基於決策樹演算法之台語連音變調預估模組

A Prediction Module for Taiwanese Tone Sandhi Based on the Decision Tree Algorithm

潘能煌　Neng-Huang Pan
建國科技大學資訊管理系
Department of Information Management
Chienkuo Technology University
nhpan@cc.ctu.edu.tw

余明興　Ming-Shing Yu
國立中興大學資訊工程學系
Department of Computer Science and Engineering
National Chung-Hsing University
msyu@dragon.nchu.edu.tw

蔡珮均　Pei-Chun Tsai
國立中興大學資訊工程學系
Department of Computer Science and Engineering
National Chung-Hsing University
tippy7346@hotmail.com

摘要

台語連音變調問題爲研究台語文轉音系統的重要問題之一。在詞的階層，大多數的詞都遵循詞尾本調，非詞尾變調的一般變調規則。在句子的階層，台語的連音變調問題會變得比較複雜，因爲一般變調規則並不能完全適用在句中的每個詞上。在這篇論文中，我們提出了一套可用於中文文句轉台語語音系統的台語連音變調預估模組。我們以決策樹 C5.0 演算法搭配三種 Special Case 來對句子中的各個音節做連音變調預估，此預估模組在內部測試和外部測試的預估正確率分別爲 93.42%和 91.13%。

Abstract

Taiwanese tone sandhi problem is one of the important research issues for Taiwanese Text-to-Speech systems. In word level, we can use the general tone sandhi rules to deal with the Taiwanese tone sandhi problem. The tone sandhi becomes more difficult in sentence level because of that the general tone sandhi rules for words may not apply at each word in a sentence. In this paper we proposed a module to deal with the Taiwanese tone sandhi problem for Chinese to Taiwanese Text-to-Speech systems. We adopt Decision tree C5.0 algorithm accompanied with three Special Cases generated from training data to predict the tone sandhi of each syllable. In this module, the accuracy of the inside test and outside test are 93.42%

and 91.13%, respectively.

關鍵詞：台語連音變調，文轉音系統，決策樹

Keywords: Taiwanese Tone Sandhi, Text-to-Speech System, Decision Tree.

一、緒論

在我國最常被使用的三大語言分別是國語、台語及客語，這三種語言都屬於聲調語言(Tonal Language)，也都存有連音變調(Tone Sandhi)的現象。所謂連音變調指的是在連續語音中，某些音節因受其前後音節的影響而不再保有原有調號的情形。例加：在中文裡，「老/ㄌㄠˇ/」和「虎/ㄏㄨˇ/」都是三聲字，而當這兩個字組成「老虎」這個詞的時候，「老」這個字的讀音會變成二聲的/ㄌㄠˊ/。「展/ㄓㄢˇ /」、「覽/ㄌㄢˇ/」、「館/ㄍㄨㄢˇ/」組成「展覽館」時「展」和「覽」的讀音都變成二聲。在台語中，「土」的讀音是/to2/，「地」的讀音是/de7/，而「土地」的讀音是/to1 de7/，我們可以發現「土」這個字的聲調由原本的二聲變成一聲。在客語中，「針」和「線」的讀音分別爲/ziim2/與/sien1/，當它們組成「針線」這個詞時，「針」的讀音要變成三聲的/ziim3/。

在台語中，所有單字詞的讀音聲調稱爲本調。大多數的文獻都認爲台語的聲調只有七種，分別爲： 東/dong1/、黨/dong2/、棟/dong3/、督/dok4/、同/dong5/、黨/dong6/、洞/dong7/、獨/dok8/，其中台語的二聲和六聲同調。可是我們發現若考慮連音變調和南腔北調的情形，台語的聲調變化高達十二種，詳見表一。舉例來說，台語三疊音中第一個字的聲調，不屬於常見聲調的任何一種，如：凍/dong0/凍凍以及入聲音的毒/dok0/毒毒。而聲調 9 和聲調 8 則爲南北腔調的不同。在詞的階層，只有少數的詞其所有的音節都讀本調，例如「頭痛/tau5 tiann3/」，而大多數的台語詞都遵循詞尾音節讀本調，非詞尾音節讀變調的台語一般變調規則。台語變調的情形可分成非入聲字與入聲字變調這兩個部分，詳細的變調規則可參考圖一。在入聲字的變調處理上，一般多認爲四聲和八聲互換，而我們認爲的入聲字變調處理應爲四聲轉爲二聲，而八聲和九聲變調後應爲三聲。台語詞內的變調範例可參考表二。台語拼音系統有很多，我們所採用的是台灣拼音系統[3]，因爲它能同時適用於國語、台語及客語等多種在台灣通行的語言，這有利於我們未來的發展。

在句子的階層，台語的連音變調情形就變得比較複雜了。有時候一個句子裏只有最後一個字讀本調，其餘的字都要變調。以句子「我買洗衣機」爲例，底下我們列出句中每個字的本調發音以及句子正確的台語讀音(經變調處理)。

例句：	我	買	洗	衣	機
本調：	ghua2	bhe2	se2	sann1	gi1
變調：	ghua1	bhe1	se1	sann7	gi1

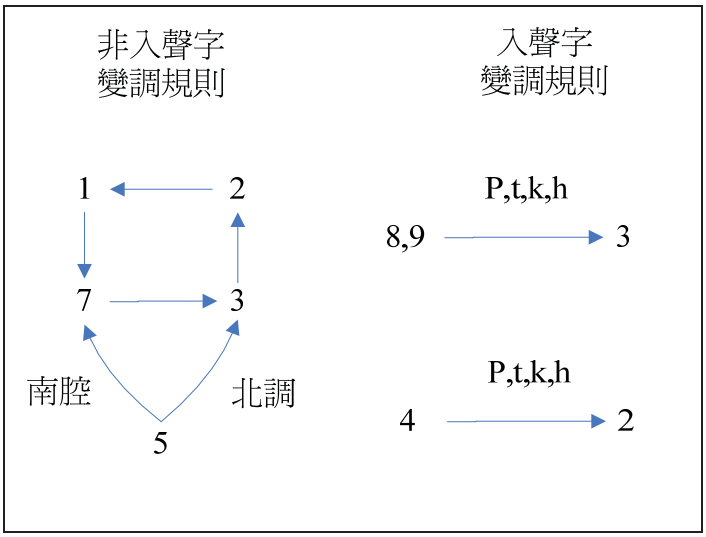
另一種情形則是我們可以將一個句子視爲由數個語法段落組成，每個語法段落稱爲變調詞組[4]。變調詞組的最後一個字讀本調，其餘的字皆變調。請看以下的例子，(底線表變調詞組)：

例句：	運	動	是	一	個	好	習	慣
本調：	un7	dong7	si7	zit8	ei5	ho2	sip8	guan3
變調：	un3	dong7	si3	zit3	ei5	ho1	sip3	guan3

句子「運動是一個好習慣」，經過中研院詞庫小組的線上斷詞器處理後可分解成六個詞，分別是「運動」、「是」、「一」、「個」、「好」、「習慣」。若依照圖一的規則來處理連音變調，那麼其結果就不會正確。因爲變調詞組是由一個或多個詞所組合而成，所以圖一的規則運用到句子上時就顯得不夠完善。我們由訓練語料發現非詞尾的音節有 95%的機率讀變調，而詞尾的音節有 60%的機率讀本調，40%的機率讀變調。這表示句中的非詞尾音節大致遵循一般的台語連音變調規則，而詞尾音節則沒有這種傾向，因此在句子階層裡需要有更佳的方法來處理台語的連音變調問題。

表一、台語聲調說明例表

聲調	例字 （台語發音）	聲調名稱與註解
dong0	凍（dong0）凍凍	高聲（三連音第一字）
dong1	東	高平
dong2 dong6	黨	高降
dong3	棟	低降
dong5	同	低緩上升
dong7	洞	中平
dok0	毒（dok0）毒毒	高入（入聲音）
dok2	剁（dok2）斷	中降（入聲音）
dok3	獨（dok3）立	低入（入聲音）
dok4	督	中入（入聲音）
dok8	獨	高入（入聲音南腔）
dok9	獨	高入（入聲音北調）



圖一、台語一般變調規則

表二、台語變調範例

本調(詞尾)	變調(非詞尾)
放心/sim1/	心/sim7/境
長久/gu2/	久/gu1/長
拖欠/kiam3/	欠/kiam2/人
委屈/kut4/	屈/kut2/服
同台/dai5/	台/dai3/北 (北調) 台/dai7/北 (南腔)
出外/ghua7/	外/ghua3/人
拘役/ik8/	役/ik3/男

目前台語連音變調問題的研究，大多數學者專家採用規則式方法來處理台語連音變調問題[1][4][8-10]。其中，Lin 與 Chen[1]利用四大規則(一般變調規則、「仔」前變調規則、輕聲變調規則、三疊形容詞變調規則)來實作連音變調模組，其測試語料的規模爲 5576 個音節，此系統的連音變調正確率爲 82.53%。楊允言等人[9]利用 20 條變調規則來處理台語的連音變調問題，在使用這些規則時會先將句中的每個音節都先設爲本調，然後再依後面的規則來修正其變調情形。原則上愈後面的規則的重要性愈強，所以後面的規則可以修正前面規則所產生的結果。此模組的測試語料規模比較小，只有 962 個音節，正確率爲 88.98%。梁敏雄[8]所發展出的文轉音系統也是採用規則式的方法來處理連音變調問題，其正確率爲 65%。洪俊詠[6]利用馬可夫語言模型來處理台語連音變調問題，其所採用的語料爲台語佛經，語料中有 6593 個句子內含 35543 個字，平均句長爲 5 個字。其預估正確率爲 84%。許書豪[7]先利用貝氏網路取得初步的連音變調預估結果，再用中文詞對應台語變調規則來修正這些結果的方式來處理台語連音變調問題。其訓練語料共有 583 句，內含 8138 個音節，外部測試正確率爲 85.84%。

在本論文中，我們提出了以決策樹 C5.0 演算法搭配三種 Special Case 來對句子中的各個音節做連音變調預估的方法。我們的模組可以應用在中文句轉台語語音系統中，用來提升文轉音系統在文句分析上的能力。本論文的章節安排如下:在第二節，我們將介紹我們的實驗方法。第三節主要介紹各種方法的實驗結果。結論在第四節。

二、實驗方法

(一)語料

我們從 Chinese Gigaword Third Edition 的繁體中文語料中隨機抽取 3672 個句子作爲實驗語料，並將語料分爲訓練語料(約佔 75.5%)和測試語料(約佔 24.5%)。並透過

先前開發的中文文句轉台語語音系統[12]，取得中文文句所對應的台語拼音，最後再使用台語文句拼音校正工具[7]來做人工校正。表三爲我們實驗語料的相關數據。Chinese Gigaword Third Edition 中的文章包含了斷詞結果與詞性標記，爲了避免資料稀疏問題，我們透過中研院詞類標記表將所有的詞性標記改爲精簡詞類。

表三、訓練語料與測試語料數據

	句子數	音節數
訓練語料	2772	38508
測試語料	900	12452
總數	3672	50960

(二)決策樹C5.0 演算法

台語連音變調問題可被視爲一種分類問題，因爲每個音節的讀音只有本調和變調這兩種可能。決策樹(Decision Tree)是使用樹狀分岔來產生分類規則，藉由一連串的問題和規則將資料做分類，藉由相似的形態來推測相同的結果。下列說明決策樹規則產生的過程[2]：

步驟1. 首先，將所有輸入資料母體作爲根節點。

步驟2. 決策樹逐一掃描所有的輸入變數，以計算每個輸入變數對應預測變數的分岔準則，然後根據分岔準則挑出最佳分岔變數，用此分岔變數產生資料分割，以產生子節點。各個子節點根據案例的預測變數分布機率，指派分類結果以及產生分類機率。

步驟3. 將子節點視爲新母體，透過同樣步驟持續讓決策樹生長，最後採用修剪技術(Pruning)修剪不必要的規則。

一般常用的決策樹演算法大致爲 C5.0、CART、CHAID 與 QUEST 等四種方法。我們的研究使用決策樹 C5.0 演算法來處理台語的連音變調問題。C5.0 演算法屬於監督式學習演算法，也可稱爲規則推理模型。它能夠對連續型變數及類別型變數做分析。C5.0 的每一個節點可以產生不同數量的分支，它會依最大資訊增益(Information Gain Value)的欄位切割樣本，重複切割直到樣本子集不能再被分割爲止，且能依使用者需求產生決策樹及規則集兩種模型[11]。

我們利用詞性、詞長、前一詞詞性、前一詞詞長、前二詞詞性、前二詞詞長、後一詞性、後一詞詞長、後二詞性、後二詞詞長、是否爲詞尾以及是否爲句尾等十二項參數透過決策樹 C5.0 演算法產生台語連音變調分類規則。以下將利用一個例句來說明我們所使用的各項參數。

例句：可(ADV)作爲(Vt)人工(N)肝臟(N)的(T)生物(N)感應器(N)。

1. 詞性：音節所在詞的詞性。例句中的「可」所在詞爲一單字詞，其詞性爲 ADV。「作」這個音節所在詞「作爲」是一個二字詞，其詞性爲 Vt。

2. 詞長：音節所在詞的字數。「可」這個音節所在詞爲一單字詞，詞長爲 1。「作」這個音節所在詞爲一個二字詞，詞長爲 2。

3. 前一詞詞性：音節所在詞的前一詞詞性。「可」這個單字詞爲句子中的第一個詞，沒有前一詞，故標記爲 Null。「作」這個音節所在詞爲「作爲」，其前一詞「可」的詞性爲 ADV。

4. 前一詞詞長：音節所在詞的前一詞所含字數。單字詞「可」爲句子中的第一個詞，沒有前一詞，故前一詞詞長記爲 0。「作」這個音節所在詞的前一詞爲「可」，其詞長爲 1。

5. 前二詞詞性：音節所在詞的前面第二個詞的詞性。「可」爲句子的第一個詞，故此項參數標記爲 Null。「作」這音節所在詞，「作爲」，爲句子的第二個詞，無前二詞，故前二詞詞性標記爲 Null。「人」這個音節所在詞，「人工」，爲句子中的第三個詞，其前面第二個詞爲單字詞「可」，因此其前二詞詞性爲 ADV。

6. 前二詞詞長：音節所在詞的前面第二個詞所含字數。「可」這音節所在詞爲句子的第一個詞，無前二詞，故前二詞詞長記爲 0。「作」這音節所在詞爲句子的第二個詞，無前二詞，故前二詞詞長記爲 0。「人」這個音節所在詞爲句子中的第三個詞，前面第二個詞爲單字詞「可」，其前二詞詞長爲 1。

7. 後一詞性：音節所在詞的後一詞的詞性。單字詞「可」的後一詞爲「作爲」，其詞性爲 Vt。「工」這音節所在的詞爲「人工」，其後一詞爲「肝臟」詞性爲 N。「感」這音節所屬的詞爲「感應器」，爲句子中最後一詞，無後一詞因此標記爲 Null。

8. 後一詞詞長：音節所在詞的後一詞所含字數。「可」的後一詞爲「作爲」其詞長爲 2。「工」這音節所在的詞爲「人工」，其後一詞爲「肝臟」字數爲 2。「感」這音節所在的詞「感應器」爲句子中最後一詞，無後一詞故此項參數記爲 0。

9. 後二詞詞性：音節所在詞的後面第二個詞的詞性。單字詞「可」的後面第二個詞爲「人工」其詞性爲 N。「工」這音節屬於二字詞「人工」，其後面第二個詞爲「的」詞性爲 T。「感」這音節所在的詞爲「感應器」，是句子中最後一詞並無後二詞，故標記爲 Null。

10. 後二詞詞長：音節所在詞的後面第二個詞所含字數。「可」的後面第二個詞爲「人工」其詞長爲 2。「工」這音節所在的詞爲「人工」，其後第二個詞爲單字詞「的」，故「工」的後二詞詞長爲 1。「感」這音節所屬詞爲「感應器」是句子中最後一詞並無後二詞，故其後二詞詞長應記爲 0。

11. 是否爲詞尾：音節是否爲其所在詞的最後一字。「可」這音節所在詞爲一單字詞，故「可」爲詞尾。二字詞「作爲」的第一個音節爲「作」，故「作」屬非詞尾。而「爲」是二字詞「作爲」的第二個音節，故「爲」爲詞尾。

12. 是否爲句尾：音節所在位置是否爲句子的最後一音節。「可」這音節所在位置不爲句子的最後一音節，故標記爲非句尾。「感」這音節所在詞爲句子的最後一個詞，但「感」這音節爲「感應器」詞中的一個音節故仍爲非句尾。「器」這音節爲句子最後一個詞「感應器」中的最後一個音節，故標記爲句尾。

(二)三種Special Case

我們從訓練語料中發現某些詞的變調情形在語料中是固定的，所以我們進一步分析詞的本身是否會影響連音變調的結果，以下是我們探討的三種 Special Case。

1. Word：在訓練語料中，若某個中文詞出現次數大於等於 2 次，且詞內各音節的本變調樣式在訓練語料中各音節本變調順序中所佔比例大於等於 94%，我們就將這樣的中文詞收錄當成規則，往後遇到完全一樣的中文詞時，可直接給定各音節本變調結果。例如，在我們的訓練語料中，「總統府」這個詞出現過 4 次且每次的讀音都是/zong1 tong1 hu2/，所以在「總統府」這個詞內的本變調樣式爲(變調，變調，本調)的機率爲 100%。往後只要遇見「總統府」這個詞，系統將會直接指定其連音變調結果爲(變調，變調，本調)。

2. POS+Word$_R$：在訓練語料中，若某一詞 W 的詞性爲 POS 且其下一個詞爲某一特定中文詞時，這樣的“詞性”加上“中文詞”的組合，在訓練語料中出現次數大於等於 2 次且 W 的詞尾音節本變調在此種情形下出現比例大於等於 94% 時，我們就會將這種組合收錄成 Special Case。我們從語料中發現，若詞 W 的詞性爲 DET 且其下一個詞爲「國」這個單字詞時，W 的詞尾音節有明顯的變調傾向。在下面的兩個例句中，詞性爲 DET 的兩個詞的詞尾都讀變調。

   例句 1：分頭 拜會 各(DET)/gerh2/ 國 駐 聯合國 代表團。
   例句 2：這 是 兩(DET)/nng3/ 國 政治 最 主要 的 分野。

3. Word$_L$+POS： 訓練語料中，若某一詞 W 的詞性爲 POS 且其上一個詞爲某一特定中文詞時，這樣的“中文詞”加上“詞性”的組合，在訓練語料中出現次數大於等於 2 次且 W 的詞尾音節本變調在此種情形下出現比例大於等於 94%，我們就會將這種組合收錄成 Special Case。我們從語料中發現，若詞 W 的詞性爲爲 ADV 且其前一個詞爲二字詞「如果」時，W 的詞尾音節有明顯的變調傾向。在下面的兩個例句中，詞性爲 ADV 的兩個詞的詞尾都讀變調。

   例句 1：政府 如果 不(ADV)/m3/ 加強 保護 淡水魚 。
   例句 2：大陸 各 地 民族 如果 都(ADV)/long1/ 使用 這 個 招數。

從訓練語料中抽出三種 Special Case 的相關規則之後，我們會把這些規則與前一小節所發展的 C5.0 連音變調預估模組結合在一起。以下我們將說明兩者結合的處理過程，假設輸入的文句爲「台灣(N) 光復(Vt) 后(N) 土地(N) 改革(Vt) 研討會(N)。」

步驟1. 首先要檢查句子的內容有無符合三種 Special Case 的規則。

a、 Word 規則：

「台灣」在我們的 Word 規則中出現，有 98.62%的機率讀(變調，本調)。

「土地」在我們的 Word 規則中出現，有 94.74%的機率讀(變調，本調)。

b、 POS+Word$_R$ 規則：

在本例中找無符合條件的規則。

c、 Word$_L$+POS 規則：

找到(土地，Vt)的規則，即當詞性爲 Vt 且前一詞爲「土地」時，有 94.74%的機率使詞性 Vt 的詞尾讀本調。故在本例中，音節「革」應讀本調。

步驟2. 利用決策樹 C5.0 演算法預估剩餘音節的連音變調結果。

本說明例的最終結果如下，在此“本”表本調，“變”表變調。

音　節	台	灣	光	復	后	土	地	改	革	研	討	會
預估結果	變	本	變	變	本	變	本	變	本	變	變	本

三、實驗結果

我們的語料共有 3672 個句子，內含 50960 個節，其中 2772 句，38508 個音節(約 75.5%)爲訓練語料，其餘 900 句，12452 個音節(約 24.5%)爲測試語料。表四爲我們的實驗結果，決策樹 C5.0 演算法的內部測試與外部測試的正確率分別爲 89.39%以及 88.84%，而決策樹 C5.0 演算法搭配三種 Special Case 方法的內部測試與外部測試的正確率分別爲 93.42%及 91.13%。我們可以發現決策樹 C5.0 演算法在加入三種 Special Case 的相關規則後其預估正確率有顯著的提升。表五列出台語連音變調問題相關研究的實驗數據，雖然各研究所採用的語料都不同，所以無法做出完全客觀的比較，但仍有一定的參考價值。我們可以從中發現本論文所提出的方法具有較高的預估正確率，且測試語料的規模也比較大。

表四、實驗結果(預估正確率)

	內部測試	外部測試
決策樹 C5.0 演算法	89.39%	88.84%
決策樹 C5.0 演算法+Special Case	**93.42%**	**91.13%**

表五、台語連音變調問題相關研究實驗數據

相關研究	正確率	測試語料大小
Lin and Chen [1]	82.53%	5576 音節
洪俊詠[6]	84%	未知
楊允言等[9]	88.98%	962 音節
許書豪[7]	85.84%	159 句
本研究	**91.13%**	**12452 音節**

四、結論

國語、台語和客語爲我國三大常用語言，這三種語言都是聲調語言，也都存有連音變調的現象。和另外兩種語言相比，台語的連音變調問題比較複雜，因此也成爲發展高品質的台語文轉音系統的一大難題。本論文提出一個以決策樹 C5.0 演算法搭配三種 Special Case 的方法來處理台語的連音變調問題，我們的模組之內部測試和外部測試的預估正確率分別爲 93.42%及 91.13%。我們的方法與其它台語連音變調問題相關研究相比，我們的模組具有較高的預估正確率。

參考文獻

[1] C. J. Lin and H. H. Chen,"A Mandarin to Taiwanese Min Nan Machine Translation System with Speech Synthesis of Taiwanese Min Nan," **Internal Journal of Computational Linguistic and Chinese Language Processing**, Vol. 4, No. 1, pp. 59-84, 1999.

[2] 尹相志，*SQL Server 2008 Data Mining 資料採礦*，悅知文化，2009。

[3] 中興大學資工系語音語言實驗室網站，2012，http://speechlab.cs.nchu.edu.tw/

[4] 李尙德，*台語辭典建構與台語變調探討*，碩士論文，資訊科學研究所，中興大學，2007。

[5] 邱玉雪，*台灣閩南語偏正結構詞組中的變調分界*，碩士論文，台灣語言與語文教育研究所，新竹師範學院，2004。

[6] 洪俊詠，*馬可夫語言模型應用 di 台語變調 gah 注音*，碩士論文，統計學研究所，清華大學，2005。

[7] 許書豪，*台語連音變調問題研究*，碩士論文，資訊網路與多媒體研究所，中興大學，2010。

[8] 梁敏雄，*台灣多語語音資料庫之建立及應用*，博士論文，電機工程學研究所，長庚大學，2008。

[9] 楊允言、李盛安、劉杰岳、高成炎，“台語變調系統實作研究，” *第十七屆自然*

語言與語音處理研討會論文集，台南，台灣， pp.293-304，2005。

[10] 鄭良偉，*台語的語音與詞法*，遠流出版公司，1997。

[11] 廖述賢、溫志皓，*資料探勘理論與應用-以 IBM SPSS Modeler 爲範例*，博碩文化，2012。

[12] 潘能煌、余明興、蔡宗謀，2008，“中文文句轉台語語音系統，” *第十三屆人工智慧與應用討論會論文集*，宜蘭，中華民國，pp. 1-5，2008。