20	592	10,326	237	19,299

由表 4.23 可以看出雙語混合聲學模型(Mix)與客語聲學模型(H)之音節辨認率比較時，雙語混合聲學模型之音節辨認率介於客語 RCD-I/F 聲學模型和客語 Tri-phone 聲學模型之間，所以雙語混合聲學模型對於客語音節的辨認率算是合理。

但是，當待辨認音節由客語音節替換成雙語混合音節(待辨認音節 1,008 個)時，與直接合併之雙語語言模型(H+M)的音節辨認率有明顯差距。針對這個現象，尋找替換待辨認音節後辨認率下降的原因(由 50.87 降至 42.20)，發現當替換成雙語混合音節時，許多客語音節容易辨認為相似的國語音節(詳細錯誤分析如表 5.3)，而直接合併之雙語聲學模型的客語音節辨認為國語相似音節的情況卻明顯比較少，推測造成此現象的原因可能是因為雙語混合聲學模型中的一些雙語共用的 HMM 模型其訓練語料來源包含了國語語料及客語語料，而直接合併之雙語聲學模型則沒有共用的 HMM 模型。

表 4.24：Substitution 明顯增加且常出現(>40 次)，前十名之音節

排名	Substitution(%)增加	正確客語音節	最大辨認錯誤音節		出現次數
			音節	錯誤次數	
1	31.5	H_zien	M_jian	10	54
2	29.8	H_me	M_mei	13	47
3	29.4	H_mien	M_mian	15	51
4	29.3	H_so	M_sou	10	41
5	28.4	H_zo	M_zou	8	67
6	28.4	H_zung	M_zong	19	116
7	28.1	H_do	M_dou	99	541
8	26.9	H_tung	M_tong	34	104
9	26.7	H_sii	M_si	47	265
10	26.5	H_gau	M_gao	20	98

表 4.24 為置換性錯誤明顯增加的客語音節，其最大錯誤的對象皆為國語相似音節，並且它們有共用的 HMM 模型。

➤ 國語音節辨認率之比較

接著，替換國語測試語料，表 4.25 為其詳細辨認率統計。由表可看出雙語混合聲學模型與國語聲學模型之音節辨認率比較時，其音節辨認率差距並不明顯，屬於合理的下降範圍。並且，當雙語混合聲學模型將待辨認音節由國語音節替換成雙語混合音節時，並沒有發生像客語如此明顯的辨認率下降，推測原因可能是因為國語的訓練語料較客語的訓練語料來的豐富，造成雙語共用的 HMM 模型比較偏向於國語特性，因此即使替換待辨認音節也不會發生像客語如此嚴重的搶音節問題。

表 4.25：不同聲學模型、待辨認音節對於國語測試語料之音節辨認率比較

Model	Unit	Syllable	Text	Cor.	Acc.	Del.	Sub.	Ins.	Total
M	Tri-ph.	M	M	71.28	70.63	416	8,553	201	31,252
Mix	Tri-ph.	M	M	70.51	69.96	478	8,739	171	31,252
H+M	Tri-ph.	H+M	M	71.20	70.56	420	8,580	202	31,252
Mix	Tri-ph.	Mix	M	69.75	69.19	476	8,978	175	31,252

➤ 雙語音節辨認率之比較

最後，對雙語測試語料做音節辨認率的比較，此語料完全是測試語料，使用此語料做音節辨認率的比較將會更為客觀。此部份主要是對雙語混合聲學模型與直接合併之雙語聲學模型的音節辨認率之比較，詳細辨認率如表 4.26 所示

表 4.26：不同聲學模型對於雙語測試語料之音節辨認率比較

Model	Unit	Syllable	Text	Cor.	Acc.	Del.	Sub.	Ins.	Total
H+M	Tri-ph.	H+M	Mix	30.60	28.76	239	6,296	174	9,417
Mix	Tri-ph.	Mix	Mix	37.58	36.75	331	5,547	78	9,417

由表 4.26 中的辨認率可發現當使用雙語語料做辨認時，雙語混合聲學模型的辨認率明顯優於直接合併之聲學模型，這也是本研究所期望的結果。推測造成此原因應該是直接合併之聲學模型對於兩語言中相似音節置換性錯誤問題十分嚴重。

4.6.2 客語語言模型實驗

此章節將特別針對客語語言模型做辨認率的比較，以評估語言模型的效能。目的在於比較各個階層的所考量的資訊對於客語語音辨認的幫助。

由於實驗要比較的是客語語言模型，所以此實驗之聲學模型設定為一個模型數量較少的客語聲學模型來進行客語語言模型實驗，以加快實驗速度。

選用的客語聲學模型為音節內右相關聲/韻母模型(Right-context-dependent Initial/Final Model, RCD)，每一個聲母之 HMM 模型採用 3 個由左至右(left-to-right)的狀態(state)表示，而韻母之 HMM 模型則採用 5 個狀態來表示。其中每一個狀態以平均 32 個高斯分布之高斯混合模型(Gaussian Mixture Model, GMM)描述其特徵參數之分布，聲學模型之 HMM 設定及音節辨認率如表 4.27 及表 4.28 所示。

表 4.27：HMM 模型之設定

HMM 模型類別	狀態個數	模型數量
Right-context-dependent Initial	3	87
Final	5	71
Silence	3	1
Short pause	1	1

表 4.28：RCD 聲學模型音節辨認率

Correct	Accuracy	Deletion	Substitution	Insertion	Total
51.38	50.19	601	8,801	230	19,338

接著，將使用上述的聲學模型對各階層的客語語言模型計算辨認率，第一個階層為加入客語詞類資訊、第二個階層為加入國語詞類資訊、第三個階層為加入國語語言模型資訊。詞、字元及音節辨認率分別如表 4.29、表 4.30 及 4.31 所示。

表 4.29：客語語言模型詞(Word)辨認率

LM	Correct	Accuracy	Deletion	Substitution	Insertion	Total
Free grammar	23.31	21.44	1,770	6,934	212	11,350
Baseline	42.05	40.79	2,026	4,551	143	11,350
Stage 1	45.81	42.99	1,436	4,715	320	11,350
Stage 2	46.99	44.45	1,530	4,487	288	11,350
Stage 3	47.28	44.56	1,476	4,508	309	11,350

表 4.30：客語語言模型字(Character)辨認率

LM	Correct	Accuracy	Deletion	Substitution	Insertion	Total
Free grammar	45.43	42.67	316	10,236	534	19,338
Baseline	60.15	58.99	475	7,232	223	19,338
Stage 1	62.11	60.79	401	6,926	255	19,338
Stage 2	63.26	62.17	546	6,559	211	19,338
Stage 3	63.28	62.12	513	6,588	225	19,338

表 4.31：客語語言模型音節(Syllable)辨認率

LM	Correct	Accuracy	Deletion	Substitution	Insertion	Total
Free grammar	60.88	57.97	344	7,221	562	19,338
Baseline	67.58	66.29	501	5,768	249	19,338
Stage 1	70.05	68.63	421	5,371	275	19,338
Stage 2	70.26	69.07	565	5,187	230	19,338
Stage 3	70.44	69.18	532	5,184	244	19,338

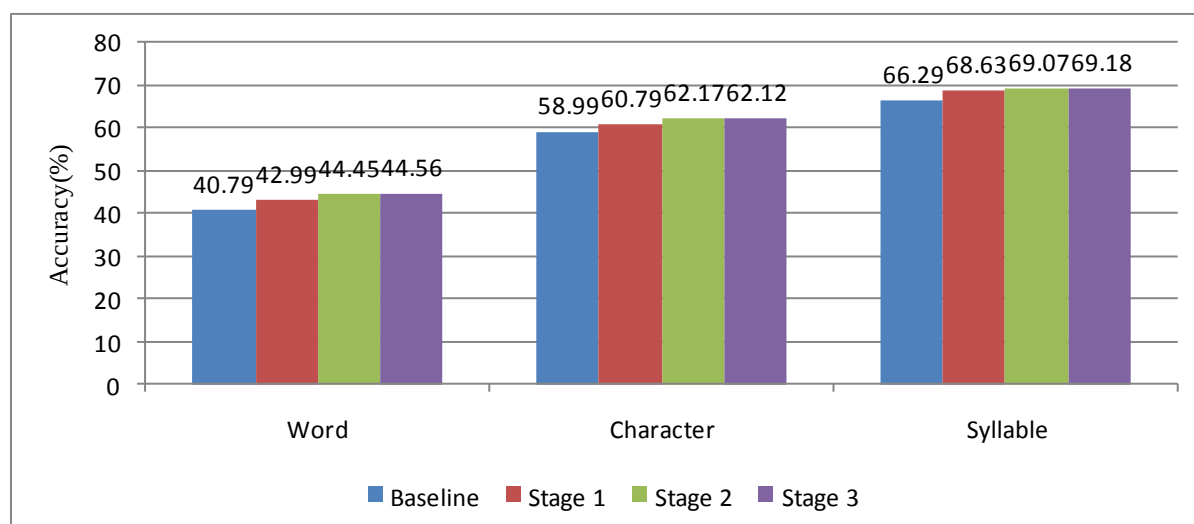


圖 4.10：客語語言模型辨認率

由圖 4.10 客語語言模型辨認率，可以看出客語基本語言模型再加入了客語詞類資訊後，不管是詞、字元還是音節辨認率，均明顯的提升，其中詞辨認率提升了 2.2%。而再加入國語詞類資訊後詞的辨認率較 Stage1LM 提升了 1.46%。對於客語辨認而言，不管是使用客語或國語的詞類資訊，對於最後的辨認結果都有明顯的改善，如此也證明了國語與客語在文法結構上確實十分相似。

接著分析詞類資訊對於哪些詞條的辨認率有顯著幫助，所以將比較各個詞條從 baseline 語言模型至 Stage 1 及 Stage 2 語言模型的辨認率，普遍詞條的辨識率皆有上升，圖 4.11 為明顯提升的詞條。

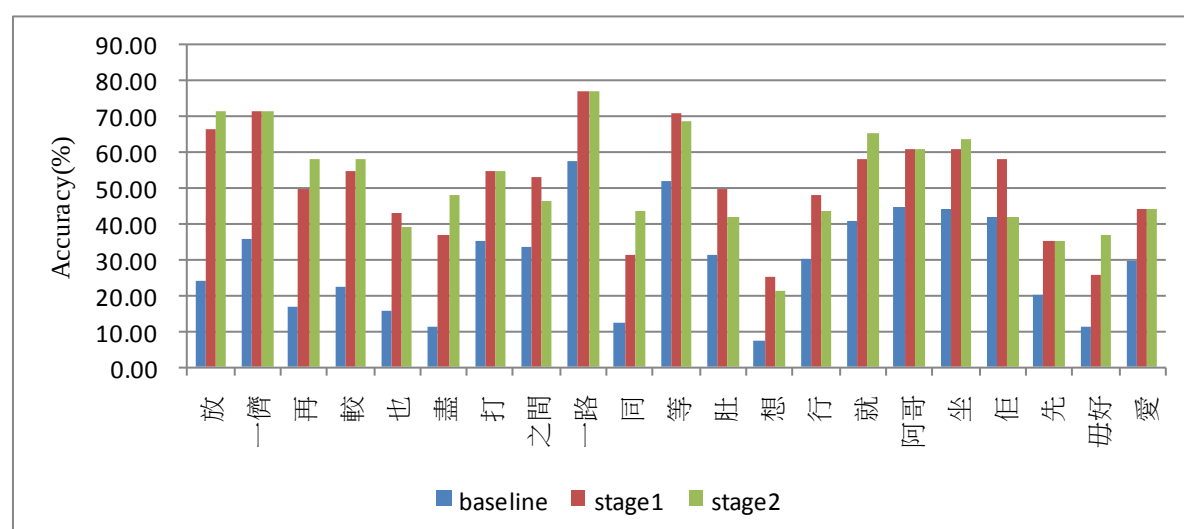


圖 4.11：詞類標記明顯幫助的詞條之 TOP20(詞條次數大於 15 次)

以詞性為單位，去觀察詞條的辨識率提升狀況：

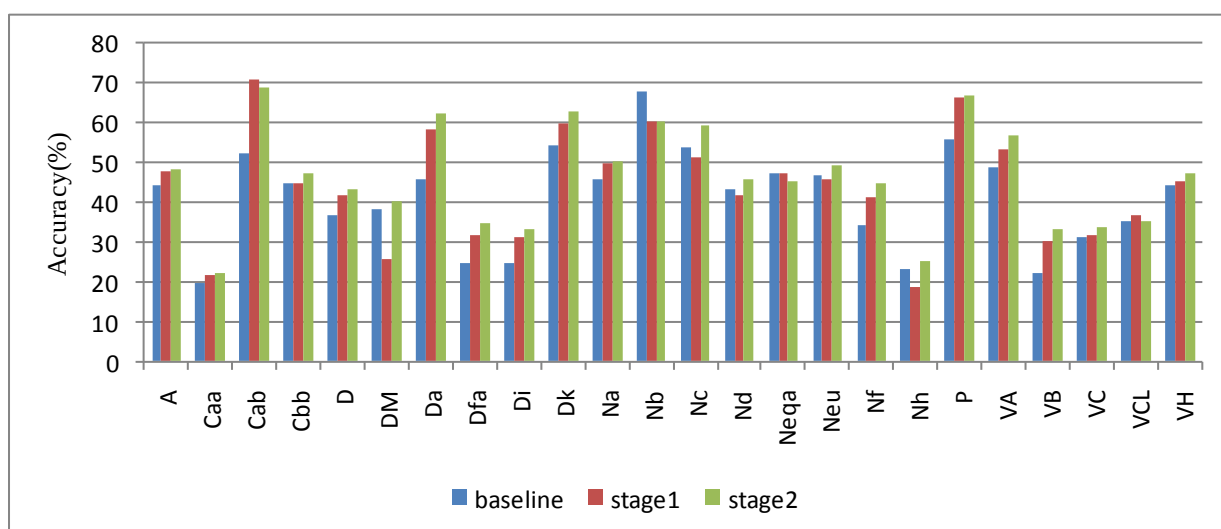


圖 4.12：詞類標記對於幫助辨認的詞類(詞類次數大於 40 次)

由圖 4.12 可看出詞類標記普遍對於各種詞類皆有幫助，加入國語詞類資訊後也有所提升。各種詞類中，以連結詞(C)及副詞(D)的幫助最為顯著，猜想是此類的詞條大部分皆為短詞，所以當加入詞類標記資訊時，辨認率提升會較為明顯。可是專有名詞(Nb)的辨認率卻下降了，推測原因可能是加入 POS 資訊時，使得專有名詞變得更為混淆，導致其辨認率下降。

此外，再加入國語語言模型資訊(Stage 3)時，對於 Stage2LM 而言，詞的辨認率只提升了 0.11%，幾乎是沒有改善。目前推測可能有下列原因：一、直接使用國語語言模型的 bigram 很有可能是『中文式』客語。二、客語詞條與翻譯的對應數量很少，其關係其實並不可靠。三、客語重要的詞彙可能沒有國語翻譯的對應。四、國語文章為廣泛性的文章，而客語為故事文章。可能基於上述的原因造成直接使用國語語言模型的資訊對於辨認率沒有明顯改善。

4.6.3 雙語語音辨認實驗

雙語的語音辨認實驗，目的除了測試本研究的整個系統外，也將特別對前述章節的客語言模型、聲學模型做交叉比對的實驗。因為 4.6.2 節客語語言模型實驗提到，加入國語語言模型資訊(Stage3)對於 Stage2 而言提升效果並不明顯。所以此部分的雙語語言模型設計是分別對 Stage2 及 Stage3 的客語語言模型，加入國語詞及語言轉換的分數。

此實驗所使用的雙語測試語料內容為日常生活中的國客語夾雜文句，是以客語為主的文句，國語與客語的比例約為 1:3(詞條數 738: 2206)。根據訓練出的雙語語言模型(Stage2 及 Stage3)，分別對雙語混合聲學模型(Mix)及直接合併的聲學模型(H+M)做辨認率的比較及分析。詳細辨認率如表 4.32 所示。由表中的雙語辨認率可明顯看出，MIX 的聲學模型辨認率明顯比 H+M 的聲學模型來的佳。接著，將針對詞的辨認率觀察其語言的辨認狀況，如表 4.33 所示：

表 4.32：雙語辨認率之比較

AM	LM	Unit	Cor.	Acc.	Del.	Sub.	Ins.	Total
Mix	None	Syllable	37.58	36.75	331	5,547	78	9,417
	Bilingual (Stage 2)	Word	40.15	37.16	349	1,413	88	2,944
		Character	56.03	54.33	137	1,934	80	4,710
		Syllable	57.98	56.24	140	1,839	82	4,710
	Bilingual (Stage 3)	Word	41.58	37.84	303	1,417	110	2,944
		Character	57.15	55.27	126	1,892	89	4,710
		Syllable	58.13	56.14	131	1,841	94	4,710
	None	Syllable	30.60	28.76	239	6,296	174	9,417
	Bilingual (Stage 2)	Word	25.44	21.33	463	1,732	121	2,944
		Character	40.08	34.86	134	2,688	246	4,710
		Syllable	43.42	38.09	140	2,525	251	4,710
H+M	Bilingual (Stage 3)	Word	26.12	21.98	443	1,732	122	2,944
		Character	40.83	35.61	142	2,645	246	4,710
		Syllable	43.84	38.47	150	2,495	253	4,710

表 4.33：雙語語音辨認之詞辨認率分析

AM	LM	比率%(個數)				總數	
		客語正確率	國語正確率	客 -> 國	國 -> 客	客語	國語
H+M	Bilingual (Stage 2)	28.56 (630)	16.12 (119)	1.04 (23)	72.49 (535)	2,206	738
	Bilingual (Stage 3)	29.10 (642)	17.21 (127)	1.04 (23)	71.82 (530)	2,206	738
Mix	Bilingual (Stage 2)	34.09 (752)	58.27 (430)	2.22 (49)	32.25 (238)	2,206	738
	Bilingual (Stage 3)	35.86 (791)	58.67 (433)	1.99 (44)	33.06 (244)	2,206	738

由上表，可觀察出 Stage 3 語言模型對於 Stage 2 而言，詞辨認率相對提高。且 Mix 聲學模型對於國語的辨認有極大的幫助，推測因為欲辨認之文句皆是以客語為主文句，所以造成 H+M 聲學模型辨認國語詞條容易辨認成客語詞條。接著，觀察其音節辨認的辨認狀況：

表 4.34：雙語語音辨認之音節辨認率分析

AM	LM	比率%(個數)				總數	
		客語正確率	國語正確率	客 → 國	國 → 客	客語	國語
H+M	Bilingual (Stage 2)	55.15 (1,782)	17.78 (263)	0.80 (26)	79.31 (1,173)	3,231	1,479
	Bilingual (Stage 3)	55.15 (1,782)	19.13 (283)	0.84 (27)	77.76 (1,150)	3,231	1,479
Mix	Bilingual (Stage 2)	55.15 (1,782)	64.16 (949)	1.80 (58)	32.86 (486)	3,231	1,479
	Bilingual (Stage 3)	55.15 (1,782)	64.64 (956)	1.86 (60)	32.45 (480)	3,231	1,479

由表 4.34 可看出 MIX 聲學模型明顯提升國語音節的辨認率，並且大幅降低國語音節辨認為客語音節的機率。此外有個現象，H+M 聲學模型與 Mix 聲學模型的音節辨認率，對於客語而言是一樣的，但是詞的辨認率卻有明顯差異，推測是因為 H+M 模型容易將國語音節辨認為客語音節，造成音節構成客語詞條時，會包含原本應是國語的音節，進而造成詞條的辨認率有如此差異。

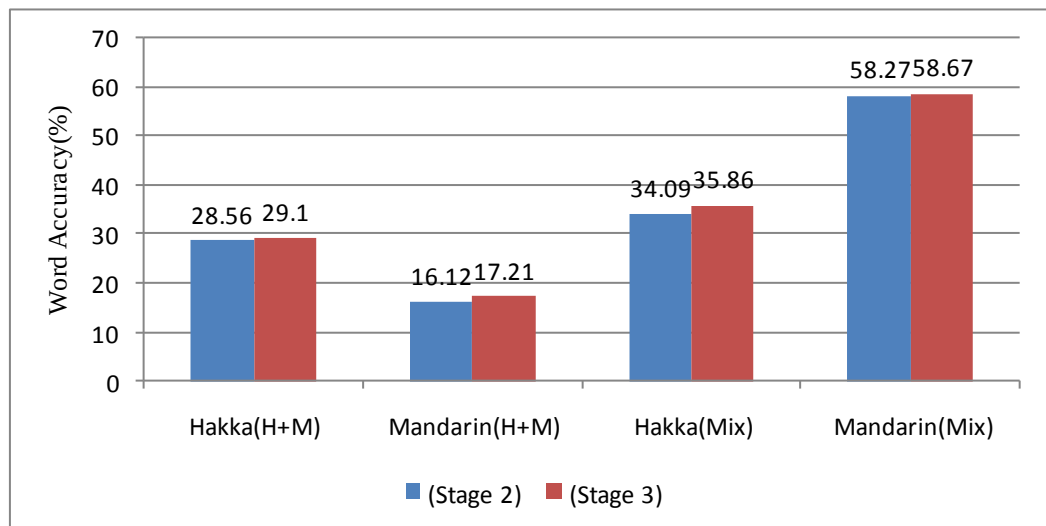


圖 4.13：不同客語語言模型對於雙語語音辨認之辨認率比較

由圖 4.13 可看出，Stage3 的語言模型對於雙語語音的辨認有明顯的改善，不管是在 Mix 聲學模型或 H+M 聲學模型上，國語及客語詞條的辨認率都有所提升。可能是因為國語語言模型是屬於廣泛性文章，所以對於這個雙語語料比對客語語料來的符合，造成

在雙語語音辨識的時候，Stage3 語言模型效能明顯比 Stage2 語言模型佳。接著，對聲學模型的比較，如圖 4.14 所示：

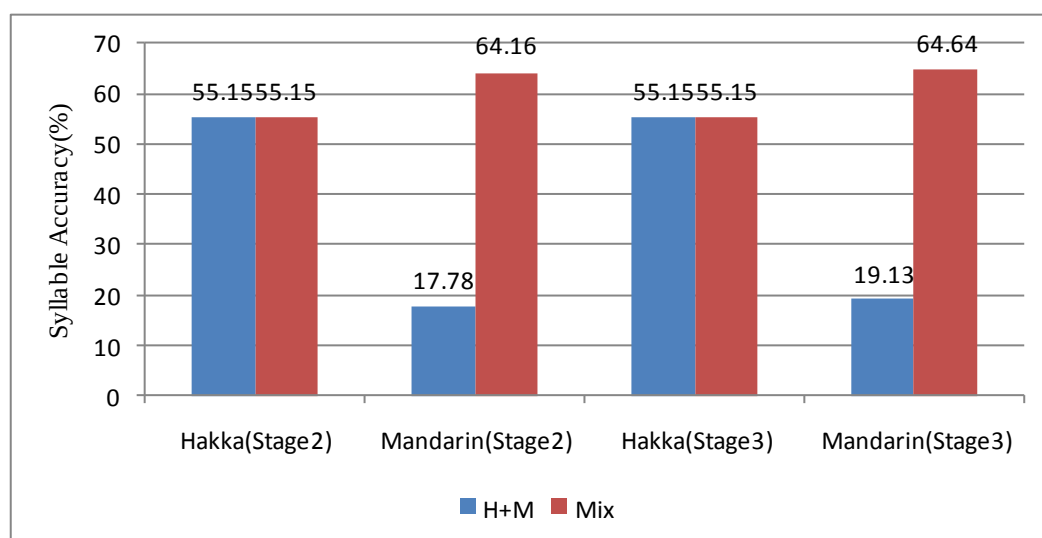


圖 4.14：不同聲學模型對於雙語語音辨認之辨認率比較

對於 H+M 聲學模型與 Mix 聲學模型而言，不管是使用 Stage2 語言模型或 Stage3 語言模型，對於國語音節的辨認率都有大幅提升，但是客語音節的辨認率沒有變化。Mix 聲學模型對於幫助非主語的詞彙，有顯著的幫助。

五、音韻學研究 - 客語「啊」時貌之間的關係探討

本研究的主要目的是解釋當客語的「啊」作為時貌標記時，其核心語意為表淺程度、短時之意，從這個語意出發經過核心語的隱喻轉換、和語意延伸，可貫穿「啊」所扮演的四個客語時貌標記（暫時貌、嘗試貌、反覆貌、以及緊密先前貌）之間的語意關連性。並將這些不同時貌對應之語音聲調表現關係建立起來，外來可用於語音合成之韻律產生規則使用。

本研究之研究法為與料庫語言學之取向，我們從收集之語音與文字語料庫，先從不同的「啊」句法功能來檢視啊的聲調表現——如「啊³¹」可作為句首感嘆詞、「啊²⁴」可作為句尾表達語者強化主觀判斷的語氣詞、「啊²¹」句中可作為語流中的停頓，讓語者有機會計畫自己接下來的語意內容，最後，「啊⁵⁵」在「V啊V結構」中，有強調某動作或狀態的功能。之後，我們將討論逐漸收斂到「啊」所扮演的四種時貌標記功能，同樣的透過啊的聲調變化及出現語境來決定啊的本調為高平調55，由此再進一步從語料例句中找出其核心語意為「一下」之表短時或淺程度之暫時貌語意（delimitative aspect），而其句法結構為：[動詞_i（動詞_i）啊⁵⁵哩²¹/啊³¹]（如我著一下衫啊⁵⁵哩²¹/啊³¹）；其次，從表嘗試貌（tentative）的例句中，我們發現其句型結構與暫時貌結構類同，但動詞必須是表嘗試語意的「試」或「看」，該句型的完整結構為[試-動詞-看-啊⁵⁵-哩²¹/啊³¹]（如你去試食看-啊⁵⁵-哩²¹/啊³¹），其中「試」或「看」，可同時出現或任選一位出現但表短時、淺程度的「啊⁵⁵-哩²¹/啊³¹」則一定要出現。第三客語的反覆貌結構「V啊V仔」也有一個啊（如細人仔蓄啊蓄仔，無多知仔，恁大欸），前人文獻將「仔」寫成客語完成時貌的「欸」或「哩」，但經過我們檢視「仔」的聲調變化，發現「仔」的聲調是隨前面的動詞的聲調來變化自己的聲調，因此是個典型的小稱詞，於是我們將反覆貌結構中的「仔」以小稱詞而非完成時貌來表示，如此，表短時、淺程度的「啊」正好可以和結構最後的小稱「仔」淺後呼應，整個構式的語意表現出「做一點、做一點」的反覆語意，因此可以表現出客語特有的反覆時貌；最後，客語表達「剛剛」之意的緊密先前貌（immediate anterior）構式中也有啊字的出現，其構式為[動詞-啊⁵⁵-補語]（如佢正行啊入），我們論證：（1）此時原本表短時「一下、一點」的啊、透過隱喻的轉換，將語意轉成之某事件發生時間與說話時間之間的距離很短，也就是國語的「剛剛」之意，但在兩個相關事件中出現時，則將「剛剛」之意延伸為指該兩事件之間前後發生的時間距離很近，因此衍生出「一。。。就。。。之語意。

總結來說，客語「啊」出現的頻率很高，可扮演許多的句法功能，但不同的句法功能會決定啊的聲調表現，由於客語是聲調語言，聲調的定奪對詞彙語意有辨意的功能，同時也能決定語者說話的語氣、和表現的態度，一旦我們確定啊的時貌義的本調及其語意，就能開始進行含有「啊」的各種時貌結構之間的語意和句法關係。這種關係的探討對語音合成和辨認來說都是十分重要的。

六、結論與未來發展

6.1 結論

在客語文字轉語音系統部分，我們實現了單機版的四縣客語文字轉語音系統，其系統包含分詞與詞性標記單元、停頓預估單元以及基於HMM之語音合成器。在整個實作過程中，我們得到了以下結論：

1. 客語語料的取得是非常不容易的，因此我們在客語斷詞系統部分無法訓練客語專屬分詞及詞性標記條件隨機域模型，本研究利用國語和客語有相同語言特性的特點，在中文斷詞系統中加入客語外掛詞典以及客語規則，完成一套客語斷詞系統。
2. 本研究在TTS系統中加入了預估詞後是否有停頓的單元，停頓預估的加入能讓合成語音在韻律表現上更加自然。
3. 由基於HMM之語音合成器所合成的客語語音在客觀及主觀評估上都有不錯的表現，這對於客語TTS系統應用在本土語言學習中是一大幫助。

在建立客語和國客夾雜語音辨認系統方面，使用國語及客語單語言的語音及文字語料庫，訓練一個可做為國語、客語及國客語混合語音辨認的系統。也改善了原本客語的單語言辨認系統，使用了客語詞類資訊、國語辭類資訊及國語語言模型資訊來強化原本客語語言模型。根據各項實驗數據及研究過程，針對客語語音辨認系統及雙語語音辨認系統歸納出以下幾項重點與成果：

- 客語語言模型：取得大量的客語文章其實是非常困難的，所以首要目標應是增加國語與客語的對譯詞典，使用本計畫所提出的方法，利用國語語言模型的資訊幫助客語語言模型的建立。
 - 雙語音素集的定義：本計畫提出不同語言音素合併的規則，根據定義的音素結構位置、相同的漢語音標對應及同種語言音素不合併等。相信對於雙語辨認會有不錯的效果，且經由辨認可以發現音節辨認率明顯提升，尤其是對於非主要語語言。
 - 雙語語言模型：目前對於非主要語言的國語詞彙，僅使用一個詞群做分類，給予相同的轉移機率。相信這方面可以更改為使用詞類標記資訊，做為語言轉換時的分數，因為由實驗可數據可看出國語詞彙與客語詞彙的詞類標記是極為相似的。
- 漢語語系語言辨認系統：相信本研究所提出的方法，可以再應用於其他腔調的客家話及台語，因為其文法結構及漢字特性都與國語十分相像。

6.2 未來展望

有以下幾點可以提供未來繼續深入探討並改進的方向：

1. 雖然目前在客語詞典的收錄已有一定的數量，但是跟中文詞典和文章數比起來還是非常少，我們希望能在客語詞典中收錄中文對照翻譯之資訊，進而從斷詞系統中訓

練出國客語之雙語斷詞系統。

2. 國客語之雙語合成、或是輸入中文字而系統合成客語語音，這都是未來的趨勢以及我們所要努力的方向。
3. 目前客語本身的辨識率和國語比較起來還是明顯偏低，必須再進一步深入探究其原因，但為了提升其效能，接下來我們還是得解決缺乏客語文字語料的問題，增加 language modeling 需要的客語文章數量，使 LM 更加強健。
4. 使用國客語對譯資訊來增加客語 LM 之強健性，除了本研究所使用的 syntax 資訊外，在下階段應再加入 semantic 及 morphology 的資訊輔助，再使辨認率提升。

七、 參考文獻

- [1] 陳運棟，1989，《台灣的客家人》，台北：臺原出版社。
- [2] 鍾榮富，2004，《台灣客家語音導論》，台北：五男圖書出版有限公司。
- [3] 羅肇錦，1990，《台灣的客家話》，台北：臺原出版社。
- [4] 江敏華，1998，《台中縣東勢鎮客語語音韻研究》，台大國文研究所碩士論文。
- [5] Yu, Hsiu-Min(余秀敏). 1984. *Aspects of the Phonology of Miaoli Hakka* (苗栗客家話音韻研究). Master Thesis. Fu Jen Catholic University.
- [6] Chung, Raung-Fu. 1992. "The Domain of Hakka tone sandhi." *Studies in Language Teaching, Linguistics, and Literature*. 113-114. Department of English, National Kaohsiung Normal University.
- [7] Hsu, Kuei Ping, and Yuchao Hsiao (徐桂平、蕭宇超). 1994. "Hakka Yinging Tone Sandhi from a Syntactic Perspective (從句法結構看苗栗絲線客家話的陰平變調)." Paper for *The Third International Symposium on Chinese Phonology and the Twelfth Annual Conference of the Association of Chinese Phonology*. Institute of Linguistics, National Tsing Hua University.
- [8] Hsu, Kuei Ping (徐桂平). 1996. *Hakka Tone Sandhi: The Interface between Syntax and Phonology*. M.A. Thesis. National Tsing Hua University.
- [9] Chen, Matthew Y. 2000. *Tone Sandhi: Patterns across Chinese Dialects*. Cambridge: Cambridge University Press.
- [10] Hsiao, E. Yuchao. 2000. Trisyllabic and Quadrasyllabic Hakka Tone Sandhi: An Optimality Theory Perspective. *Proceedings of the NCCU Teacher's Conference on Linguistics Research*. National Chengchi University.
- [11] Chen, Matthew, Yan Xiuhong and L.H. Wee. 2004. "Hakka Tone sandhi –Corpus and Analytical Challenges." *Journal of Chinese Linguistics Monograph* 21.
- [12] Chen, Matthew Y. 2004. "Changting Hakka Tone Sandhi: Analytical Changes." In *Language and Linguistics* 5.4: 799-820.
- [13] Hsiao, Yuchao and Yunyi Chiu(蕭宇超，邱昀儀). 2006. "Trisyllabic and Tetrasyllabic Tone Sandhi in Dongshi Hakka (東勢客語變調：三至四字組分析)." In *Language and Linguistics* 7.2: 455-482.

- [14] D. H. Klatt, "Review of Text-to-Speech Conversion for English," *J. Acoust. Soc. Am.*, vol. 82, no. 3, pp. 137-181, Sept. 1987.
- [15] R. Sproat, *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, Kluwer Academic Publishers, 1998.
- [16] Fu-Chiang Chou, Chiu-Yu Tseng, Lin-Shan Lee, "A Set of Corpus-Based Text-to-Speech Synthesis Technologies for Mandarin Chinese", *IEEE Transactions on Speech and Audio Processing*, Volume 10, Issue 7, Oct 2002, pp. 481- 494.
- [17] S. H. Chen, S. H. Hwang, and Y. R. Wang, "A Mandarin Text-to-Speech System," *Computational Linguistics and Chinese Language Processing*, Vol. 1, No. 1, pp. 87-100, August, 1996. [18] 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.
- [18] Min Chu, Hu Peng and Eric Chang, "A Concatenative Mandarin TTS System without Prosody Model and Prosody Modification," in *Proc. of 4th ISCA Workshop on Speech Synthesis*, Scotland 2001
- [19] Min Chu, Hu Peng, Yong Zhao, Zhengyu Niu and Eric Wang, "Microsoft Mulan –A Bilingual TTS System," *ICASSP 2003*,
- [20] Min Chu, Hu Peng, Hong-yun Yang, Eric Chang, "Selecting non-uniform units from a very large corpus for concatenative speech synthesizer", *ICASSP 2001*, Vol. 2, SPEECH-L2.2, 2001.
- [21] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, *Spoken Language Processing*, Prentice-Hall, Inc.
- [22] Tseng, Chiu-yu, Pin, Shao-huang, Lee, Yeh-lin, Wang, Hsin-min and Chen, Yong-cheng (2005). "Fluent speech prosody: framework and modeling," *Speech Communication*, Vol. 46, issues 3-4, (July 2005), Special Issue on Quantitative Prosody Modelling for Natural Speech Description and Generation, 284-309.
- [23] S.-H. Chen, W.-H. Lai, and Y.-R. Wang, "A New Duration Modeling Approach for Mandarin Speech," *IEEE Trans. On Speech and Audio Processing*, vol. 11, no. 4, July 2003.
- [24] Sin-Horng Chen, Wen-Hsing Lai and Yih-Ru Wang, "A statistical pitch contour model for Mandarin speech," *J. Acoust. Soc. Am.*, 117 (2), pp. 908-925, Feb. 2005.
- [25] 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)
- [26] 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)

- [27] 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
- [28] 洪國興，以語料庫為基礎之中文文句翻語音系統實現，國立交通大學電信工程學系碩士論文，2006
- [29] K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi, “Hidden Markov Models Based on MultiSpace Probability Distribution for Pitch Pattern Modeling,” Proc. of ICASSP, 1999.
- [30] Heiga Zen and Tomoki Toda, “An Overview of Nitech HMM-based Speech Synthesis System for Blizzard Challenge 2005,” Proc. of InterSpeech 2005, pp. 93-96
- [31] J. Lafferty, A. McCallum, and F. Pereira. “Conditional random fields: Probabilistic models for segmenting and labeling sequence data, ” Proc. of ICML, pp.282-289, 2001.
- [32] Hai Zhao, Chang-Ning Huang, and Mu Li, “An Improved Chinese Word Segmentation System with Conditional Random Field,” Proc. of SIGHAN-5, pp.162-165, Sydney, Australia, July 22-23, 2006.
- [33] CRF++, available on <http://crfpp.sourceforge.net/>
- [34] 龔萬炆，2004，“臺灣客家文學選集小說”。
- [35] J. Köhler, “Multilingual phone models for vocabulary independent speech recognition task,” Speech Communication, Vol. 35(1-2), pp. 21-30, 2001.
- [36] B. H. Juang, L. R. Rabiner, “A probabilistic distance measure for hidden Markov model,” Bell System Technical Journal, 64(2), pp. 391-408, 1985.

附錄一

四縣腔客語 671 個音節、17 個聲母、71 個韻母對照表

音碼	拼音	子音	母音	音碼	通用	子音	母音	音碼	通用	子音	母音
1	bi	b	i	30	bok	b	ok	59	piak	p	iak
2	be	b	e	31	buk	b	uk	60	pok	p	ok
3	ba	b	a	32	pi	p	i	61	piok	p	iok
4	bia	b	ia	33	pa	p	a	62	puk	p	uk
5	bo	b	o	34	pia	p	ia	63	mi	m	i
6	bu	b	u	35	po	p	o	64	me	m	e
7	biu	b	iu	36	pu	p	u	65	ma	m	a
8	bai	b	ai	37	piu	p	iu	66	mia	m	ia
9	boi	b	oi	38	pai	p	ai	67	mo	m	o
10	beu	b	eu	39	poi	p	oi	68	mu	m	u
11	bau	b	au	40	peu	p	eu	69	mai	m	ai
12	bin	b	in	41	pau	p	au	70	moi	m	oi
13	ben	b	en	42	pin	p	in	71	meu	m	eu
14	bien	b	ien	43	pen	p	en	72	mau	m	au
15	ban	b	an	44	pien	p	ien	73	miau	m	iau
16	bun	b	un	45	pan	p	an	74	min	m	in
17	bang	b	ang	46	pon	p	on	75	men	m	en
18	biang	b	iang	47	pun	p	un	76	mien	m	ien
19	bong	b	ong	48	pang	p	ang	77	man	m	an
20	biong	b	iong	49	piang	p	iang	78	mun	m	un
21	bung	b	ung	50	pong	p	ong	79	mang	m	ang
22	bit	b	it	51	piong	p	iong	80	miang	m	iang
23	bet	b	et	52	pung	p	ung	81	mong	m	ong
24	biet	b	iet	53	pit	p	it	82	miong	m	iong
25	bat	b	at	54	pet	p	et	83	mung	m	ung
26	bot	b	ot	55	piet	p	iet	84	mit	m	it
27	but	b	ut	56	pat	p	at	85	met	m	et
28	bak	b	ak	57	put	p	ut	86	miet	m	iet
29	biak	b	iak	58	pak	p	ak	87	mat	m	at

音碼	通用	子音	母音	音碼	通用	子音	母音	音碼	通用	子音	母音
88	mut	m	ut	117	ve	v	e	146	deu	d	eu
89	mak	m	ak	118	va	v	a	147	dau	d	au
90	mok	m	ok	119	vo	v	o	148	diau	d	iau
91	muk	m	uk	120	vu	v	u	149	dem	d	em
92	fi	f	i	121	vai	v	ai	150	dam	d	am
93	fe	f	e	122	voi	v	oi	151	diam	d	iam
94	fa	f	a	123	van	v	an	152	din	d	in
95	fo	f	o	124	von	v	on	153	den	d	en
96	fu	f	u	125	vun	v	un	154	dien	d	ien
97	fai	f	ai	126	vang	v	ang	155	dan	d	an
98	foi	f	oi	127	vong	v	ong	156	don	d	on
99	feu	f	eu	128	vung	v	ung	157	dun	d	un
100	fam	f	am	129	vet	v	et	158	dang	d	ang
101	fin	f	in	130	vat	v	at	159	dong	d	ong
102	fen	f	en	131	vut	v	ut	160	diong	d	iong
103	fan	f	an	132	vak	v	ak	161	dung	d	ung
104	fon	f	on	133	vok	v	ok	162	dep	d	ep
105	fun	f	un	134	vuk	v	uk	163	dap	d	ap
106	fang	f	ang	135	di	d	i	164	diap	d	iap
107	fong	f	ong	136	de	d	e	165	dit	d	it
108	fung	f	ung	137	da	d	a	166	det	d	et
109	fap	f	ap	138	dia	d	ia	167	diet	d	iet
110	fit	f	it	139	do	d	o	168	dat	d	at
111	fet	f	et	140	dio	d	io	169	dot	d	ot
112	fat	f	at	141	du	d	u	170	dut	d	ut
113	fok	f	ok	142	diu	d	iu	171	dak	d	ak
114	fut	f	ut	143	dai	d	ai	172	diak	d	iak
115	fuk	f	uk	144	doi	d	oi	173	dok	d	ok
116	vi	v	i	145	dui	d	ui	174	diok	d	iok

音碼	通用	子音	母音	音碼	通用	子音	母音	音碼	通用	子音	母音
175	duk	d	uk	204	tet	t	et	233	nap	n	ap
176	ti	t	i	205	tiet	t	iet	234	nit	n	it
177	te	t	e	206	tat	t	at	235	net	n	et
178	ta	t	a	207	tot	t	ot	236	nat	n	at
179	to	t	o	208	tut	t	ut	237	not	n	ot
180	tu	t	u	209	tak	t	ak	238	nak	n	ak
181	tiu	t	iu	210	tok	t	ok	239	nuk	n	uk
182	tai	t	ai	211	tuk	t	uk	240	li	l	i
183	toi	t	oi	212	ni	n	i	241	le	l	e
184	tui	t	ui	213	ne	n	e	242	la	l	a
185	teu	t	eu	214	na	n	a	243	lo	l	o
186	tiau	t	iau	215	no	n	o	244	lio	l	io
187	tiem	t	iem	216	nu	n	u	245	lu	l	u
188	tam	t	am	217	niu	n	iu	246	liu	l	iu
189	tiam	t	iam	218	nai	n	ai	247	lai	l	ai
190	tin	t	in	219	nui	n	ui	248	loi	l	oi
191	ten	t	en	220	neu	n	eu	249	lui	l	ui
192	tien	t	ien	221	nau	n	au	250	leu	l	eu
193	tan	t	an	222	niau	n	iau	251	lau	l	au
194	ton	t	on	223	nem	n	em	252	liau	l	iau
195	tun	t	un	224	nam	n	am	253	lim	l	im
196	tang	t	ang	225	nin	n	in	254	lem	l	em
197	tong	t	ong	226	nen	n	en	255	lam	l	am
198	tiong	t	iong	227	nan	n	an	256	liam	l	iam
199	tung	t	ung	228	non	n	on	257	lin	l	in
200	tep	t	ep	229	nun	n	un	258	lien	l	ien
201	tap	t	ap	230	nang	n	ang	259	lan	l	an
202	tiap	t	iap	231	nong	n	ong	260	lon	l	on
203	tit	t	it	232	nung	n	ung	261	lion	l	ion

音碼	通用	子音	母音	音碼	通用	子音	母音	音碼	通用	子音	母音
262	lun	l	un	291	zio	z	io	320	zep	z	ep
263	lang	l	ang	292	zu	z	u	321	zap	z	ap
264	liang	l	iang	293	ziu	z	iu	322	ziap	z	iap
265	long	l	ong	294	zai	z	ai	323	ziit	z	iit
266	liong	l	iong	295	zoi	z	oi	324	zit	z	it
267	lung	l	ung	296	zui	z	ui	325	zet	z	et
268	liung	l	iung	297	zeu	z	eu	326	ziet	z	iet
269	lip	l	ip	298	zau	z	au	327	zat	z	at
270	lep	l	ep	299	ziau	z	iau	328	zot	z	ot
271	lap	l	ap	300	ziim	z	iim	329	zut	z	ut
272	liap	l	iap	301	zim	z	im	330	zak	z	ak
273	lit	l	it	302	zem	z	em	331	ziak	z	iak
274	let	l	et	303	zam	z	am	332	zok	z	ok
275	liet	l	iet	304	ziam	z	iam	333	ziok	z	iok
276	lat	l	at	305	ziin	z	iin	334	zuk	z	uk
277	lot	l	ot	306	zin	z	in	335	ziuk	z	iuk
278	lut	l	ut	307	zen	z	en	336	cii	c	ii
279	lak	l	ak	308	zien	z	ien	337	ci	c	i
280	liak	l	iak	309	zan	z	an	338	ce	c	e
281	lok	l	ok	310	zon	z	on	339	ca	c	a
282	liok	l	iok	311	zun	z	un	340	cia	c	ia
283	luk	l	uk	312	zang	z	ang	341	co	c	o
284	liuk	l	iuk	313	ziang	z	iang	342	cio	c	io
285	zii	z	ii	314	zong	z	ong	343	cu	c	u
286	zi	z	i	315	ziong	z	iong	344	ciu	c	iu
287	ze	z	e	316	zung	z	ung	345	cai	c	ai
288	za	z	a	317	ziung	z	iung	346	coi	c	oi
289	zia	z	ia	318	ziip	z	iip	347	cui	c	ui
290	zo	z	o	319	zip	z	ip	348	ceu	c	eu

音碼	通用	子音	母音	音碼	通用	子音	母音	音碼	通用	子音	母音
349	cau	c	au	378	cut	c	ut	407	sen	s	en
350	ciau	c	iau	379	cak	c	ak	408	sien	s	ien
351	ciim	c	iim	380	ciak	c	iak	409	san	s	an
352	cim	c	im	381	cok	c	ok	410	son	s	on
353	cam	c	am	382	ciok	c	iok	411	sun	s	un
354	ciam	c	iam	383	cuk	c	uk	412	sang	s	ang
355	ciin	c	iin	384	ciuk	c	iuk	413	siang	s	iang
356	cin	c	in	385	sii	s	ii	414	song	s	ong
357	cen	c	en	386	si	s	i	415	siong	s	iong
358	cien	c	ien	387	se	s	e	416	sung	s	ung
359	can	c	an	388	sa	s	a	417	siung	s	iung
360	con	c	on	389	sia	s	ia	418	siip	s	iip
361	cion	c	iong	390	so	s	o	419	sip	s	ip
362	cun	c	un	391	sio	s	io	420	sep	s	ep
363	cang	c	ang	392	su	s	u	421	sap	s	ap
364	ciang	c	iang	393	siu	s	iu	422	siap	s	iap
365	cong	c	ong	394	sai	s	ai	423	siit	s	iit
366	ciong	c	iong	395	soi	s	oi	424	sit	s	it
367	cung	c	ung	396	sui	s	ui	425	set	s	et
368	ciung	c	iung	397	seu	s	eu	426	siet	s	iet
369	cip	c	ip	398	sau	s	au	427	sat	s	at
370	cap	c	ap	399	siau	s	iau	428	sot	s	ot
371	ciap	c	iap	400	siim	s	iim	429	sut	s	ut
372	ciit	c	iit	401	sim	s	im	430	sak	s	ak
373	cit	c	it	402	sem	s	em	431	siak	s	iak
374	cet	c	et	403	sam	s	am	432	sok	s	ok
375	ciet	c	iet	404	siam	s	iam	433	siok	s	iok
376	cat	c	at	405	siin	s	iin	434	suk	s	uk
377	cot	c	ot	406	sin	s	in	435	siuk	s	iuk

音碼	通用	子音	母音	音碼	通用	子音	母音	音碼	通用	子音	母音
436	gi	g	i	465	gang	g	ang	494	kia	k	ia
437	ge	g	e	466	giang	g	iang	495	kua	k	ua
438	gie	g	ie	467	gong	g	ong	496	ko	k	o
439	ga	g	a	468	giong	g	iong	497	kio	k	io
440	gia	g	ia	469	gung	g	ung	498	ku	k	u
441	gua	g	ua	470	giung	g	iung	499	kiu	k	iu
442	go	g	o	471	guang	g	uang	500	kai	k	ai
443	gio	g	io	472	gip	g	ip	501	kuai	k	uai
444	gu	g	u	473	giep	g	iep	502	koi	k	oi
445	giu	g	iu	474	gap	g	ap	503	kioi	k	ioi
446	guai	g	uai	475	giap	g	iap	504	kui	k	ui
447	goi	g	oi	476	git	g	it	505	keu	k	eu
448	gui	g	ui	477	giet	g	iet	506	kau	k	au
449	gau	g	au	478	guet	g	uet	507	kieu	k	ieu
450	gieu	g	ieu	479	gat	g	at	508	kiau	k	iau
451	giau	g	iau	480	guat	g	uat	509	kim	k	im
452	gim	g	im	481	got	g	ot	510	kam	k	am
453	giem	g	iem	482	gut	g	ut	511	kiam	k	iam
454	gam	g	am	483	gak	g	ak	512	kiem	k	iem
455	giam	g	iam	484	giak	g	iak	513	kin	k	in
456	gin	g	in	485	gok	g	ok	514	ken	k	en
457	gen	g	en	486	giok	g	iok	515	kien	k	ien
458	gien	g	ien	487	guk	g	uk	516	kan	k	an
459	guen	g	uen	488	giuk	g	iuk	517	kuan	k	uan
460	gan	g	an	489	guak	g	uak	518	kon	k	on
461	guan	g	uan	490	ki	k	i	519	kun	k	un
462	gon	g	on	491	kie	k	ie	520	kiun	k	iun
463	gun	g	un	492	kue	k	ue	521	kang	k	ang
464	giun	g	iun	493	ka	k	a	522	kiang	k	iang

音碼	通用	子音	母音	音碼	通用	子音	母音	音碼	通用	子音	母音
523	kong	k	ong	552	nguai	ng	uai	581	ngok	ng	ok
524	kiong	k	iong	553	ngoi	ng	oi	582	ngiok	ng	iok
525	kung	k	ung	554	ngui	ng	ui	583	ngiuk	ng	iuk
526	kiung	k	iung	555	ngau	ng	au	584	hi	h	i
527	kip	k	ip	556	ngieu	ng	ieu	585	he	h	e
528	kep	k	ep	557	ngiau	ng	iau	586	ha	h	a
529	kap	k	ap	558	ngim	ng	im	587	hia	h	ia
530	kiap	k	iap	559	ngiem	ng	iem	588	ho	h	o
531	kit	k	it	560	ngam	ng	am	589	hio	h	io
532	kiet	k	iet	561	ngiam	ng	iam	590	hiu	h	iu
533	kat	k	at	562	ngin	ng	in	591	hai	h	ai
534	kut	k	ut	563	ngien	ng	ien	592	hoi	h	oi
535	kiut	k	iut	564	ngan	ng	an	593	heu	h	eu
536	kak	k	ak	565	nguan	ng	uan	594	hau	h	au
537	kiak	k	iak	566	ngion	ng	ion	595	hieu	h	ieu
538	kok	k	ok	567	ngiun	ng	iun	596	hiau	h	iau
539	kiok	k	iok	568	ngang	ng	ang	597	him	h	im
540	kuk	k	uk	569	ngiang	ng	iang	598	hem	h	em
541	kiuk	k	iuk	570	ngong	ng	ong	599	ham	h	am
542	kuak	k	uak	571	ngiong	ng	iong	600	hiam	h	iam
543	ngi	ng	i	572	ngiung	ng	iung	601	hin	h	in
544	ngie	ng	ie	573	ngip	ng	ip	602	hen	h	en
545	nga	ng	a	574	ngap	ng	ap	603	hien	h	ien
546	ngia	ng	ia	575	ngiap	ng	iap	604	han	h	an
547	ngo	ng	o	576	ngit	ng	it	605	hon	h	on
548	ngio	ng	io	577	ngiet	ng	iet	606	hiun	h	iun
549	ngu	ng	u	578	ngat	ng	at	607	hang	h	ang
550	ngiu	ng	iu	579	ngut	ng	ut	608	hong	h	ong
551	ngai	ng	ai	580	ngiak	ng	iak	609	hiong	h	iong

音碼	通用	子音	母音	音碼	通用	子音	母音	音碼	通用	子音	母音
610	hiung	h	iung	639	am	INULL	am	668	ng	ng	FNULL
611	hiing	h	iing	640	iam	INULL	iam	669	iuì	INULL	iuì
612	hip	h	ip	641	in	INULL	in	670	ngung	ng	ung
613	hap	h	ap	642	en	INULL	en	671	nguk	ng	uk
614	hiap	h	iap	643	ien	INULL	ien				
615	hit	h	it	644	an	INULL	an				
616	het	h	et	645	on	INULL	on				
617	hiet	h	iet	646	iun	INULL	iun				
618	hat	h	at	647	ang	INULL	ang				
619	hot	h	ot	648	iang	INULL	iang				
620	hak	h	ak	649	ong	INULL	ong				
621	hok	h	ok	650	iong	INULL	iong				
622	hiuk	h	iuk	651	iung	INULL	iung				
623	i	INULL	i	652	ip	INULL	ip				
624	e	INULL	e	653	ep	INULL	ep				
625	ie	INULL	ie	654	ap	INULL	ap				
626	a	INULL	a	655	iap	INULL	iap				
627	ia	INULL	ia	656	it	INULL	it				
628	o	INULL	o	657	et	INULL	et				
629	io	INULL	io	658	iet	INULL	iet				
630	iu	INULL	iu	659	at	INULL	at				
631	ai	INULL	ai	660	ot	INULL	ot				
632	oi	INULL	oi	661	iut	INULL	iut				
633	eu	INULL	eu	662	ak	INULL	ak				
634	au	INULL	au	663	iak	INULL	iak				
635	ieu	INULL	ieu	664	ok	INULL	ok				
636	iau	INULL	iau	665	iok	INULL	iok				
637	im	INULL	im	666	iuk	INULL	iuk				
638	em	INULL	em	667	m	m	FNULL				

附錄二

新增至客語詞典之一字多義一字詞表

客語詞	詞性組合	客語詞	詞性組合
個	DE	肚	Ncd_Ng
愛	D_VC_VE	當	Dfa_P
係	SHI	儕	Na_Nf
等	Di_VC	同	Caa_P
該	Dk_Ncd_Nep_T	歸	Nb_Neqa
異	Dfa	畜	Na_VC
到	Caa_P_VCL	摻	Caa_P
正	Da_D	屋	Na_Nc_Nf
項	Ncd_Nf_Ng	適	P_VJ
食	Na_VC	歇	VA_VCL
好	D_VH	驚	Na_VK
逐	D_Nes	企	Nc_VA
諗	Di_T_VJ	識	D_VJ
盡	Dfa	喊	P_VA_VC_VE_VF_VG_VL
分	P_VD_VL	供	VC_VH
擺	Na_Nf_VC	奈	Ncd_Nep_VJ_T
著	P_VC	遽	D_VH
恁	D_Dfa_VD	忒	Dfa_Di_VCL_VH
兜	Dfb_Nf	仰	D_VC_VH
仔	DE_Na_Nf		

附錄三

詞類標記								
編號	詞類	代號	編號	詞類	代號	編號	詞類	代號
1	非謂形容詞	A	18	特指定詞	Nes	34	動作句賓動詞	VE
2	對等連接詞	Caa	19	指代定詞	Nep	35	動作謂賓動詞	VF
3	連接詞，如：等等	Cab	20	數量定詞	Neqa	36	分類動詞	VG
4	連接詞，如：的話	Cba	21	後置數量定詞	Neqb	37	狀態不及物動詞	VH
5	關聯連接詞	Cbb	22	量詞	Nf	38	狀態使動動詞	VHC
6	數量副詞	Da	23	後置詞	Ng	39	狀態類及物動詞	VI
7	動詞前程度副詞	Dfa	24	代名詞	Nh	40	狀態及物動詞	VJ
8	動詞後程度副詞	Dfb	25	感嘆詞	I	41	狀態句賓動詞	VK
9	時態標記	Di	26	介詞	P	42	狀態謂賓動詞	VL
10	句副詞	Dk	27	語助詞	T	43	有	V_2
11	副詞	D	28	動作不及物動詞	VA	44	的，之，得，地	DE
12	普通名詞	Na	29	動作使動動詞	VAC	45	係	SHI
13	專有名詞	Nb	30	動作類及物動詞	VB	46	外文標記	FW
14	地方詞	Nc	31	動作及物動詞	VC	58		DM
15	位置詞	Ncd	32	動作接地方賓語動詞	VCL	59		Nv
16	時間詞	Nd	33	雙賓動詞	VD	60		PA
17	數詞定詞	Neu						

附錄四

聲母音素問題集		
	國語	客語
鼻音	M_m, M_n	H_m, H_n, H_ng, H_NH
塞音	M_p, M_t, M_ch, M_c, M_k, M_b, M_d, M_g, M_q	H_p, H_t, H_c, H_k, H_b, H_d, H_g, H_q
塞音清	M_p, M_t, M_ch, M_c, M_k	H_p, H_t, H_c, H_k
塞音濁	M_b, M_d, M_g, M_q	H_b, H_d, H_g, H_q
擦音	M_f, M_s, M_sh, M_x, M_h, M_z, M_zh	H_f, H_s, H_x, H_HH, H_h, H_v, H_z
擦音清	M_f, M_s, M_sh, M_x, M_h	H_f, H_s, H_x, H_HH, H_h
擦音濁	M_z, M_zh	H_v, H_z
近音	M_j	H_j
邊近音	M_l	H_l
閃音	M_r	
唇	M_m, M_p, M_b, M_f	H_m, H_p, H_b, H_f, H_v
雙唇	M_m, M_p, M_b	H_m, H_p, H_b
唇齒	M_f	H_f, H_v
齶音	M_n, M_t, M_d, M_s, M_z, M_l, M_r	H_n, H_t, H_d, H_s, H_z, H_l
捲舌	M_ch, M_sh, M_zh	
硬顎	M_c, M_j	H_c, H_j
	M_k, M_g, M_x, M_q	H_ng, H_k, H_g, H_x, H_NH, H_q
軟顎	M_k, M_g, M_x	H_ng, H_k, H_g, H_x
小舌	M_q	H_NH, H_q
	M_h	H_HH, H_h
會厭		H_HH
聲門	M_h	H_h
客語韻母音素問題集(依照韻母結構設定問題)		
元音 H_a	H_ek, H_em, H_en, H_eng, H_ep, H_et, H_wu3, H_yi3	
元音 H_eh	H_em, H_en, H_ep, H_et, H_wu3	
元音 H_ii	H_em, H_en, H_eng, H_ep, H_et	
元音 H_o	H_ek, H_en, H_eng, H_et, H_yi3	
元音 H_yi2	H_em, H_en, H_ep, H_et	
介音 H_wu1	H_en, H_et	
介音 H_yi1	H_ek, H_em, H_eng, H_ep, H_wu3	
韻尾 H_ek	H_a, H_o, H_wu1, H_wu2, H_yi1	
	H_wu1, H_yi1	
	H_a, H_o, H_wu2	
韻尾 H_em	H_a, H_eh, H_ii, H_yi1, H_yi2	
	H_yi1	
	H_a, H_eh, H_ii, H_yi2	
韻尾 H_en	H_a, H_eh, H_ii, H_o, H_wu1, H_wu2, H_yi1, H_yi2	

	H_wu1, H_yi1
	H_a, H_eh, H_ii, H_o, H_wu2, H_yi2
韻尾 H_eng	H_a, H_ii, H_o, H_wu1, H_wu2, H_yi1
	H_wu1, H_yi1
	H_a, H_ii, H_o, H_wu2
韻尾 H_ep	H_a, H_eh, H_yi1, H_yi2
	H_yi1
	H_a, H_eh, H_yi2
韻尾 H_et	H_a, H_eh, H_yi1
	H_yi1
	H_a, H_eh
韻尾 H_wu3	H_a, H_eh
韻尾 H_yi3	H_a, H_eh, H_o, H_wu2
無韻尾	H_a, H_eh, H_ii, H_o, H_wu1, H_wu2, H_yi1, H_yi2, H_FNULL
	H_wu1, H_yi1
	H_a, H_eh, H_ii, H_o, H_wu2, H_yi2,
	H_FNULL
國語韻母音素問題集(依照韻母結構設定問題)	
元音 M_e	M_ng, M_en
元音 M_eh	M_yi3
元音 M_o	M_wu3
元音 M_a	M_ng, M_en, M_wu3
介音 M_yi1	M_ng, M_en, M_yi3, M_wu3
	M_ng, M_en
介音 M_yu1	M_ng, M_en
	M_en
介音 M_wu1	M_ng, M_en, M_yi3
	M_ng, M_en
	M_yi3
韻尾 M_ng	M_e, M_a, M_yi1, M_yu1, M_wu1
	M_yi1, M_yu1, M_wu1
	M_e, M_a
韻尾 M_en	M_e, M_a, M_yi1, M_yu1, M_wu1
	M_yi1, M_yu1, M_wu1
	M_e, M_a
韻尾 M_yi3	M_eh, M_a, M_yi1, M_wu1
	M_yi1, M_wu1
	M_eh, M_a
韻尾 M_wu3	M_o, M_a, M_yi1
	M_yi1
	M_o, M_a
無韻尾	M_e, M_eh, M_o, M_a, M_er, M_yi1, M_yi2, M_yu1, M_yu2, M_wu1, M_wu2, M_FNULL1, M_FNULL2
	M_yi1, M_yu1, M_wu1
	M_e, M_eh, M_o, M_a, M_er, M_yi2, M_yu2, M_wu2
	M_FNULL1, M_FNULL2

附錄五

發表之論文依序為

- [1] Yi-Ling Tsai, Hsiu-Min Yu, Yih-Ru Wang, Chen-Yu Chiang, Lieh-Shih Lo, and Sin-Horng Chen, “An HMM-based Hakka Text-to-Speech System,” accepted by OCOCOSDA 2010.
- [2] Tsai-Lu Tsai, Chen-Yu Chiang, Hsiu-Min Yu, Lieh-Shih Lo, Yih-Ru Wang and Sin-Horng Chen, “A Study on Hakka and Mixed Hakka-Mandarin Speech Recognition,” accepted by ISCSLP 2010.
- [3] 余秀敏, “客語啊時貌之間的關係探討”, 第八屆台灣語言及其教學國際研討會論文集, pp. 216~227.

An HMM-based Hakka Text-to-Speech System

Yi-Ling Tsai, ^{*}Hsiu-Min Yu, Yih-Ru Wang, Chen-Yu Chiang, ⁺Lieh-Shih Lo, and Sin-Horng Chen

Institute of Communication Engineering, NCTU, Hsinchu, Taiwan

⁺*College of Hakka, NCTU, Hsinchu, Taiwan*

^{*}*Language Center, Chung Hua University, Hsinchu, Taiwan*

yiling750124@yahoo.com.tw, kuo@chu.edu.tw, yrwang@cc.nctu.edu.tw, gene.cm91g@nctu.edu.tw, asiisl@mail.nctu.edu.tw, and schen@mail.nctu.edu.tw

Abstract

In this paper, a Hakka Text-to-Speech (TTS) system is implemented. It is an HMM-based speech synthesis system. The work focuses on the realization of a Hakka parser to tag the input text into word and POS sequences. The difficulty lies in the lack of a large text corpus to train a robust Hakka parser. Motivated by the fact that Hakka is a dialect of Mandarin Chinese so as to share many linguistic properties of Chinese, we adopt a new approach to constructing a Hakka parser via extending an existing CRF-based Chinese parser to attach a Hakka dictionary and incorporate some Hakka word construction rules. Besides, a pause predictor is designed to estimate the inter-syllable locations to insert pauses for improving the fluency of the synthesized speech. A subjective quality test confirms that the Hakka TTS system is a promising one.

1. Introduction

In Taiwan, there are three major languages including Mandarin, Min-Nan/Taiwanese and Hakka. Over 90% people speak Mandarin, about 70% whose mother-tongue are Taiwanese, and only 10% speak Hakka. In the past, there are only few studies in both Min-Nan and Hakka TTS [1,2]. This is due in part to the lack of a unified written form representation and in part to the difficulty of collecting a large speech database. In recent years, Taiwan government started to pay more attention to mother-tongue education including Hakka. This motivates us to realize a useful Hakka TTS system for the society in this study.

Hakka includes several sub-dialects, namely Hoi-Liuk, Si-Rhan, Ta-Pu, etc. We choose Si-Rhan as the target in this study. Just like Mandarin and Min-Nan, Hakka speech is a monosyllabic and tonal language. Each character is pronounced as a syllable carrying a lexical tone. There are only 671 base-syllables and 6 tones. All base-syllables have the same *initial-final* structure like Mandarin and Min-Nan base-syllables. There are 17 *initials* and 72 *finals*. Although Hakka

speech has similar linguistic characteristics as Mandarin speech, its written form is just going to be standardized. This is mainly pushed by the Council for Hakka Affairs of Taiwan. But, the written system is still not popular nowadays in Taiwan. Three main types of word written forms can be found. Many Hakka words directly use the written forms of the Chinese words with the same meaning. Some Hakka words are represented by their own specific Chinese character strings. Usually, the character string can be pronounced in Mandarin with sound similar to the Hakka word it represents. Lastly, some extraordinary syllables are represented in the Romanization form. Due to the fact that the written form of Hakka is not widely used yet, it is difficult to collect a large text corpus. This makes the text analysis very difficult for Hakka language.

In this paper, the design and implementation of a Hakka TTS system is presented. The HMM-based Text-to-Speech (HTS) [3,4] technique is adopted to realize the system. The study pays special efforts on the issue of realizing a reliable Hakka parser using insufficient Hakka texts with the help of an existing Chinese parser. Besides, a pause predictor is also implemented to improve the HTS-based speech synthesis.

The paper is organized as follows. Section 2 presents the proposed Hakka TTS system. All functional blocks of the system are discussed in detail. Experimental results that evaluate the system's performance are discussed in Section 3. Some conclusions are given in the last section.

2. The Implementation of Hakka TTS

The Hakka TTS system is composed of four main parts: parser, pause predictor, context analyzer and HMM-based synthesizer. The input text is firstly tagged in the text analyzer into word and POS sequences. Then, the pause predictor estimates the inter-syllable locations to insert pauses. The context analyzer then generates the synthesis unit and some language parameters. Lastly, the HMM-based

synthesizer produces duration, pitch, and spectral parameters to generate the output synthesized speech. In the following, we discuss each part in more detail.

2.1 The Hakka Parser

In this study, we adopt a new approach to constructing the Hakka parser because we do not have a large text corpus to train a robust Hakka parser. The Hakka parser is an extended version of an existing Conditional Random Field [5] (CRF)-based Chinese parser attaching with a Hakka dictionary and incorporating with some Hakka word construction rules. The use of this approach is motivated by the fact that Hakka is a dialect of Mandarin Chinese so as to share many linguistic properties of Chinese, including using the same Chinese character set, using many Chinese words, and having a similar syntactic structure. An analysis of a well-tagged Hakka text corpus, which consists of many short articles of stories with total size of 63,158 syllables, reveals that the percentage of using Chinese words is as high as 48.97%. So, a Chinese parser can be used to assist in tagging Hakka texts, especially for processing those words commonly used in both Chinese and Hakka languages. As for the words with different form, we attach a Hakka dictionary and incorporate some Hakka word construction rules to process them.

Figure 1 displays a block diagram of the proposed Hakka parser. It consists of five parts, including symbol normalization, word segmentation, POS tagger, word construction and grapheme to phone. Since the proposed Hakka parser is an extended version of an existing Chinese parser, we first introduce the Chinese parser and then describe its extension to Hakka language in the following.

2.1.1 The CRF-based Chinese Parser

2.1.1.1 Symbol normalization

Since the Chinese parser takes BIG5 tokens to process word segmentation, the symbol normalization converts the format of input character sequence from ASCII to BIG5. Besides, it normalizes various BIG5 codes of each punctuation mark (PM) type to a single representative BIG5 code.

2.1.1.2 Word segmentation and POS tagging

Word segmentation and POS tagging can be formulated by

$$\begin{aligned} W^*, S^* &= \arg \max_{W, S} P(W, S | C) \\ &= \arg \max_{W, S} P(S | W, C) P(W | C) \end{aligned} \quad (1)$$

where W , S and C represent word, POS and character sequences, respectively; $P(W | C)$ and

$P(S | W, C)$ are correspondingly word segmentation and POS tagging model. In this study, both word segmentation and POS tagging models are implemented by the technique of Conditional Random Field (CRF) [5,6]. To simplify the decoding process in Eq. (1), top-N ($N=3$ in this study) word segmentation sequences are firstly generated by $P(W | C)$ and then rescored by $P(S | W, C)$.

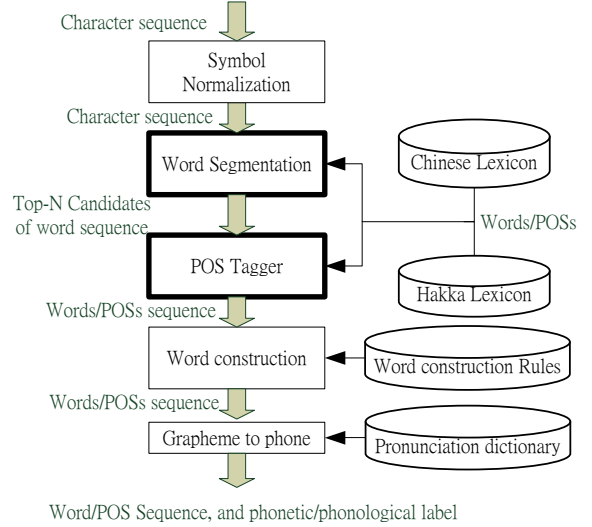


Fig.1: A block diagram of the proposed Hakka parser.

The CRF-based word segmentation can be viewed as a label-tagging problem. The word sequence W is expressed by a tag sequence Y that represents character positions in words. In this study, a six-tag label set [6] that includes B_1 (1st character in a word), B_2 (2nd in a word), B_3 (3rd in a word), M (middle in a word), E (end in a word), and S (single character word) is utilized. Hence, the CRF-based word segmentation model is expressed by

$$\begin{aligned} P(W | C) &= P(Y | C) \\ &= \frac{1}{N(C)} \exp \left(\sum_{t=1}^T \sum_{i=1}^I \lambda_i f_i(Y_t = y, Y_{t-1}, C) \right) \end{aligned} \quad (2)$$

where $N(C)$ is a normalization factor to ensure $\sum_Y P(Y | C) = 1$; I represents the number of feature functions; and $f_i(Y_t = y, Y_{t-1}, C)$ is a feature function defined by

$$f_i(Y_t = y, Y_{t-1}, C) = \begin{cases} 1, & \text{if } C = h_j \text{ is satisfied and } y = y_k \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where h_j represents the j -th possible feature context; y_k is the k -th possible tag to be labeled. Generally, feature contexts are organized into several groups,

referred to as “feature templates”. Table 1 displays the feature templates used in this study.

Table 1: Feature templates used for word segmentation

Feature	description
C_n	Character
R_n	Radical of Chinese character
PM_n	Pronunciation mark(0,1)
WL_n	Length of the longest sub-character string containing the current character that matches the entry in the Chinese lexicon.
WT_n	Six-tag label of the current character that is inside of the longest sub-character string containing the current character that matches the entry in the Chinese lexicon.
DP_n	length of maximum repeated string in the sentence (long-term parameter)
DT_n	Six-tag label tags of current character in the maximum repeated string in the sentence (long-term parameter)
D_n	Is the current character a digit?
E_n	Is current character a English alphabet?
Feature template (separated by comma)	
Character context: $C_{n-1}, C_n, C_{n+1}, C_{n-1}^n, C_n^{n+1}, (C_{n-1} C_{n+1})$	
Radical context: $(R_{n-1} R_n), (R_{n+1} R_n), (R_{n+2} R_{n+1})$	
Lexicon information: $(WL_n WT_n), WT_{n-1}^{n+1}, (WL_n WT_n DP_n DT_n)$	
Digit information $(D_{n-1} C_n D_{n+1})$	
alphabet information: $E_{n-1}^n, E_n^{n+1}, E_{n-1}^{n+1}$	

Notice that the four features ($C_n, PM_n, D_n, E_n, WL_n$ and WT_n) were proposed by Zhao [6]. We suggest adding the other three features (i.e. R_n, DP_n and DT_n) to model words that is related to their radicals or proper nouns that are frequently repeated in a paragraph but not included in the lexicon.

The POS tagging modeling is also realized by CRF with targets of 46 POS classes and features of word, and candidate POSs, etc. Table 2 displays the feature templates used for POS tagging.

2.1.1.3 Word construction

Some word types cannot be directly identified by the CRF-based word segmentation due to their high productivity. They include determinative-measure compound words, derivational words and reduplication words. Some word construction rules are incorporated to identify these words using some components, such as stem, affix and POS, obtained by the CRF-based word segmentation and POS tagging.

2.1.1.4 Grapheme to phone

The function of grapheme-to-phone is to generate the phonetic and phonological transcriptions, like the syllable type and lexical tone of each syllable, for

speech synthesis. Here, this is simply implemented by looking up the pronunciation dictionary for each word tokenized by the word segmentation and POS tagger.

Table 2: Feature templates used for POS tagging.

Feature	description
PM_n	Type of punctuation marks
WL_n	Length of the current word in character.
LS_n	All possible POSs of the current word (available if the current word can be found in the Chinese/Hakka lexicon)
FC_n	The first character in the current word(available if the current word can NOT be found in the Chinese/Hakka lexicon)
LC_n	The last character in the current word(available if the current word can NOT be found in the Chinese/Hakka lexicon)
Feature template (separated by comma)	
PM context: PM_{n-1}^{n+1}	
Candidate POS context: $LS_{n-1}, LS_n, LS_{n+1}, LS_{n-2}^{n-1}, LS_{n-1}^n, LS_n^{n+1}, LS_{n+1}^{n+2}$	
Special features for words not included in the lexicon: WL_n, FC_n, LC_n	

2.2 Extension to Realize the Hakka Parser

We now extend the above CRF-based Chinese parser to a Hakka parser by attaching a Hakka dictionary and incorporating some Hakka word construction rules. The Hakka dictionary contains 34541 entries. Words in the Hakka dictionary are of higher priority to use than those of the Chinese dictionary. Two sets of word construction rules are used. One is the rules commonly used for Chinese and Hakka languages, and another is the parallel rules used respectively for the two languages. They are listed in Tables 3 and 4, respectively.

2.3 Pause Predictor

Silent pause (SP) insertion plays a very important role in TTS to influence the fluency and naturalness of the synthesized speech. When conducting the HTS training, SP is treated as an HMM unit when its duration is larger than a threshold (set to be 0.03 second in this study); otherwise it is ignored. But, in the test phase, the existence/absence of SP is not known and can only be predicted from the input linguistic features. Thus, a pause insertion predictor is essential for TTS. It is generally agreed that all inter-syllable locations accompanying punctuation marks should be mapped to SPs. Besides, some word or syntactic boundaries may also need to insert SPs for improving the naturalness of the synthesized speech. This is especially true for synthesizing a long sentential utterance. In this study, a

CRF-based pause predictor is constructed with the targets of the two classes (SP existence and SP absence) and the contextual linguistic features around the current word juncture. The feature and template set used to construct the CRF-based pause predictor is summarized in Table 5. In the test phase, the existence/absence of SP on each word juncture is determined by the well-trained CRF-based predictor using the input linguistic features. Then, the SP labels to be used in the HTS synthesizer are formed.

Table 3: The set of word construction rules commonly used for Chinese and Hakka languages.

word construction rules	Chinese example	Hakka example
XXY(Z)	說說看	日日大
X-XY-Y	乾乾淨淨	洗洗湯湯
V+了/著/過/看/看看	想過	歇過
名詞後詞綴：氏/度/性/家/長/師/員/們/子/倆	女性朋友們	細妹伴仔
排行:數字+姐/妹/哥/弟/伯/叔/姑	二伯	二伯
VG+為/成/作/不了/得了	身為	身為

Table 4: The set of parallel word construction rules used respectively for Chinese and Hakka languages.

Chinese word construction rules	Chinese example	Hakka word construction rules	Hakka example
V+不+到/起/出	看不起	V+毋+到/起/出	看毋起
越+來+越	越來越少	緊+來+緊	緊來緊少
越+X+越+Y	越穿越多	緊+X+緊+Y	緊著緊多
VB/VH+不休/今 今/不絕/不下/不 得/開來	蜿蜒不絕	V+毋得	捨毋得
老/小/阿+Na	小妹妹	細+Na	細妹仔

2.4 Context Analyzer

The context analyzer is designed to form the labels representing the contextual linguistic information for each HMM synthesis unit by using the outputs of the Hakka parser. In the study, syllable initials, syllable finals, and SP are taken as the basic synthesis unit. In the training of HTS, the labels hence contain the initial/final-level segmentation information, the contextual phonetic information, and some word-level and sentence-level linguistic features. Table 6 lists the labels generated by the context analyzer. In the test phase, the labels contain the same information as in the training phase, except for the two labels indicating the beginning/end timing information of HMM units

because these two labels refer to HMM synthesis units' durations to be predicted by the duration model in HTS.

Table 5: Features and templates used in the CRF-based SP predictor.

Feature	description
PM_n	Pronunciation mark following word n
WL_n	Word length in syllable of word n
POS_n	POS of word n
I_n	Initial type of the first syllable in word n
F_n	Final type of the last syllable in word n
Template	
$PM_{n-1}, PM_n, PM_{n+1}, PM_{n-1}^{n+1}, PM_{n-2}^n$	
$WL_{n-2}, WL_{n-1}, WL_n, WL_{n+1}, WL_{n+2}, WL_{n-2}^{n+2}$	
$POS_{n-1}, POS_n, POS_{n+1}$	
$I_{n+1}, F_n, (I_{n+1} + F_n), (POS_n + I_{n+1} + F_n)$	

Table 6: The labels generated by the context analyzer.

T1/T2	Beginning/end times of HMM unit
p1/p2/p3	Previous(PRE)/current(CUR)/following(FOL) Initial/Final/SP
t1/t2/t3	Lexical tones of PRE/CUR/FOL syllable
w1/w2	Syllable position in a lexical word (LW) (forward/backward)
s1/s2	Syllable position in a sentence (forward/backward)
PM	Punctuation mark after the current syllable
w3/w4/w5	Lengths of PRE/CUR/FOL LWs in syllable
P1/P2/P3	POs of PRE/CUR/FOL LWs
s3/s4/s5	Lengths of PRE/CUR/FOL sentences in syllable

2.5 HMM-based Speech Synthesizer

The standard context-dependent HMM training for speech synthesis [4,5] is adopted here to simultaneously construct spectral, f0 and state duration models using the labels generated by the context analyzer and SP predictor. Five-state left-to-right HMMs are used to model each synthesis unit of syllable initial and final. Speech signal is converted into 24-dimensional mel-generalized cepstral (MGC) sequence in 5ms interval. The question set used for the decision-tree based context clustering of HMMs is formed by using the context labels. To achieve a better tree clustering result, we merge some individual linguistic features to form several complex questions according to their effects on spectral and prosody productions. For examples, the initial/final types are classified by the manner or place of articulation; lexical tones are tied according to their F0 level in the syllable

beginning or ending; POS categories are merged into broader classes, etc. There are in total 2001 questions formed for the decision-tree based context clustering of HMMs in this study.

3. Experimental Results

In this section, some experiments were conducted to examine the performances of the CRF-based parser, the pause predictor, and the Hakka TTS.

3.1 The CRF-based Parser

3.1.1. The Chinese Parser

The training and test corpora used for the Chinese CRF-based parser are those used in the Chinese Word Segmentation Competition Test hosted by the SIGHAN of the Association for Computational Linguistics [6]. Their sizes are 8,350,369 and 196,867 words, respectively. The corpus used for training and testing the POS tagger is the Academia Sinica Balanced Corpus of Modern Chinese V.3.0 [7]. Its size is 5,841,942 words. It is divided into two sets: a 4,426,886-word training set and a 484,070-word test set. The number of POS classes is 46. The Chinese lexicon used contains 88,142 entries. The experimental results of word segmentation by the CRF-based Chinese parser are displayed in Table 7. High F-scores of 96.4 and 98.9 were achieved in the inside and outside tests. For POS parser, accuracy rates of 96.0% and 94.7% were obtained in the inside and outside tests, respectively.

Table 7: Experimental results of word segmentation by the CRF-based Chinese parser.

	precision	recall	F-score
Inside test	96.0%	96.8%	96.4
Outside test	98.9%	98.8%	98.9

3.1.2. The Hakka Parser

A novel entitled “阿啾箭个故乡” in the book by Gong [8] is used to test the performance of the Hakka parser. It is composed of 42 short articles with total length of 63,158 syllables or 36,450 words. Two experiments were examined. Exp A is conducted by the Chinese parser and Exp B by the Hakka parser. The experimental results of Hakka word segmentation are displayed in Table 8. It can be found from the table that the F-score of directly using the Chinese parser is 74 and raises to 82.9 as the Hakka parser is used. This confirms the effectiveness of the proposed Hakka parser. Moreover, the accuracy rates of POS tagging is 59.4% for Exp A and 77.1% for Exp B.

Table 8: Experimental results of Hakka word segmentation.

	precision	recall	F-score
Exp A	71.1%	77.2%	74.0
Exp B	84.0%	81.8%	82.9

3.2 The Pause Predictor

A read-speech corpus was used to examine the performance of the proposed CRF-based pause predictor. It was generated by a male speaker. The associated texts are the articles of the above text corpus. It is divided into two sets: 36,991 words for training and 4,095 words for test. We set a durational threshold of 25 ms to determine whether there is a pause (SP) at an inter-word boundary. A scheme to simply set pauses at all PMs was also tested and taken as the baseline. An accuracy rate of 85.7% was obtained by the baseline scheme. For the proposed predictor, accuracy rates of 87% and 88% were achieved in the inside and outside tests, respectively. The confusion matrix for the outside test is displayed in Table 9.

Table 9: The confusion matrix of pause prediction for the outside test. (# and * denote true and predicted ones, respectively)

	SP*	Non-SP*
SP#	1035	363
Non-SP#	150	2409

3.3 The Quality Test of Synthesized Speech

Two speech data sets were used to test the speech quality of the synthesized speech. They were generated by a male speaker and a female respectively, and consisted of 63,158 and 42,337 syllables. Table 10 shows the results of objective RMSE tests for the HMM-based syntheses of syllable duration and syllable pitch contour.

Table 10: Experimental results of objective RMSE tests

		Inside test	Outside test
Duration RMSE	Female	37.35(Hz)	45.64(Hz)
	Male	16.08(Hz)	19.4(Hz)
Pitch RMSE	Female	37.31(ms)	33.13(ms)
	Male	45.13(ms)	47.57(ms)

Then, a subjective test was conducted to evaluate the naturalness of the two TTS systems using different pause prediction schemes. The baseline system is the one with pauses inserted to all PM locations. Another is

the one using the proposed CRF-based pause predictor. The number of subjects involved in the test was ten (five females and five males). All the subjects' mother tongues are Hakka. Utterances of ten sentences in the test set were chosen. The shortest and longest sentences had 20 and 72 characters, respectively. Subjects were requested to grade each test utterance in a five-level mean opinion score (MOS) as shown in Table 11.

Table 11: Five-level MOS

MOS	5	4	3	2	1
Quality	Excellent	Good	Fair	Poor	Bad

Figure 2 displays the experimental results. It can be found from the figure that the system with the CRF-based pause predictor performed better than the baseline system. We also find that the male voices sounded much more natural than the female voices. This is mainly resulted from the higher quality of the training dataset of the male speech.

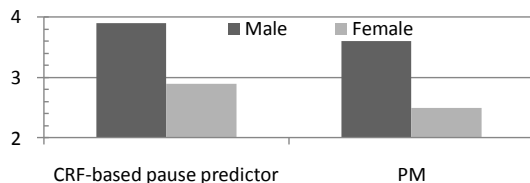


Fig. 2: MOSs of two TTS systems.

4. Conclusion

The design and implementation of an HMM-based Hakka TTS system is presented in this paper. This study focuses on the realization of a reliable Hakka parser based on an existing CRF-based Chinese parser. Experimental results showed that the proposed Hakka parser is effective as we incorporate some Hakka word construction rules and attach a Hakka lexicon to the Chinese parser. Furthermore, the naturalness of the synthesized speech was improved by using a CRF-based pause predictor. Further study to

use more Chinese language information to assist in improving the Hakka parser is worth doing in the future.

ACKNOWLEDGMENT

This work was supported by NSC of Taiwan under contract NSC98-2221-E-009-075-MY3. The authors would like to thank the Council for Hakka Affairs and MOE of Taiwan, H. M. Yu, and W. Z. Gong for supplying the lexicon and the text corpus.

References

- [1] Wei-Chih Kuo, Xiang-Rui Zhong, Yih-Ru Wang and Sin-Horng Chen, "A High-Performance Min-Nan/ Taiwanese TTS System", ICASSP2003.
- [2] Hsiu-Min Yu, Lie-Shih Lo, Hsin-Te Hwang, Hsi-Chun Hsiao, and Sin-Horng Chen, "On Constructing Speech Corpus for Implementing Hakka Text-To-Speech Synthesis," OCOOSDA, Dec. 2007.
- [3] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, K. Tokuda, *The HMM-based speech synthesis system version 2.0*, Proc. of ISCA SSW6, Bonn, Germany, Aug. 2007.
- [4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, *Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis*, Proc. of Eurospeech, pp.2347-2350, Sept. 1999.
- [5] J. Lafferty, A. McCallum, and F. Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", In Proc. of ICML, pp.282-289, 2001.
- [6] Hai Zhao, Chang-Ning Huang, and Mu Li, "An Improved Chinese Word Segmentation System with Conditional Random Field", Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing (SIGHAN-5), pp.162-165, Sydney, Australia, July 22-23, 2006.
- [7] Hsu, Hui-li and Chu-Ren Huang, 1995. Design Criteria for a Balanced Modern Chinese Corpus. Proceedings of ICCPOL'95, Hawaii.
- [8] Wan-Zao Gong, The Taiwan Hakka Literature Collections (Global, Taipei, Taiwan, 2004), in Chinese/Hakka.

A Study on Hakka and Mixed Hakka-Mandarin Speech Recognition

Tsai-Lu Tsai, Chen-Yu Chiang,*Hsiu-Min Yu, [†]Lieh-Shih Lo, Yih-Ru Wang and Sin-Horng Chen
Institute of Communication Engineering, National Chiao Tung University, Hsinchu, Taiwan, R.O.C.

[†]College of Hakka Studies, National Chiao Tung University, Hsinchu, Taiwan, R.O.C.

*Language Center, Chung Hua University, Hsinchu, Taiwan, R.O.C.

remnantsnow.cm97g@nctu.edu.tw, gene.cm91g@nctu.edu.tw, kuo@chu.edu.tw, asiilsl@mail.nctu.edu.tw,
yrlwang@cc.nctu.edu.tw and schen@mail.nctu.edu.tw

Abstract—A first study on Hakka and mixed Hakka-Mandarin speech recognition (SR) is reported in this paper. The main focus of the study is on solving the problem of the lack of a large text corpus for training a reliable language model. In the Hakka SR, several methods to use the information of part of speech and Hakka-Chinese word translation to assist in language modeling are proposed. For mixed language SR, a method to train a mixed Hakka-Mandarin acoustic model is suggested. Experimental results show that the proposed language and acoustic modeling approaches are promising for Hakka and mixed Hakka-Mandarin SR.

Keywords—*speech recognition; language model; acoustic model; Hakka; Mandarin*

I. INTRODUCTION

There are three major languages in Taiwan, including Mandarin, Min-Nan, and Hakka. Only 10% people's mother-tongues are Hakka. In the past, there are only few studies in Hakka speech processing [1,2] due to the difficulty in collecting Hakka speech and text corpora. Recently, both Taiwan government and society started to pay more attention to mother-tongue education, especially in Hakka. Now, there are many broadcast media, like the Hakka-TV [3] and the Hakka-radio [4], broadcast news in Hakka or mixed Hakka and Mandarin. The need for a Hakka/mixed Hakka-Mandarin dictation system is expected in the near future. This motivates us to accomplish two speech recognition prototype systems: (1) a Hakka speech recognition system and, (2) a mixed Hakka-Mandarin speech recognition system.

Hakka includes several sub-dialects, namely Hoi-Liuk, Si-Rhan, Ta-Pu, etc. We choose Si-Rhan in this study. Just like Mandarin, Si-Rhan Hakka speech is a monosyllabic and tonal language. Although Hakka speech has similar linguistic characteristics as Mandarin speech, it does not have a standard written form. The written form is a hybrid one which uses Chinese characters to represent ordinary words and represents some extraordinary syllables in the Romanization form. Unfortunately, only few Hakka experts could write in Hakka texts. This makes the collection of a large-size Hakka text corpus for constructing a reliable language model (LM) very difficult. To overcome this problem in the Hakka speech recognition task, we incorporate POS (part of speech)-based LM, Hakka-Mandarin word translation, and Mandarin LM to help to construct a more reliable Hakka LM. This is mainly motivated by the fact that Hakka has similar linguistic

characteristics as Mandarin in syntactic structure. The effectiveness of the proposed Hakka language modeling approaches will be examined by simulations using a Hakka speech corpus.

Besides, to accomplish the mixed Hakka-Mandarin speech recognition task, a shared phone set is required. In this study, the shared phone set is obtained by the log-likelihood measurement method [5,6] using a large-size Mandarin speech corpus and a median-size Hakka speech corpus. Then, a standard decision-tree based context clustering of HMMs is conducted to construct a set of state-clustered, cross-word tri-phone HMMs. As for the LM used in this bilingual speech recognition task, we simply extend the Hakka LM utilized in the Hakka speech recognition task to include some frequently-used Mandarin words. Its performance will be evaluated by simulations using a mixed Hakka-Mandarin speech corpus with its texts being mainly Hakka words embedded with one or two Mandarin words. We believe that the proposed approach can be applied to speech recognition of other dialects of Chinese in Taiwan, like Min-Nan, Hoi-Liuk Hakka, and Ta-Pu Hakka, etc.

The paper is organized as follows. Section II presents the speech and text corpora used in this study. The Hakka speech recognition system is described in Section III. In Section IV, the study on mixed Hakka-Mandarin speech recognition is discussed. Some conclusions are given in the last section.

II. DATABASE

This section briefly describes the database used in the study. It contains two parts: the speech corpus and the text corpus.

A. Speech corpora

The experimental speech corpora are all recorded in reading style, including the Si-Rhan Hakka speech corpus, the TCC300 corpus, and a mixed Hakka-Mandarin speech corpus. All the speech data are recorded through microphone and digitized at 16kHz sampling rate with 16-bit resolution. The Si-Rhan Hakka speech corpus is used to train the HMM-based acoustic models for both Hakka and mixed Hakka-Mandarin speech recognitions. It consists of 110 utterances spoken by 92 speakers with its texts being selected from the book 'The Taiwan Hakka Literature Collections' [7]. The TCC300 is used in the acoustic models training for the mixed Hakka-Mandarin speech recognition. It is generated by 300 speakers including 150 females and 150 males. Each speaker

reads 37 utterances in average. The mixed Hakka-Mandarin speech corpus is used to conduct a test of mixed Hakka-Mandarin speech recognition. It is generated by 19 speakers and each speaker uttered 40 utterances in average. Its texts are mainly written in Hakka and embedded with Mandarin words in a ratio of 2206/738. Table 1 summarizes the main information about the speech corpora mentioned above.

Table 1. SUMMARY OF SPEECH CORPUS

Corpus		No. of Speakers	No. of Syllables
Hakka (H)	Training	68	134,612
	Test	24	19,388
Mandarin (M)	Train	300	332,708
Mixed H-M	Test	19	4,710

B. Text corpora

The text corpora used in this study can be divided into two parts: Mandarin and Hakka. The Hakka part consists of the following three types of texts:

- *TextA*: It contains many short articles collected from Internet. It is word-segmented by an automatic word matching algorithm [8] with the *Dict* lexicon discussed below.
- *TextB*: It is a Hakka-Chinese parallel-text corpus. It is constructed by firstly selecting Hakka articles from the book ‘The Taiwan Hakka Literature Collections’ [7], then tagged into word and POS sequences, and lastly translated into Chinese text by a Hakka expert.
- *Dict*: It is a 43,521-word Hakka lexicon. Each word entry is associated with the pronunciation and POS tag. About 26,000 entries have also Chinese translation.

The Chinese corpus (*TextM*) is a large database with texts collected from Sinorama magazine [9] and the CIRB [10]. CIRB is a database built for information retrieval competition. Table 2 summarizes the main information about the text corpora.

Table 2. SUMMARY OF THE TEXT CORPORA

Corpus	No. of Words	No. of word entries
<i>TextA</i>	143,341	15,115
<i>TextB</i>	46,581	8,068
<i>Dict</i>		43,521
<i>TextM</i>	117,163,131	591,825

III. HAKKA SPEECH RECOGNITION SYSTEM

A. Acoustic model

The acoustic model (AM) is trained using the Hakka speech corpus. It is composed of 87 Right-context-dependent (RCD) 3-state initial HMM models and 71 5-state final HMM models. Each state is a 32-mixture GMM. Acoustic feature vector used in this study consists of 38 components (12 MFCC parameters, and their first and second order time derivatives, and energy parameter’s first and second order time derivatives), which is analyzed at a 10-msec frame rate with a 32-msec Hamming window size. Cepstrum Mean

Normalization (CMN) is employed to compensate the bias of channel and/or speaker. Table 3 summarizes the main information about the AM.

Table 3. SUMMARY OF THE HAKKA AM

Classification	No. of models	No. of states	No. of mixtures
Initial	87	261	8,352
Final	71	355	11,360
Silence	1	3	96
Short pause	1	1	32

B. Baseline language model

The baseline language model (LM) is a conventional n-gram model. An n-gram LM uses the last n-1 words of the history as its sole information source. Thus a bigram predicts w_i from w_{i-1} by

$$P_{bsl}(w_i | w_{i-1}) = \frac{C_{TextA}(w_{i-1}, w_i)}{C_{TextA}(w_{i-1})} \quad (1)$$

where $C_{TextA}(\bullet)$ denotes the count of samples in *TextA*.

In a particle realization, Katz’ backoff smoothing method [11] is adopted to obtain a more robust bigram LM. Due to the small size of *TextA*, most bigrams are backoffed to unigrams. Only 19,213 bigrams are estimated from *TextA*. Table 4 displays the performance of the baseline Hakka speech recognition system.

Table 4. PERFORMANCE OF THE BASELINE SYSTEM ON THE HAKKA TEST CORPUS

LM	Accuracy (%)		
	Word	Character	Syllable
Free grammar	21.44	42.67	57.97
Baseline	40.79	58.99	66.29

C. Improvement on LM

Since the data of *TextA* is not sufficient to construct a robust bigram language model, some useful syntactic and Hakka-Chinese translation information is incorporated to help to construct a robust Hakka LM. The proposed improved LM is formulated by

$$P(w_i | w_{i-1}) = \begin{cases} P_{bsl}(w_i | w_{i-1}), & C_{textA}(w_{i-1}, w_i) \geq k \\ \alpha_1(w_{i-1})P_{pos}(w_i | w_{i-1}) \\ \quad + \alpha_2(w_{i-1})P_{trans}(w_i | w_{i-1}), & \text{otherwise} \end{cases} \quad (2)$$

where $P_{pos}(\bullet)$ is an LM using POS information and $P_{trans}(\bullet)$ is an LM using translation information.

1) Stage 1: Incorporating POS information from Hakka text corpus

We first use the POS information to construct a class-based LM. Each POS category is treated as a class. The model is formulated by

$$P_{pos}(w_i | w_{i-1}) = \sum_{G(w_i), G(w_{i-1})} P(w_i | G(w_i)) \cdot P(G(w_i) | G(w_{i-1})) \cdot P(G(w_{i-1}) | w_{i-1}) \quad (3)$$

where $G(w_i)$ denotes the POS of word w_i ; $P(G(w_i)|G(w_{i-1}))$ is a POS bigram model estimated directly from the manually-labeled Hakka text corpus (*TextB*); $P(w_i|G(w_i))$ and $P(G(w_{i-1})|w_{i-1})$ are word-POS and POS-word models describing the relationships between words and their possible syntactic roles. The two models are estimated from *TextB* and the Hakka dictionary (*Dict*). Although *TextB* contains well-annotated word and POS information, not all word-POS statistics can be directly obtained from it because it is relatively small to contain only 8,068 word entries. To obtain more robust estimations of word-POS statistics, the information stored in *Dict* is incorporated to assist in providing the word-POS information. Therefore, the POS-word model is estimated by

$$P(G(w_i)|w_i) = \begin{cases} P_{TextB}(G(w_i)|w_i), & C_{TextB}(G(w_i), w_i) \geq k \\ \lambda \frac{P_{Dict}(G(w_i)|w_i)P_{TextB}(G(w_i))}{\sum_{G(w_i)} P_{TextB}(G(w_i))}, & otherwise \end{cases} \quad (4)$$

where

$$\lambda = 1 - \frac{\sum_{w_i \in C_{TextB}(w_i, G(w_i)) \geq k} P_{TextB}(G(w_i)|w_i)}{\sum_{w_i} P_{TextB}(G(w_i)|w_i)};$$

$P_{TextB}(G(w_i)|w_i)$ is the POS-word statistics directly estimated from *TextB*; $P_{Dict}(G(w_i)|w_i)$ represents the POS-word statistics in *Dict*; and $P_{TextB}(G(w_i))$ is the unigram POS LM estimated from *TextB*. Since *Dict* contains only POS-word information, the POS-word statistics in *Dict* is given by a fair floor value set to be the inverse of the count of POS entries listed in *Dict* given the word. Then the word-POS model can be estimated by

$$P(w_i|G(w_i)) = \frac{P(G(w_i)|w_i)}{P_{TextB}(G(w_i))} \cdot P(w_i) \quad (5)$$

where $P(G(w_i)|w_i)$ is estimated by Eq. (4); and $P(w_i)$ is a word unigram LM estimated from *TextA* and *TextB*.

2) Stage 2: Incorporating POS information from Mandarin text corpus and Mandarin-Hakka bilingual dictionary

In this stage, the properties that Hakka and Mandarin Chinese are similar in their syntax and bilingual word translation information are utilized to construct a more robust Hakka LM. Since Hakka is a dialect of Chinese, they are very similar in the syntactic structure. To use the information, we assume that the POS bigram of Hakka is identical to that of Mandarin Chinese. Therefore, the parameter of POS bigram $P(G(w_i)|G(w_{i-1}))$ in Eq. (3) can be directly replaced by the one estimated from the Chinese text corpus (*TextM*).

Furthermore, to improve the estimation robustness of the POS-word model and the word-POS model shown in Eqs. (4) and (5), Hakka-Chinese word translation information in *Dict* is used. In *Dict*, 26,524 out of 43,521 Hakka words are mapped to 25,532 Chinese entries. After analyzing those translation pairs, we classify them into the following three cases:

➤ Case 1

Each Hakka word is mapped to one or more Chinese entries. Each Chinese entry is composed of one lexical word and can be found in the Mandarin LM. Some examples are given below:

Hakka	Mandarin translation entries	English
阿公	爺爺, 祖父, 老人家, 阿公	grandpa
阿啾箭	大卷尾鳥, 烏鶯, 黑捲尾, 大捲尾	a bird
哩	了, 囉, 啦, 哩, 呢, 呀	interjections
三十暗晡	除夕	New Year's Eve
八月半	中秋節	Moon Festival
工廠	工廠	factory
用具	用具	tool

➤ Case 2

A Chinese translation entry is composed of a sequence of words, a phrase, or a sentence. Those Chinese translations can not be directly found in the Mandarin LM. But after decomposing them into word strings, all their constituent words can be found in Mandarin LM. Some examples are given below:

Hakka	Mandarin translation entries	English
甲打半	一甲半 + 左右	about one and half acres
奈項	哪 + 一項	which item
正做	剛 + 做	just doing
其	他 + 的	his

➤ Case 3

Mandarin translation entries of this case are both similar to those of Case 1 and Case 2, but some of their (decomposed) words can not be found in Mandarin LM. Most of them are very low frequency words. Some examples are given below:

Hakka	Mandarin translation entries	English
蹩腳貨	蹩腳貨	junk
抹草	稀簽草	a plant

Table 5 lists the statistics of these three cases.

Table 5. SUMMARY OF MANDARIN TRANSLATION ENTRIES

Classification	No. of entries (%)
Case 1	14,636 (57.3%)
Case 2	9,487 (37.1%)
Case 3	1,419 (5.6%)
Total	25,532

Based on the above analysis, the improved POS-word model and the word-POS model can be obtained by

$$P^*(G(w_i)|w_i) = \begin{cases} \sum_{M_i} P_M(G(M_i)|M_i)P_T(M_i|w_i), & M_i \in \text{Case 1} \\ P(G(w_i)|w_i), & M(w_i) \notin \text{Case 1} \end{cases} \quad (6)$$

and

$$P^*(w_i | G(w_i)) = \frac{P^*(G(w_i) | w_i)}{P_M(G(w_i))} \cdot P^*(w_i) \quad (7)$$

where $M_i = M(w_i)$ denotes a possible Chinese translation of the Hakka word w_i found in *Dict*; $P_T(M_i | w_i)$ is the Hakka-to-Chinese word translation probability obtained by

$$P_T(M_i | w_i) = \begin{cases} P_{TextB}(M(w_i) | w_i), & C_{TextB}(M(w_i), w_i) \geq k \\ \lambda \frac{P_{Dict}(M(w_i) | w_i) P_{TextM}(M(w_i))}{\sum_{M(w_i)} P_{TextM}(M(w_i))}, & \text{otherwise} \end{cases} \quad (8)$$

$P_{TextB}(M(w_i) | w_i)$ represents the probability of w_i being translated as a Mandarin word $M(w_i)$ estimated from the Hakka-Mandarin parallel text corpus (*TextB*); $P_{Dict}(M(w_i) | w_i)$ represents the Mandarin-Hakka word pair statistics in *Dict*; and $P_{TextM}(M(w_i))$ is the unigram LM estimated from *TextM*. Since *Dict* contains only Mandarin-Hakka word translation pair information, the statistics $P_{Dict}(M(w_i) | w_i)$ is given by a fair floor value set to be the inverse of the number of possible Mandarin translation entries listed in *Dict* mapped with w_i . $P_M(G(M_i) | M_i)$ is the POS-word model of Mandarin Chinese that can be directly estimated from *TextM*. The resulting improved Hakka word unigram model is expressed by

$$P^*(w_i) = \begin{cases} P_M(M(w_i)), & M(w_i) \in \text{Case 1} \\ P_M(M(w_i) = M_1, \dots, M_n) \\ \approx P_M(M_1) P_M(M_2 | M_1) \dots P_M(M_n | M_{n-1}), & M(w_i) \in \text{Case 2} \\ P(w_i), & \text{otherwise} \end{cases} \quad (9)$$

where $P_M(M(w_i))$ and $P_M(M_n | M_{n-1})$ denote the Chinese unigram and bigram LMs estimated from *TextM*; and $P(w_i)$ is the unigram Hakka LM estimated in stage 1.

3) *Stage 3: Incorporating the information of Chinese word bigram model and Mandarin-Hakka bilingual dictionary.*

This stage utilizes word-by-word Hakka-Mandarin translation information given by *TextB* and *Dict*. The Hakka LM using word translation information $P_{trans}(\bullet)$ is formulated by

$$P_{trans}(w_i | w_{i-1}) = \begin{cases} \sum_{M_{i-1}^i} P_T(w_i | M_i) \cdot P_M(M_i | M_{i-1}) \cdot P_T(M_{i-1} | w_{i-1}), \\ C_{TextB}(w_{i-1}^i, M_{i-1}^i) \geq k \text{ and } M_{i-1}^i \in \text{Case 1} \\ P^*(w_i), & \text{otherwise} \end{cases} \quad (10)$$

where $P_T(w_i | M_i)$ and $P_T(M_{i-1} | w_{i-1})$ are Hakka-Mandarin and Mandarin-Hakka translation probability estimated from *TextB* (the latter is shown in Eq. (8)); $P_M(M_i | M_{i-1})$ represents the Mandarin bigram LM; $P^*(w_i)$ is the improved Hakka word unigram model shown in Eq. (9). It can be seen from Eq. (10) that, as more Hakka-Mandarin translation pairs occur in *TextB*, the Hakka bigram will relies more on the corresponding Mandarin bigram ($P_M(M_i | M_{i-1})$); otherwise it will be backoffed to the unigram $P^*(w_i)$.

D. Experimental Results

The performances of these three proposed approaches to construct the improved LM for Hakka speech recognition are evaluated using the Si-Rhan Hakka speech corpus introduced in Section II. Fig. 1 displays the experimental results. It can be seen from the figure that the recognition performance was gradually improved as more information is incorporated in a higher stage to assist in constructing the improved LM. Specifically, the word recognition rate increased 2.2% over the baseline system as we used the POS information of Hakka language in Stage 1. Another 1.5% increase was gained as we added the POS information of Chinese in Stage 2. Lastly, a small improvement of 0.1% was gained when we further incorporated the information of Chinese word bigram model through the Hakka-Chinese word translation. These results show that the reliable Chinese LM can be used to assist in constructing more reliable LMs for improving the Hakka speech recognition. We believe that this is mainly resulted from the property that Hakka and Chinese have very similar syntactic structure.

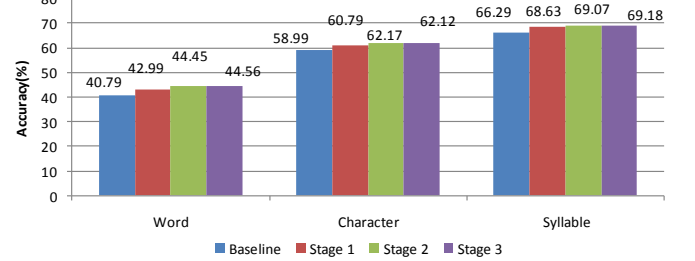


Figure 1. Comprison of language model performance.

It is noted that the improvement of Stage 3 is negligibly small. Some possible reasons are discussed as follows. First, only about 61% of Hakka words have the translated Chinese words in the Hakka-Chinese dictionary (*Dict*). Many frequent or important Hakka words may have no Chinese word mappings so that the abundant information of Chinese LM is nothing to do with them. Second, the text corpora of Hakka and Chinese contain different types of articles. The Chinese text corpus is large and general to contain plenty of types of articles, while the Hakka text corpus is small and contains articles of Hakka-specific historic stories. So, their information contents are not matched. We hence need to put more efforts on this issue in the future.

IV. A STUDY ON MIXED HAKKA-MANDARIN SPEECH RECOGNITION

In this preliminary study, the mixed Hakka-Mandarin speech corpus is used to evaluate the proposed acoustic and language modelings. Since the size of the mixed Hakka-Mandarin speech corpus is not large enough to construct a reliable bilingual acoustic model, both the Si-Rhan Hakka speech corpus and TCC300 Mandarin corpus are used to construct a more robust bilingual acoustic model. In language modeling, due to the fact that there is no sufficient mix Hakka-Mandarin text for LM training, the Hakka LM obtained in III.C is modified to include the Mandarin words embedded in the mixed Hakka-Mandarin speech corpus. These embedded Mandarin words are some frequently-used proper nouns or new words. The details of the acoustic model and the language model used in the mixed Hakka-Mandarin speech recognition are described in the following sections.

A. Bilingual acoustic model

1) Phone combination

To conduct a mixed Hakka-Mandarin speech recognition task, a suitable bilingual phone set is required. The bilingual phone set used in this study is created by combining 38 Mandarin phones and 38 Hakka phones using the log-likelihood measurement [5,6]. The distance (log-likelihood) between two 3-state left-to-right mono-phone models λ_i and λ_j is defined by:

$$D(\lambda_i, \lambda_j) = \log f(\bar{O}_i | \lambda_i) - \log f(\bar{O}_i | \lambda_j) \quad (11)$$

where $\bar{O}_i$ denotes a sequence of observations labeled as phone i ; and $f(\bar{O}_i | \lambda_i)$ is the probability density function (PDF) of the observations. Since the distance is not symmetric, the average distance can be calculated as follows:

$$D_{i,j} = \frac{1}{2} (D(\lambda_i, \lambda_j) + D(\lambda_j, \lambda_i)) \quad (12)$$

The combinations of Hakka-Mandarin phone pairs in this study must satisfy the following two principles: (1) one Hakka phone can only be combined with one Mandarin phone; (2) the phones to be combined are located in the same phonetic context in the syllable structure (i.e. *initial*, *medial*, *vowel*, or *coda*). After calculating the log-likelihood distance in Eq. (12), the top-22 nearest Hakka-Mandarin phone pairs are merged to form a 54-phone bilingual phone set. Among these 22 merged phone pairs, 14 phones are located at *initial*, 1 at *medial*, 5 at *vowel* and 2 at *coda*. Table 6 lists the top-10 nearest Hakka-Mandarin phone pairs.

2) Tri-phone model training

After obtaining the bilingual phone set (54 phones), a standard decision-tree based context clustering of HMMs is conducted to construct state-clustered, cross-word tri-phone HMMs. Three-state left-to-right HMMs are used to model each clustered tri-phone. The question set used for the decision-tree based context clustering is formed by using the left and right phonetic contexts. For comparison, mono-lingual

tri-phone models of Mandarin and Hakka are also constructed. Table 7 shows the statistics of these three tri-phone models.

Table 6. LOG-LIKELIHOOD DISTANCES OF TOP-10 NEAREST PHONE PAIRS

Rank	Distance	Hakka		Mandarin	
		model		model	
1	4.264	H_eng	coda	M_ng	coda
2	4.408	H_a	vowel	M_a	vowel
3	4.444	H_en	coda	M_en	coda
4	4.815	H_yi2	vowel	M_yi2	vowel
5	4.972	H_eh	vowel	M_eh	vowel
6	5.065	H_l	initial	M_l	initial
7	5.251	H_wu2	vowel	M_wu2	vowel
8	5.472	H_s	initial	M_s	initial
9	6.127	H_m	initial	M_m	initial
10	6.469	H_yi1	medial	M_yi1	medial

Table 7. THE STATISTICS OF THE TRI-PHONE MODELS (MIX: HAKKA-MANDARIN COMBINED TRI-PHONE MODEL, H: HAKKA TRI-PHONE MODEL, M: MANDARIN TRI-PHONE MODEL)

Model	Phone #	Question #	Logical state #	Physical state #	Total mixture #
Mix	54	454	31868	3291	52,627
H	38	196	15521	620	19,784
M	38	180	19806	3301	52,807

B. Language model

Each utterance of the mixed Hakka-Mandarin speech corpus is mainly a sentential Hakka utterance embedded with one or several Mandarin words. Two examples are given in the following:

頭擺人毋知麼安到 (網路遊戲)
(摩托車) 要買較大台正會較穩

where words in the bracket are Mandarin words and are delimited by spaces. To include Mandarin words in the mixed Hakka-Mandarin speech recognition task, the space of the Hakka word bigrams obtained in III.C is discounted for additionally modeling Hakka-Mandarin, Mandarin-Hakka and Mandarin-Mandarin word transitions. Therefore, the bilingual LM is formulated by

$$P_{bl}(B_i | B_{i-1}) = \begin{cases} P_{bl}(w_i | w_{i-1}) = (1-k)P(w_i | w_{i-1}) \\ P_{bl}(M_i | w_{i-1}) = \frac{k}{\|DictM\|} \\ P_{bl}(M_i | M_{i-1}) = \frac{k}{\|DictM\|} \\ P_{bl}(w_i | M_{i-1}) = (1-k)P(w_i) \end{cases} \quad (13)$$

where $B_i \in (DictM, DictH)$ can be a Mandarin word $M_i \in DictM$ or a Hakka word $w_i \in DictH$; $DictM$ and $DictH$ represent the Mandarin word set and the Hakka word set in the bilingual dictionary, respectively; k is the ratio of Mandarin word count to Hakka word count in the mixed Hakka-Mandarin speech corpus; $\|DictM\|$ represents the count of Mandarin word in the bilingual dictionary. In this

study, the bilingual dictionary contains 43,521 Hakka words and 738 Mandarin words.

C. Experiments and Results

The experimental results are displayed in Table 8 and Fig. 2. It can be found from Table 8 that the scheme using the bilingual mixed AM (Mix) performs much better than that using the directly-combined AM (H+M) for both bilingual LMs modified from the two Hakka LMs derived in Stages 2 and 3. We also find from Fig. 2 that both Hakka and Mandarin words are improved by using the mixed AM. Moreover, the improvement is more significant for Mandarin words.

Table 8. PERFORMANCES ON BILINGUAL CORPUS BY LANGUAGE MODEL

Accuracy (%)				
AM	LM	Word	Character	Syllable
Mix	Stage 2	37.16	54.33	56.24
	Stage 3	37.84	55.27	56.14
H+M	Stage 2	21.33	34.86	38.09
	Stage 3	21.98	35.61	38.47

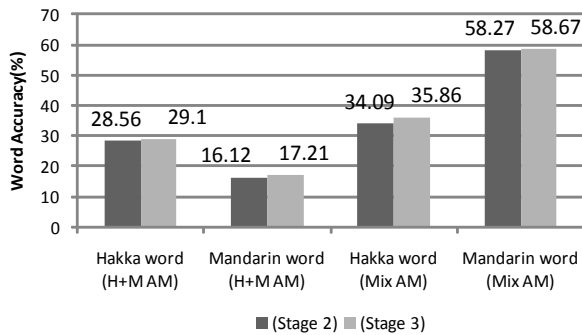


Figure 2. Comprison of language model performance.

Table 9 displays the analysis on syllable accuracy and inter-language syllable confusions. It can be seen from the table that the use of Mix AM can greatly improve the accuracy of Mandarin syllables and reduce the confusion of recognizing Mandarin syllables as Hakka syllables ($M \rightarrow H$). Although the accuracy of Hakka syllables by using the Mix AM remains the same as the one using the H+M AM, the accuracy of Hakka words is significantly improved (as shown in Figure 2). This is probably because the reduction of $M \rightarrow H$ syllable confusion makes less insertion errors of Hakka words. This result reconfirms the effectiveness of the proposed Mix AM modeling.

V. CONCLUSION

In this paper, a first study on Hakka and mixed Hakka-Mandarin speech recognition has been discussed. Several methods of language modeling have been proposed to help to solve the problem of lacking large text corpus in Hakka SR. For mixed Hakka-Mandarin SR, a new method of training a single mixed AM was discussed. Experimental

results confirmed the effectiveness of these proposed methods. Further study to use more Chinese language information to assist in improving the Hakka LM is worth doing in the future.

Table 9. ANALYSIS ON SYLLABLE ACCURACY OF HAKKA (H), MANDARIN (M), AND INTER-LANGUAGE CONFUSIONS OF $H \rightarrow M$ AND $M \rightarrow H$ IN THE BILINGUAL SPEECH RECOGNITION.

AM	LM	% (count)			
		H	M	$H \rightarrow M$	$M \rightarrow H$
H+M	Stage 2	55.15 (1,782)	17.78 (263)	0.80 (26)	79.31 (1,173)
	Stage 3	55.15 (1,782)	19.13 (283)	0.84 (27)	77.76 (1,150)
Mix	Stage 2	55.15 (1,782)	64.16 (949)	1.80 (58)	32.86 (486)
	Stage 3	55.15 (1,782)	64.64 (956)	1.86 (60)	32.45 (480)
Total syllable number = 3,231 (H) + 1,479 (M) = 4,730					

ACKNOWLEDGMENT

This work was supported by NSC of Taiwan under contract NSC98-2221-E-009-075-MY3. The authors would like to thank Council for Hakka Affairs Executive Yuan, Republic of China (Taiwan), Ministration of Education, Taiwan, and College of Hakka Studies, National Chiao Tung University, Taiwan, H. M. Yu and W. Z. Gong for supplying the lexicon and the text corpus.

REFERENCES

- [1] Hsiu-Min Yu, Lie-Shih Lo, Hsin-Te Hwang, Hsi-Chun Hsiao, and Sin-Hong Chen, "On Constructing Speech Corpus for Implementing Hakka Text-To-Speech Synthesis," OCOCODA, Dec. 2007.
- [2] Dau-Cheng Lyu, Chun-Nan Hsu, Yuang-Chin Chiang and Ren-Yuan Lyu, "Acoustic Model Optimization for Multilingual Speech Recognition," Computational Linguistics and Chinese Language Processing, Vol. 13, No. 3, September 2008, pp. 363-386.
- [3] Hakka Television Service Online, available at <http://www.hakkatv.org.tw/>
- [4] Formosa Hakka Radio Online, available at <http://www.formosahakka.org.tw/>
- [5] J. Köhler, "Multilingual phone models for vocabulary independent speech recognition task," Speech Communication, Vol. 35(1-2), pp. 21-30, 2001.
- [6] B. H. Juang, L. R. Rabiner, "A probabilistic distance measure for hidden Markov model," Bell System Technical Journal, 64(2), pp. 391-408, 1985.
- [7] Wan-Zao Gong, The Taiwan Hakka Literature Collections (Global, Taipei, Taiwan, 2004), in Chinese/Hakka.
- [8] Chen, Keh-jiann, Shing-Huan Liu, "Word Identification for Mandarin Chinese Sentences," Proceedings COLING '92, pp.101-105, Nantes, France, 1992
- [9] Sinorama magazine, available at <http://dblink.ncl.edu.tw/sinorama/index.htm>
- [10] Chen, Kuang-hua and Chen, Hsin-His, 2004, Overview of CIRB030 Information Retrieval Test Collection, available at <http://lips.lis.ntu.edu.tw/cirb/index.htm>.
- [11] S.M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recogniser", IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-35, vol. 3, pp 400-401, March 1987.

客語 ㄞ時貌之間的關係探討*

余秀敏

中華大學語言中心

kuo@chu.edu.tw

摘要

本研究旨在探討客語 ㄞ所表現的多重動詞時貌語意之間的關係。本文以一小型語料為基礎，蒐集並分析出現在不同時貌構式中的 ㄞ的語意，我們認為 ㄞ具有類似客語動後小稱詞 仔 的降低程度的意義、或是等同國語動量詞 一下 的短時、弱勢語的語意。因此，暫時貌構式[動詞_i(動詞_i) ㄞ⁵⁵哩²¹/ㄞ³¹]的 ㄞ表達一點、稍微的進行某一動作(見1)、嘗試貌構式[試-動詞-看-ㄞ⁵⁵哩²¹/ㄞ³¹]的 ㄞ表達淺程度的嘗試某一行動(見2)、重複貌構式[動詞_i ㄞ⁵⁵動詞_i仔]中的 ㄞ呼應第二個動詞之後的小稱詞 仔 (見3)，以及緊密先前貌構式[動詞-ㄞ⁵⁵-補語]中 ㄞ⁵⁵則從短時的核心語意轉化為兩事件之間的時間近距離的關係意義 (relational meaning)：在單事件句子中，ㄞ表達‘剛剛’之意，是某一過去事件與說話時刻之間的緊密關係(見4)，在雙事件句子中，則表達‘一...就...’之意，強調兩個連續事件之間的時間距離很近(見5)，正由於這種關係，在相關雙事件出現頻率很高的語用情境中，雙事件中的第一個事件很容易被詮釋成造成第二事件的條件，此時，很容易發展出習慣性的語意(見6)，這也就能解釋為何客語子句動 ㄞ 著 會 熟語化為描述令人不悅的習慣性行為，類似英語出現進行式中的 *always* 的動不動、老是的負面語意(見7)。

- 1) 該隻細妹仔恁讚，毋行兼 ㄞ，斯會分人追走喔！
那位女孩很棒，(你若)不(和她)走近一點，(她可是)會被人追跑囉！
- 2) 毋怕哪！去試食看 ㄞ，又無喊你買！
沒關係啦！去試吃看看，又沒叫你買！
- 3) 佢跌倒以後，這下行路跛 ㄞ 跛仔。
他跌倒以後，現在走路跛 ㄞ 跛的。
- 4) 你看！日頭出 ㄞ 來，還靚哪！
你看！太陽剛出來，真美啊！
- 5) 佢去 ㄞ 到台北，就會打電話分你。
他一去到台北，就會打電話給你。
- 6) 佢堵 ㄞ 著人，就愛搭人个肩頭。

*本文是受國科會計畫《客語文句轉語音及語音辨認之研究》(96-2221-E-009-030-MY3)之補助完成的。

他一遇到人，就喜歡搭人家的肩膀。

- 7) 阿旺伯動啊著就怨人,沒人愛同佢行兼
阿旺伯老是報怨他人，沒人喜歡和他走近。

關鍵詞：客語時貌、客語啊、暫時貌、嘗試貌、反覆貌、緊密先前貌

In this paper we concern ourselves with the multiple aspectual functions denoted by the Hakka aspect marker $-a^{55}$ -啊, which widely distributes in the delimitative construction, the tentative construction, the iterative construction, and the immediate anterior construction. Based on a mini corpus and the cross-aspectual analyses on $-a^{55}$ -啊, we propose that there is a semantic core shared by $-a^{55}$ -啊 in the above-mentioned Hakka aspectual constructions, which can be defined as *scaling the extent of an action down*, a definition like that indicated by the Hakka postverbal diminutive *e 仔*, or like the meaning of Mandarin verbal classifier *yi-xia* 一下, referring to a short period of time and pragmatically serving as a downtoner. With this essential meaning, $-a^{55}$ -啊 expresses the meaning of *doing something a little bit* in the delimitative construction $[\text{verb}_i(\text{verb}_i)-a^{55} \text{le}^{21}/-a^{31}]$ as in (1), conveys the meaning of slightly engaging tentative actions in the tentative construction $[\text{try-verb-see-}a^{55} \text{le}^{21}/-a^{31}]$ as in (2), refers to repeated actions by collocating with a reduplicated verb and the classical Hakka diminutive *e 仔* in the iterative construction of $[\text{verb}_i-a^{55}-\text{verb}_i-e^{31}]$ as in (3), and functions as an immediate anterior marker in the construction of $[\text{verb-}a^{55}-\text{complement}]$ where the core meaning of $-a^{55}$ -啊 has been metaphorically transformed from a short period of time into a tight temporal relational meaning between two events: in the single-event sentence it profiles an immediate relationship between the moment of speech and the event time indicated by the sentence, generating the meaning of *just*. See example (4). Nonetheless, in a complex-event sentence with the first event clause containing the construction of $[\text{verb-}a^{55}-\text{complement}]$, a^{55} -啊 indicates an immediate sequential order of the two events. Example (5) illustrates. Also, due to the temporally close relationship denoted by $-a^{55}$ - between two events, a pragmatic habitual reading is derived from contexts with the first event interpreted as the condition under which the second event occurs as shown in (6). This explains why the clause *tung-a-do* ‘動啊著 *as soon as one moves*’ has been idiomatized into a phrase bearing the meaning of English *always* in progressive sentences, indicating one’s annoying habitual behavior. See (7).

- (1) *gie-zag se-moi-e an-zam, m hang-giam-a^{55}, sii voi bun-ngin giug-zeu o!*
The girl is wonderful. If you are not in close contact with her, she’ll be chased away.
- (2) *m pa na! hi cii-siid-kon-a^{55}, iu mo hem ng mai!*
Go ahead! Just have a taste for it. You are not forced to purchase it.

- (3) *gi died-do i-heu, ia-ha hang-lu bai-a⁵⁵-bai-e.* .
After falling over, he is now hobbling around.
- (4) *ng-kon! ngid-teu cud-a⁵⁵-loi, han ziang na!*
Look! The sun has just risen. What a beautiful sight!.
- (5) *gi mug-zu sab-a⁵⁵-log, ziu soi-ted le*
As soon as his eyes closed, he fell asleep (the next minute).
- (6) *gi du-a⁵⁵-do ngin, ciu oi dab ngin-ne gien-teu .*
Whenever he meets someone, he likes to put his arm around his/her shoulder.
- (7) *A-vong-baf tung-a⁵⁵-do ziu ien ngin, mon ngin oi tung gi hang giam.*
Uncle A-Vong is always complaining. No one likes to be near to him.

Key words: Hakka aspect markers, Hakka *a⁵⁵*, delimitative, tentative, iterative, immediate anterior

1. 引言

根據前人的研究，客語各次方言都有分佈很廣、語意豐富、聲調多元的啊（見賴維凱 2006 的整理），例如作為句首感嘆詞讀為中降調的啊³¹，如

- 1) 啊³¹！壞特哩！毋記得帶遮仔¹。

啊！糟糕了！不記得帶傘。

作為句尾助詞、強化語者各人看法的語氣詞的升調啊²⁴（見例 2），以及表停頓的低降調的啊²¹（見 3、4）：

- 2) 我就講啊²⁴！無愛去較好啊²⁴！
我就說嘛！不要去比較好嘛！
- 3) 有聽老人家講過：人啊²¹，毋使恁慶腳，懵懵懂懂乜過一生²。
我曾聽老人說過：人啊，不須太精明，懵懵懂懂也過一生。
- 4) 你讀書恁毋認真，係考得著高中啊²¹，日頭都會對西片出來哩³！
你讀書那麼不認真，如果考得上高中的話，太陽都會打西邊出來哩！

此外，梅縣（林立芳 1997）、六堆（賴維凱 2006）的客語次方言中，讀高平調的啊⁵⁵也做介詞用，功能像北四縣的介詞到，後接時間、處所、方位詞，如：

- 5) 佢同電話號碼寫啊⁵⁵紙項⁴。

¹選自《客語實用手冊》：100。

²選自《客語能力認證基本詞彙-中級、中高級暨語料選粹》：9-152。

³選自《客語能力認證基本詞彙-中級、中高級暨語料選粹》：2-174。

他把電話號碼寫在紙上。

另外，在重複動詞之間加入啊形成[V 啊 V]的構式，往往是強調某狀態的程度（見 6、7），如：

- 6) 佢汝生來恁呢無人緣啊⁵⁵無人緣！醜醜啦⁵！
你生來很沒人緣，醜醜的。
- 7) 佢摺間項捋到淨利啊⁵⁵淨利。
他將坊間整理得十分乾淨。

除此之外，研究客語啊的學者都提到，客語的啊字有作為動詞時貌的功能，如讀中降調的啊³¹有扮演動詞暫時貌（deliminatives，見例 8）和嘗試貌（tentatives，見例 9），高平調的啊⁵⁵也和國語的[V 啊 V 的]結構一樣有[V 啊 V 仔]（中研院詞庫小組，2006：4）的說法，此時，此時啊⁵⁵表示動作的反覆時貌（見例 10）；以及讀成高平調啊⁵⁵的緊密先前貌（immediate anterior，見例 11）：

- 8) 新正年時，乜係行春時節，大家來去客家莊邊邊啊³¹⁶。
新春時候，也是踏春時節，大家來去客家莊走一走。
- 9) 先生講：用客語來寫文章，毋會當難，試寫看啊³¹毋使驚⁷。
老師說：用客語寫文章，並不難，可以試寫看看不必害怕。
- 10) 佢昨晡日打籃球撓著腳，行路跛啊⁵⁵跛仔⁸。
他昨天打籃球扭傷腳，走路一跛一跛地/跛啊跛的。
- 11) 落雨過啊⁵⁵忒，山頂个路滑溜溜仔，行路愛細義⁹。
剛下完雨，山路滑溜溜地，走路要小心。

儘管前人的研究都有指出啊在這四種時貌中出現，但卻無人深究其中的語意關連性，因此本研究旨在探討以上四個時貌之間的語意關係，我們將指出：暫時貌和嘗試貌的啊³¹其實是高平調啊⁵⁵在句尾的變體，這些啊⁵⁵的核心語意都有具短時和降低某個動作的程度的語意，並由此核心語意透過隱喻的轉化，表達兩個事件之間時間上的緊密性。

⁴選自賴維凱 2006：46，但這種用法北四縣沒有。

⁵選自賴維凱 2006：49。

⁶選自《客語能力認證基本詞彙-中級、中高級暨語料選粹》：4-33。

⁷選自《客語能力認證基本詞彙-中級、中高級暨語料選粹》：14-99。

⁸選自《客語能力認證基本詞彙-中級、中高級暨語料選粹》：8-114。

⁹選自《客語能力認證基本詞彙-中級、中高級暨語料選粹》：26-202

2. 文獻回顧

由於客語^啊的出現頻率很高，因此很容易吸引客語研究者的注意，如引言中的例子，若不論聲調的差異，^啊可作為句首感嘆詞、表各種語氣的句尾助詞、後接時間、地點的介詞、動後的時貌詞等等，然而不同的客語次方卻呈現出不同的^啊的功能，例如梅縣（林立芳 1997）和六堆（賴維凱 2006）的客語次方言的^啊具有後接地方詞、處所詞的介詞功能，但以苗栗客語為代表的北四縣次方言則沒有這種用法。

若針對^啊所表達的時貌義來看，各次方言也有不同的構式，研究者之間也使用不同的分類術語，例如林立芳（1997：135-138）舉出的^阿¹⁰的時貌義有：（一）出現在短時貌的構式中，a類短時貌構式為[V 阿欵]¹¹，^{阿欵}共現的結構表達「稍微做一下的意思」¹²，b類短時貌構式為[V+阿+趨向/結果補語]¹³，此時^阿是表示「強調動補結構實現的時間短暫，而且緊接著就發生另一動作或情況」¹⁴，（二）嘗試貌，有多種構式，其中一種和短時貌一樣為[V 阿欵]¹⁵、另一種構式為[V 阿勝/阿勝欵]¹⁶，（三）是出現在反覆貌的構式[V₁ 阿 V₁ 欵 V₂]¹⁷，語意等於國語的V₁著V₁著，最後（四）是出現在祈使句中的連續貌構式[緊V 阿分佢]¹⁸，其中的^{阿分佢}沒有實質的語意。

此外，以長汀、梅縣、和五華等客語次方言的例句為基礎，饒長溶（2001）則將客語^啊的時貌義分為五類，如：（一）剛然貌、（二）方式持續貌、（三）著落持續貌、（四）間接持續貌、以及（五）直接持續貌。若比較林立芳和饒長溶的分類和定義，我們發現同樣的^啊的時貌義卻以不同的術語來描寫的地方是：饒長溶的「剛然貌」相當於林立芳的b類「短時貌」的[V+阿+結果補語]構式、其「方式持續貌」則等於林立芳的b類「短時貌」的[V+阿+結果補語]；其「間接持續貌」和「直接持續貌」都指向林立芳[V₁ 阿 V₁ 欵 V₂]的反覆貌，只是饒長溶的客語次方言將構式中的「欵」讀成「哩」。這兩位作者沒有重複的地方是林立芳描述的客語次方言中的^啊有表達嘗試貌和出現在[緊V 阿分佢]的連續貌構式之中，這是饒長溶所研究的次方言沒有的現象；但饒長溶的著落持續貌構式[V + 啊 + 在/ 到 + N_處(+ V₂)]¹⁹，根據饒長溶的說法，該構式可表示「動作完成後，著落某處的持續狀態」，並認為此時^啊表達類似北京話的「了」。

另外，賴維凱（2006）以南四縣代表六堆次方言語料為基礎的研究，針對^啊

¹⁰ 為尊重原作者的寫法，以^阿代表^啊。

¹¹ 林立芳的例句之一：我來求阿欵阿叔，看你有辦法無？（我來求裘叔叔，看看你有沒有辦法？）

¹² 出處同上：136。

¹³ 例句之一：新娘行阿入屋就端茶端飯。（新娘一走進家門就送茶送飯）

¹⁴ 出處同上：136。

¹⁵ 例句之一：大齊家先聽阿欵再來做。（大家先聽一聽再動手幹。）

¹⁶ 例句之一：汽車分我開阿勝欵。（汽車讓我開一開試試看。）

¹⁷ 例句之一：阿三古食糖子，嚼阿嚼欵，嚼倒舌嫲。（老三吃著糖果，嚼著嚼著，嚼到了舌頭。）

¹⁸ 例句之一：你緊背阿分佢，唔會背唔識个。（你不停地背，不會背不出來的。）

¹⁹ 例句之一：一陣風吹啊來，紙撒啊到一地。（一陣風吹了來，紙散到了一地。）

所表現的時貌義方面，除了舉出同以上兩位作者的短時貌義（[V啊補語]：表達剛剛及一...就...的意思）之外，他還列出嘗試貌的構式為[V看啊^{55/31}]，這和林立芳的嘗試貌構式[V阿欸]和[V阿勝/阿勝欸]不同；此外，還有表達像國語V—V時貌義的[VV啊]、以及表示「剛完成某一連續動作」的啊構式[V₁啊V₁]，該構式表達的意思和林立芳的重複貌以及饒長溶的間接和直接持續貌十分類似，只是賴維凱的構式中沒有最後的「欸/哩」。

至於以苗栗次方言為基礎的啊研究，都指出啊中插在動補結構之中，有表示‘剛剛’或‘一...就...’的意思（羅肇錦 1995；張雁雯 1998）。其中張雁雯認為：啊在此位置其實是一種停頓，「藉由停頓，表現動作生動的效果」並「強調動作的進行」²⁰；此外，該作者也提出啊在 V 啊^{55e31}結構中表示嘗試貌的語意，這和林立芳的[V 阿欸]的嘗試貌構式一樣。

由前人的研究我們可以觀察到，（1）啊的時貌義在不同客語次方言之間有同亦有異，（2）就算是同一個時貌義，不同研究者也各自有一套術語來定義啊的時貌義，以及（3）除了賴維凱運用了李文光等編輯的《李文谷再現江湖》的劇本作為語料分析的基礎之外，大部分研究啊的論文，都依照作者的客語語感所想到的單句話或詞組來舉例、並且都是以條列式的方式一一舉出啊的各種語意。然而，個人語感容易有分歧的感受、而語意要素卻又往往受限於例句的影響而呈現紛歧的歸類。為突破前人研究的句現，本研究先將語料腔調限定在苗栗次方言—北四縣客語的代表—，並以小型語料庫所收集到 342 個例句來作為研究的基礎，以避免因個人語感造成可能的不周延或偏見的產生，並將焦點放在由啊所引導出來的四種時貌義之間的關連。

3. 研究方法

基本上本研究的方法是以苗栗四縣客語的小型語料庫為基礎，為呈現客語啊的四個時貌義：暫時貌、嘗試貌、重複貌、和緊密先前貌之間的關連性，首先即是收集相關的語料來建構四者之間的關係，語料來源有四：龔萬灶先生的《阿啾箭个故鄉》和《客語實用手冊》、以及網路上的兩項官方的資源：行政院客委會提供的《客語能力認證基本詞彙-中級、中高級暨語料選粹》以及教育部國推會編輯的《台灣客家語常用詞辭典》；若需要額外補充例句，作者將自行編造句子，以支持一些論點。

4. 客語啊的四種時貌義的分析

根據本文的統計，四種啊的時貌義之例句數的分佈是：暫時貌 103 句、嘗試貌 47 句、重複貌 20 句、以及緊密先前貌 172 句。我們先在 4.1 節推論啊的

²⁰張雁雯 1998：44。

本調應為 55 高平調，而且在這些時貌構式中的 *啊* 的核心語意是暫時貌的降低某動詞的進行的程度，類似客語動詞後的小稱詞 *仔* 表達的功能(鍾榮富 2005)。此外，透過隱喻的方式將淺程度之意轉化為兩事件之間的緊密時間關係。

4.1 暫時貌的分析及 *啊* 的本調的釐清

暫時貌 (deliminative) 根據語言學和語音學字典的定義²¹是指某一動詞的狀態或活動的期間很短，因此是描述某動詞進行的程度被降低，而是稍微、一下而已。國語的暫時貌構式為[V 一 V] (Li & Thompson 1982: 214)或是[V 一下]，語用上可以做為緩和口氣的弱勢語 (謝方怡 2007)。從我們收集到的語料中，表達暫時貌義的例句有 103 句，其構式為[動詞_i (動詞_i) 啊⁵⁵ 哩²¹/啊³¹]，動詞之後可以是中降調的 *啊*³¹ (見例 12、13)、另一種則是加高平的 *啊*⁵⁵ 之後再加 *哩*²¹ (同國語的句尾助詞-了，讀做 le²¹，見例 14、例 15)，動詞可重複也可不必：

- 12) 甜板冰到硬殼殼仔，拿出來餹餹啊³¹²²。
年糕冷藏得硬梆梆地，拿出來蒸一蒸吧！
- 13) 食東西過後牙齒愛磋淨啊³¹，較毋會蛀牙、發牙包²³。
吃東西過後牙齒要刷乾淨，較不會蛀牙、牙齦發炎。
- 14) 愛又毋鬥緊啊⁵⁵哩²¹，分人搶走正來眼扁扁仔²⁴。
想要又不快一點，被人搶走了才眼巴巴的不甘心。
- 15) 外背毋知做麼个恁吵，ngai 去看啊⁵⁵哩²¹。
外面不知為什麼這麼吵，我去看一下。

我們認為 *啊*³¹ 其實應該讀成高平調 *啊*⁵⁵，但因為位處句尾，容易受到句尾聲調下降效應的影響，而讀成降調，這可由以上同樣表暫時貌的 *啊*⁵⁵ 哩²¹ 看出，*啊*⁵⁵ 哩²¹ 中的 *啊* 由於不在句尾，因此保留了高平調的特質；此外，另一個證據是 *啊* 也可修飾狀語，若 *啊* 不在句尾則讀成高平調 (見例 16、17)、若在句尾則讀成降調 (見 18 例)：

- 16) 分人請愛早啊⁵⁵去，過晝鉢lat正到，人食飽嘍²⁵！
出門作客要早點兒去，過了中午才到的話，人家都吃飽飯啦！
- 17) 喊佢遽啊⁵⁵來，佢半晝前後正到，實在還無下落。

²¹ Crystal, David. 2006. *A Dictionary of Linguistics and Phonetics*. 5th ed. Malden: Blackwell Publishing.

²² 選自《客語能力認證基本詞彙-中級、中高級暨語料選粹》：10-256。

²³ 選自《客語能力認證基本詞彙-中級、中高級暨語料選粹》：7-52。

²⁴ 選自《客語能力認證基本詞彙-中級、中高級暨語料選粹》：9-162。

²⁵ 選自《客語能力認證基本詞彙-中級、中高級暨語料選粹》：2-163。

叫他快點兒來，他十點左右才來，實在很不經心。

- 18) 阿義仔，喊你睡當晝，你就走去搞水，皮弓緬**啊**³¹，等等分若姆修理²⁶。
阿義，叫你睡午覺，你就跑去玩水，皮拉緊些，等著給你媽媽修理。

以上例句的**啊**分別修飾狀語‘早’、‘遽’（快）、和‘緬’（緊），都表示一些、一點之意，同修飾動詞時一樣，處於句中則讀為高平調、句尾則為中降調。

4.2 嘗試貌的分析

語料中，客語的嘗試貌構式的完全形式為[試-動詞-看-**啊**⁵⁵-**哩**²¹/**啊**³¹]，共有47句。該構式基本上和暫時貌的構式一樣，只是暫時貌的動詞，在嘗試貌中是[試-動詞-看]，其中動詞**試**或**看**之一以及句尾助詞**哩**可以省去，但中降調的**啊**³¹卻一定要保留，例如：

- 19) 該細人仔毋知做麼个，噉到抽骸另命，過去試探看**啊**⁵⁵**哩**²¹²⁷。（全形）
那個小孩子不知為什麼，哭得聲嘶力竭，過去探一探吧。
- 20) 先生講：用客語來寫文章，毋會當難，試寫看**啊**³¹，毋使驚²⁸。（省去**哩**）
老師說：用客語寫文章，並不難，可以試寫看看不必害怕。
- 21) 阿忠仔，你去鄉公所問看**啊**³¹，農地休耕个錢撥下來旨²⁹？（省去**試**、**哩**）
阿忠，你到鄉公所問問看，農地休耕的錢撥下來沒？
- 22) 汝去試食**啊**³¹！異像異好時樣哩。（省去**看**、**哩**）
你去試吃看看，好像很好吃似的。

從以上的語料看來，雖然嘗試貌的嘗試語意主要是由**試**或**看**貢獻的，但**啊**⁵⁵**哩**²¹或**啊**³¹基本上都要出現，若將以上例子的**啊**³¹省去，這些句子將會被四縣客語的直覺判斷為有問題的句子。如：

- 23) ？該細人仔毋知做麼个，噉到抽骸另命，過去**試看哩**。
- 24) ？先生講：用客語來寫文章，毋會當難，**試寫看**，毋使驚。

²⁶選自《客語能力認證基本詞彙-中級、中高級暨語料選粹》：8-149。

²⁷原句為<該細人仔毋知做麼个，噉到抽骸另命，過去試看**啊**⁵⁵**哩**²¹>，選自《客語能力認證基本詞彙-中級、中高級暨語料選粹》：9-104，原句並無主動詞為**看**而非**探**，為了呈現嘗試貌構式全形，作者加入**探**字。

²⁸選自《客語能力認證基本詞彙-中級、中高級暨語料選粹》：14-99。

²⁹選自《客語能力認證基本詞彙-中級、中高級暨語料選粹》：8-163。

- 25) ? 阿忠仔，你去鄉公所問看，農地休耕个錢撥下來官？
26) ? 汝去試食！異像異好時樣哩。

作者認為在真實世界裡頭，當人們嘗試去作某行動時，往往不會一開始即大幅度或深度的去進行，而是初期只做一些、一點、或稍微的程度，因此，暫時貌很容易讀到嘗試貌的語意，因此我們可以推論，客語嘗試貌構式 [試-動詞-看-啊⁵⁵-哩²¹/啊³¹]⁵⁵的啊之所以一定要出現，是因為客語的嘗試貌不僅要求要表達出嘗試之意（如試、看之一要保留），同時也要兼具動作只是小幅度進行的語意，這一點客語的嘗試貌和國語以重複動詞[VV看]的嘗試貌表現十分不同。

4.3 重複貌的分析

重複貌的構式為[動詞|啊動詞|仔]，收集到的例句共20句。在第2節中有提到，林立芳和饒長溶為此時貌擬出構式為[V啊V欸/哩]，其中欸/哩都是客語的完成時貌標記，在我們收集的語料中有兩種寫法：諗及仔。但我們認為重複貌構式中應該是小稱詞仔、而非完成貌的欸/哩/諗。這是因為客語小稱詞仔的聲調有隨前變調的現象，仔出現在具有高調的陰平、去聲、和陽入調之後時，讀成中降調³¹，例如：箱仔、柚仔、鹿仔，在其他含低調的名詞之後則讀成低降調²¹，如：李仔、桃仔、桌仔。同樣的，出現在重複貌構式[動詞|啊動詞|仔]的仔，也呈現出隨前變調的情形，在含有高調的動詞之後讀成中將調³¹，如：

- 27) 冷鑊炒番豆一罇啊罇仔³⁰。

冷鍋炒花生，攪拌啊攪拌的。

在含有低調的動詞之後便讀成低降調 21，如：

- 28) 佢無牙齒，時東西該下糜啊糜仔。

他沒有牙齒，吃東西的時候，用牙齦磨啊磨的³¹。

因此，我們可以確認重複貌構式中最後的詞彙是小稱詞仔，也因為這個字的認定，我們便主張[動詞|啊動詞|仔]中的啊也是像仔一樣表達降低程度的語意，藉以前後呼應小稱詞仔。因此，啊的核心語意從暫時貌到嘗試貌到重複貌可以一以貫之。

4.4 緊密先前貌的分析

根據Bybee等人（1994）綜合前人的相關研究，所謂的緊密先前貌（immediate anterior）的先前貌（anterior）是指與當下有關的過去行動（a past action with current relevance），例如，英文的先前式的典型例子（摘自Bybee et al. 1994: 61）：

- 29) Carol has taken statistics. (so she can help us.)

³⁰ 取自《實用客語手冊》：143。

³¹ 改編自教育部常用辭典中的例句。

- 30) I've just eaten dinner. (so I don't want any more food.)

除了以上基本的定義之外，當先前貌配合情境因素並引入最新狀態、或最新消息時(new situation, hot news)，也具有強調的功能 (McCawley 1971; Anderson 1982)，如：

- 31) Mt. St. Helens has erupted again!
32) Nixon has resigned!

據此，Li及Thompson（1982）也認為國語句尾「了」同樣表達了與說話情境相關的事件，因此，也可以算是國語先前式的時貌。

而所謂的「先前」搭配「緊密」(immediate)是指先前的事件與當下的時間距離很短，類似國語‘剛剛’英語‘just’之意。在客語裡頭表示動詞緊密先前的時間意義的構式是[動詞-啊⁵⁵-補語]，其中動補結構中的補語（包括趨向詞、時貌及時相詞、和結果補語）貢獻了動詞的終結點（telicity）、讓動補結構表現出終結式（accomplishment）的時貌義，表示該動作已經完結（Vendler 1957, Levin 1993: 105-106）；而緊密義則由原來表短時的暫時貌啊⁵⁵透過隱喻的轉換，成為動詞事件結束時間和說話的當下的時間距離很短，因此貢獻了‘剛剛’之意，例如以下的對話：

- 33) A：若 爸 呢？
你的父親呢？
B：佢行出去咧²¹。
他走出去了。
B’：佢正行-啊⁵⁵-出去。
他剛走出去。

B 與 B’的差別在於，B 句是先前貌而已，以情態語助詞咧²¹（葉瑞娟 2000）結尾，表達父親先前已經出門，是造成現在不在家的原因，並不強調這個出去的動作與當下時間的距離很短；但 B’的時間意義，除了點出父親之前已經出去之外，更指出父親出門的時間與說話的當下很近。

除了，單一事件的句子之外，構式 [動詞-啊⁵⁵-補語]也出現在雙事件的複雜句子當中，如：

- 34) 佢目珠 眨-啊⁵⁵-落, 就睡-忒哩。
他眼睛 一闔下去, 就睡-著了。
35) 寒天飯菜煮-啊⁵⁵-著, 無跔等食, 一下仔就冷泱泱仔³²。

³²選自《客語能力認證基本詞彙-中級、中高級暨語料選粹》：28-126。

冬天飯菜煮好沒有馬上吃，一會兒就冷冰冰的。

上兩句的意思是‘他眼睛剛剛闔上，接著就睡著了’、‘冬天飯菜剛剛煮好，若沒有馬上吃，一會兒就變成冷冰冰的’，在雙事件的複雜句中，*啊*⁵⁵將‘剛剛’的語意延伸、從語境中得到像國語的時間連結詞‘一...，就...’或英語的‘as soon as’的語意，表示前一個動作剛結束，另一個事件又發生了，因而將兩個先後順序緊密相連的子句整合成一個句子單元，因此出現在句尾的助詞*哩*，只能放在第二個事件之後，不能出現在第一個子句的句尾，如：

36) *佢目珠 眨-*啊*⁵⁵-落哩，就睡-忒。

此外，由於先前式主要的定義是與當下有關的過去行動，因此，動補結構中的動詞基本上都是動態動詞（dynamic verbs），在我們收集到的 172 個具有[動詞-*啊*⁵⁵-補語]結構的例句中，只有兩個例句的動詞是靜態動詞，當靜態動詞進入這個結構時，便被賦予了動態的語意（Bybee *et al.* 1994: 76），在客語的構式裡頭靜態動詞進入緊密先前式的結構後，一定要加上補語，如此則能讓靜態動詞得到使成式的意義（inchoative meaning），如：

37) *ngai* 講，*en* 佢自家來去東部，官識坐飛行機，爽*啊*⁵⁵去坐一擺，你乜講愛恁多錢，又無去成³³。

我說，我們自己去東部，沒做過飛機，我們一（變得）奢侈去坐一趟，你也說要那麼多錢，結果又沒去成。

38) 旦勢就在人食晝邊仔，一陣烏雲暗*啊*⁵⁵轉來，跔等一陣大雨倒*啊*過，日頭又走出來，像無落過雨樣仔³⁴。

大部分都是在一個人吃中飯左右，一陣烏雲飄過來讓天空變暗，接著一陣大雨下過後，太陽又出來，就像沒下過雨一樣。

如此以來，原本不可能出現的*[靜態動詞-補語]的結構，在動補之間加*啊*⁵⁵之後就成為剛剛變成的某個狀態，如：

39) 佢暢*啊*⁵⁵去，就去出國旅行。

他一（變得）高興，就去出國旅行。

40) *佢暢去，就去出國旅行。

因此，*啊*⁵⁵在緊密先前式裡頭佔有重要的地位。

³³ 《阿啾箭个故鄉》：146。

³⁴ 《阿啾箭个故鄉》：69。

另外值得注意的是，172 個[動詞-啊⁵⁵-補語]例句中，有 21 個例句含有‘動啊著’的句子，如：

- 41) a. 同學恁久了，莫動啊著就發關，火氣恁大³⁵。
同學那麼久了，別動不動就生氣，火氣那麼大。
- b. 佢性情古怪，動啊著就變面。
他性情古怪，動不動就生氣。
- c. 毋好動啊著就愛當天咒誓³⁶。
不要動不動就對天發誓。
- d. 毋好動啊著就飯匙堵貓，堵死人³⁷。
不可以動不動就以歪理堵塞人家的說法。

我們認為動啊著之所以演化成熟語，是因為在雙事件句子中，啊表達‘一...就...’之意，強調兩個連續事件之間的時間距離很近，正由於這種關係，在相關雙事件出現頻率很高的語用情境中，雙事件中的第一個事件很容易被詮釋成造成第二事件的條件，此時，很容易發展出習慣性的語意，這也就能解釋為何子句動啊著會熟語化為描述令人不悅的習慣性行為，類似英語出現進行式中的 *always* 的動不動、老是的負面語意（見 7）。

結語

本文介紹了客語與表達短時、降低程度的啊⁵⁵ 相關的四個時貌之間的語義關連。首先，相對於國語的暫時貌格式[V — V]（如想一想、搖一搖），以動詞的重複和最小的數字一來表示動作只做一些或一下而已，客語則以[V（V）啊⁵⁵ 哩²¹/啊³¹]的方式表示，其中非句尾的啊⁵⁵ 和出現在句尾受到聲調下降影響的啊³¹ 都是表示某動作進行的程度很淺、幅度很小的樣子。和客語關係很近的廣東話也有這種[V 吓⁵⁵]的結構，如：諗吓‘考慮一下：think for a while; consider’、斟吓‘說一下：have a brief talk’廣東話的吓（讀成 ha⁵⁵）³⁸和客語的啊⁵⁵ 一樣都是表示淺程度之意。

此外，客語的暫時貌的啊⁵⁵ 在嘗試貌裡頭一定要出現，因為，客語的嘗試構式不僅要表現嘗試之意，同時也要求要表現出淺程度之意，因此，嘗試構式[試-V-看-啊⁵⁵ 哩²¹/啊³¹]中，表嘗試意義的試或看之一可以省去，但啊⁵⁵ 哩²¹/啊³¹ 則一定要保留。

³⁵選自《台灣客家語常用詞辭典》。

³⁶選自《客語能力認證基本詞彙-中級、中高級暨語料選粹》：16-83。

³⁷選自《客語能力認證基本詞彙-中級、中高級暨語料選粹》：8-23。

³⁸選自 <http://www.cantonese.sheik.co.uk/dictionary/characters/1005/>

客語的重複貌[動詞；啊動詞；仔]中的仔，有學者標成完成貌的欸/哩/誼，但經我們以仔的隨前變調的現象檢測之後，確定重複貌的最後一個字應為小稱詞仔，因此，我們可以推論重複貌[動詞；啊動詞；仔]中的啊的語意也和仔一樣，是個降低程度之意，藉以前後呼應。

最後，客語的緊密先前構式[動詞-啊⁵⁵-補語]，啊⁵⁵一樣是必備的成分，此時原來表小幅度、淺程度的啊⁵⁵透過隱喻的轉換，由短時動作的語意轉化成事件剛結束、或是兩個連續事件之間的時間距離很近，因此得到‘剛剛’和‘一...就...’的語意。

總結來說，客語啊作為時貌詞時，有其基本的語意、延伸的語意、和隱喻的語意，啊並不是因為表現動作生動效果而產生的停頓（張雁雯 1998），強調的效果其實是來自先前時貌的特質。本研究有別於以往羅列各種有關客語啊字的功能的方式，透過小型語料庫的收集，儘量避免個人語感的偏差效果，將焦點放在暫時貌、嘗試貌、重複貌和緊密先前貌之間啊的語意的連慣性，藉以呈現客語獨特的時貌構式。然而，一個完整的時貌系統的研究，仍須包含更多時貌的結構，因此本研究尚未探討這些構式與其它時貌詞（如-過、-等）、或否定時間詞（如表還沒的‘官’）之間的關連性，也尚未全面的探討構式內部動詞本具的時間意義、動詞論元和補語搭配的條件等等；同時也需補強跨次方言或其它方言之間的比較，因為很可能這些時貌構式也經歷了像音韻變化一樣的歷史演變，這些都是將來仍須補充這些方面的探討。

中文參考書目

- 林立芳. 1997. 《梅縣方言語法論稿》。廣東：中華工商聯合出版社。
- 侯復生. 2002. 〈梅縣方言調詞後面的“阿”〉，《第四屆客家方言研討會論文集》。廣州：濟南大學出版社。
- 教育部國語推行委員會編輯，《台灣客家語常用詞辭典》。網址：
<http://hakka.dict.edu.tw/>
- 張雁雯. 1998. 《台灣四縣客家話構詞研究》，國立台灣大學中國文學研究所碩士論文。
- 賴維凱. 2006. 〈南北客家移民的源流、語言變化與虛詞啊〉。《第六屆客家研究研究生學術論文研討會》論文集，頁 245~259。
- 鍾萬梅等. 2009. 《客語能力認證基本詞彙-中級、中高級暨語料選粹》。台北：行政院客委會。
- 鍾榮富. 2005. 《湖廣地區漢語方言鼻音尾/鼻化小稱語法化》。研究國科會整合型計畫報告。
- 羅肇錦. 1988. 《客語語法》。台北：五南圖書出版股份有限公司。
- 饒長溶. 2001. 〈關於客家方言體貌助詞“啊”〉，《韶關學院學報》22.11。
- 龔萬灶. 2003. 《客語實用手冊》。苗栗：翰文印刷。
- . 2004. 《阿啾箭个故鄉》。苗栗：翰文印刷。

英文參考書目

- Anderson, Lloyd B. 1982. The 'Perfect' as a universal and as a language-particular category. In Hopper 1982:227-64.
- Bybee, Joan, Revere Perkins and William Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago and London: The University of Chicago Press.
- Crystal, David. 2006. *A Dictionary of Linguistics and Phonetics*. 5th ed. Malden: Blackwell Publishing.
- Levin, Beth. 1995. *English Verb Classes and Alternations – A Preliminary Investigation*. Chicago and London: the University of Chicago Press.
- Hsieh, Fang-yi (謝芳怡). 2007. 《漢語「V 一下」格式的探討》 (*Exploration of V yi-xia Constructions in Chinese: A Construction-based Account*.) Unpublished M.A. Thesis. Hsinchu, Taiwan: National Tsing Hua University.
- Li, Charles N. and Snadra A. Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. Berkeley and Los Angeles: University of California Press.
- McCawley, James D. 1971. Tense and time reference in English. In Fillmore and Langendoen 1971: 96-113.
- Vendler, Zeno. 1957. Verbs and Times, *The Philosophical Review*. Vol. 66, No. 2 (Apr., 1957), pp. 143-160, Published by: [Duke University Press](http://www.dukeupress.edu/) on behalf of

Philosophical Review.