

# 破音字發音的預測方法

王文俊<sup>†</sup> 黃紹華<sup>‡</sup> 李俊曉<sup>†</sup> 劉繼謚<sup>†</sup>

<sup>†</sup>交通部電信研究所 基本科技研究室

<sup>‡</sup>國立交通大學 電信工程研究所

email: wjwang%tl9000@tlrouter.motctl.gov.tw

## 摘要

目前在中文語音合成研究的主要方向大都偏重在如何提高清晰度及自然度，而字轉音的正確率問題則較少討論到。事實上語音合成的最主要目的是要讓聽者了解合成語音的意義，而不致產生誤解。因此字轉音正確率的提高實有其重要性，本文就是希望能利用有效的分析方法結合語言學的知識，針對破音字的發音提出解決的方法。本實驗所採用的方法是結合字的聯想與樹狀語言模式兩項處理。由實驗結果發現總合預測正確率約為 85%，較傳統猜測法提高了約 20% 以上。其中字的聯想是一種類似句型的觀念，而樹狀分析模式則是為了能有效地運用詞類訊息。這些方法均具有延展性及擴充性，可適用於其它破音字的處理或相關的語音及語言處理。

## 1. 緒論

目前在中文語音合成研究的主要方向大都偏重在如何提高清晰度及自然度，而字轉音的正確率問題則較少討論到。事實上語音合成的最主要目的是讓聽者了解合成語音的意義，而不致產生誤解。因為就語音認知程序而言，聽者是藉由正確的音來組成相對應的字及詞，進而了解文句或文章的意義。因此為了達成此一目標，就必須降低字轉音的錯誤率。而造成國語字轉音錯誤的最大原因就是破音字的問題，因為錯誤的音形成不同意義的字及詞，導致誤解文句的原意。所以本文就是想利用有效的分析方法結合語言學的知識，針對破音字的發音提出解決的方法。

對於中文而言詞才是句法上及語意上的最小元素，因此在文字翻語音的處理中，輸入文字串必須先經過斷詞及構詞處理。而傳統上破音字的問題就是希望能在斷詞時一併解決，藉由事先已經經過檢查且存在詞庫中的各詞的相對注音決定各詞的發音，而破音字的發音也隨著詞發音的決定而跟著決定。這樣的處理除非在斷詞混淆或音異義異的破音詞情況下，的確可以解決絕大部份的破音字問題。而本篇文章所要探討的主題則正是經過上述斷詞處理後成為單字詞的破音字的發音問題，至於斷詞混淆或異音異義的破音詞問題則不在此討論。

## 2. 中文文法的特性

破音字並非中文固有的特性，相同的情況也出現在英文，稱之為 Homography 其義為同形異義字。對文句翻語音的系統而言，在處理過程中可資利用的資訊應就是上下文的文字及詞類。以「為」字為例從中研院中文詞庫小組所發表的資料〔1〕中可看出「為」字的發音可以根據詞類不同而有效的區別，如書中所述當「為」字在文句中的詞類為 P02 或 VG2 時此字應為第二聲，而當其詞類為 VJ1 或 P03 時此字應為第四聲。但事實上如此書所列之總數 178 類的詞類對電腦處理甚至非語言學家而言都並非很容易的事；而至於在縮小詞類數目下所發展出的詞類標示系統，其正確率亦尚無法達到百分之百同時也並未就此一破音字混淆問題做深入探討。因此本文的目的就是希望在容忍一些詞類標示混淆的情況下，仍能藉由其他相關的資訊幫助作破音字的分析，以建立一套有效的處理模式。本實驗初期的目標仍先選定由「為」字開始，因為在中研院所公佈的辭典中「為」字的詞類是佔整個八萬詞目詞典中的第九位，而且其他大部分的破音字成詞的機會很大，形成單字詞造成混淆的機率反而較小。

決定破音字發音的困難究其原因主要來自斷詞及構詞的錯誤，使得應成詞的破音字在錯誤的詞位，而造成發音的混淆。另一點則是詞類標示有問題所造成，而詞類標示的問題則與下列兩點有關〔2〕：一為大部分的中文詞均可作為動詞、名詞或形容詞但卻沒有任何的構詞變化。二為中文無嚴謹的文法限制，在文句中允許各種詞類的組合，甚至不合文法的情況亦常存在，因而增加了所有語言處理的困難。以下所提出的方法即是希望能補償屬於中文文法的這兩項缺失。

### 3. 樹狀語言模式

文字翻語音的處理是由一整串的文字中去預測各字的發音，因此可資利用的資訊應包括文字及隱藏的詞類。本章就是要討論如何有效的運用這兩項資訊，首先處理的對象當然是上下文中的文字，對於此項字的處理，本文所提出的方法可稱為字的聯想 (lexical association)，所謂 lexical association 是一種根據相鄰字的出現頻率幫助決定破音字的發音，事實上和破音字成詞的字就可以視為是一種距離為 1 的 lexical association 的情況，而當把距離延長時也可以找出其他一些高頻率的字。這些字和破音字的組合可視為是一種常用句型，其中在這些常以組合方式出現的字中間可以擺上不同的字或詞，此種架構在文法上雖不屬於詞，但就實用而言在文章及口語中常會搭配出現，此種類似句型的觀念可用來解決詞庫過度膨脹的問題，而適當地運用句型及複合詞的文法更可有助於增加系統處理文句的深度和廣度。不過由於中文的結構允許許多詞可以合併為新詞，因此在這種不一致的構詞處理下，此種句型架構的資料結構表示方式並不容易。

接下來要談到如何利用詞類訊息，雖然我們承認詞類標示的正確對所有語言處理都有很大的助益，但在此將不作如何提升詞類標示正確率的討論。我們秉持的是一種 partial parsing 的觀念，將不易作詞類標示易造成混淆的部份，由更高層次的處理來解決，在此我們所採取的分析方法是樹狀語言模式。此種分析模式已被廣泛利用在多方面的研究 [3] [4]，且得到不錯的結果。樹狀語言模式可視為是 N-gram model 的變形，當 N=3 時此模式就變成所謂的三連文法模式。眾所週知 N-gram model 明顯地較三連文法模式精確，因為此種模式考慮到更多的資訊，但是在大部分的情況下，三連文法模式已經夠用，因而傳統的語言模式通常只用到雙連文法模式或三連文法模式及其之間的平滑模式。在 [5] 中就曾提及利用類似雙連及三連文法模式的條列式相似率比較法，來解決破音字問題。所謂條列式相似率的作法係以所欲處理的破音字在文句中不同位置的相鄰詞及其所具詞類為依據，並利用此破音字的正確注音作區分，計算出個別的相似值；在比較之後，留下一些具有高鑑別率的規則作為其後判斷的依據。這種作法的缺點是對詞的位置的處理缺乏彈性，同時只能建立簡單的分類規則。為了使破音字發音預測的正確率提高，我們必須考慮更多的相鄰字或詞，並且要有更有效的分析方法，因此 N 值就有必要提高並建立更有效率的樹狀語言模式。此模式的分析步驟為：首先必須整理出所有影響的因素形成 Questions set，理論上這些影響「為」字發音的因素應由語言學家設計，但在本實驗進行時並無法找到相關的資料，而且即使有了類似資料其是否適於程式化執行亦未可知，因此本實驗所使用的 Questions 是在考慮系統複雜度及前節所述中文的特性和「為」字的語法關係之下所制定出來的，這些規則再經由檢測語料庫的步驟篩選出適當的規則組合以作為樹狀語言模式的產生依據。接著再以二分分裂法的方式持續地在所有 Questions set 中選擇最有效的 Question，將資料分割成數個內部分佈更一致的小單元。本篇文章所討論的樹狀模式建立程序主要是利用 Greedy algorithm，分裂的判斷依據則是採用 Gini criterion [6]。至於更詳細的樹狀結構的運算及修剪將不在此討論。樹狀語言模式的優點是利用多變化的 Question 組合，可以用更有效的方式來實現 N-gram model，另外在 N-gram model 下由於考慮語料中無法包含各種

可能的組合因此必須使用到各種平滑的技巧以避免機率值出現為零的這種考慮，在樹狀語言模式中可以被輕易地克服，原因是此系統具有容錯的能力在無符合分析時所有的情況下，依然能產生一個最接近的結果。

## 4. 實驗結果與討論

首先我們先介紹實驗所使用的詞庫及語料，詞庫的總數約為八萬詞目，其中包含單字詞、雙字詞、三字詞、四字詞及五字詞。訓練及測試的語料庫則是取自報紙新聞，整個語料庫總計約為 35 萬句，詞數約為 260 萬，字數約為 400 萬。所有的文句先經過初步的斷詞處理 [7] 後，整理出含有成詞的「為」字共 11400 句，而含有單字詞的「為」字共 12200 句，我們所要處理的就是這 12200 句的語料。整個處理的步驟如下：第一步是對所有語料中的破音字進行標示注音的工作，接著統計相鄰文字的出現頻率，得出表一的結果，由表一可看出滿足表中所列句型約 3000 句而「為」字發音正確率可達 98%，其中屬於相鄰字的詞可以考慮加在系統的詞庫中，至於某些具有高頻出現率的詞，如「為了」為何不在原先之詞庫中，理由是我們所用的原始斷詞方法，希望能將某些具有規則的字留在構詞處理時解決。第三步則是建立屬於「為」字的樹狀語言模式，如圖一所示為分析後所得結果之示意圖，本系統所使用的詞類分類係參考 [8] 其中分類 0 至分類 11 為述詞，分類 39 為述補式複合詞，分類 20 為代名詞。在圖中所出現的幾個 Question 是經過上節所述的處理步驟而產生的，在此略述一下其所代表的意義：由字面意義分析可知，「為」字的單字意義可區分為「是」、「被」或「替」前兩者發第二聲，而後者則發第四聲；一般而言當「為」字有「是」的意義時，其前通常會緊接著動詞，而當「為」字有「替」的意義時，其後通常會有代名詞或動詞，而很多情況下原本屬於動詞的詞，在不同的文句中會轉化為其他詞類，因此有必要考慮這些特殊的情況。

表二所示可視為是傳統破音字發音猜測法和樹狀語言模式所得結果之比較，如表中語料庫資料分佈的數據，就正是傳統作法以在大量語料庫中出現機率最大者為主的預測正確率數據，而利用樹狀語言模式所得到的預測正確率則約為 80%，若再加上由 lexical association 所得到的結果正確率更可提高為 85% 左右。

## 5. 結論

字轉音的正確率對語音合成系統而言可以視為是一個基本的要求，雖然對國語共約 1300 個基本單音及 5401 個常用字而言，破音字所佔的數量也許並不大，但錯誤的破音字發音對系統仍是實現語意了解的一大障礙。本文所提出的解決破音字發音的方法是結合字的聯想與樹狀語言模式兩項處理，由實驗結果發現總合預測正確率約為 85% 較傳統猜測法提高了約 20% 以上。其中字的聯想是一種類似句型的觀念，而樹狀分析模式是為了能有效地運用詞類訊息；這

些方法均具有延展性及擴充性，可適用於其它破音字的處理或相關的語音及語言處理。

未來的研究方向將朝向更詳細的樹狀結構分析以提高預測的正確率，另外也將擴大研究至所有的破音字，並考慮增加詞意的資訊來分辨如「一行人」或「一行字」這種相鄰詞類相同的情況以及包括音異義異的破音詞處理。

## 致謝

感謝交通部電信研究所所長王金土博士，副所長周義昌博士以及基本科技研究室主任鄭伯順博士對語音信號處理研究的持續支持與鼓勵。也感謝交通大學電信工程研究所教授陳信宏博士的指導與建議。

## 參考文獻

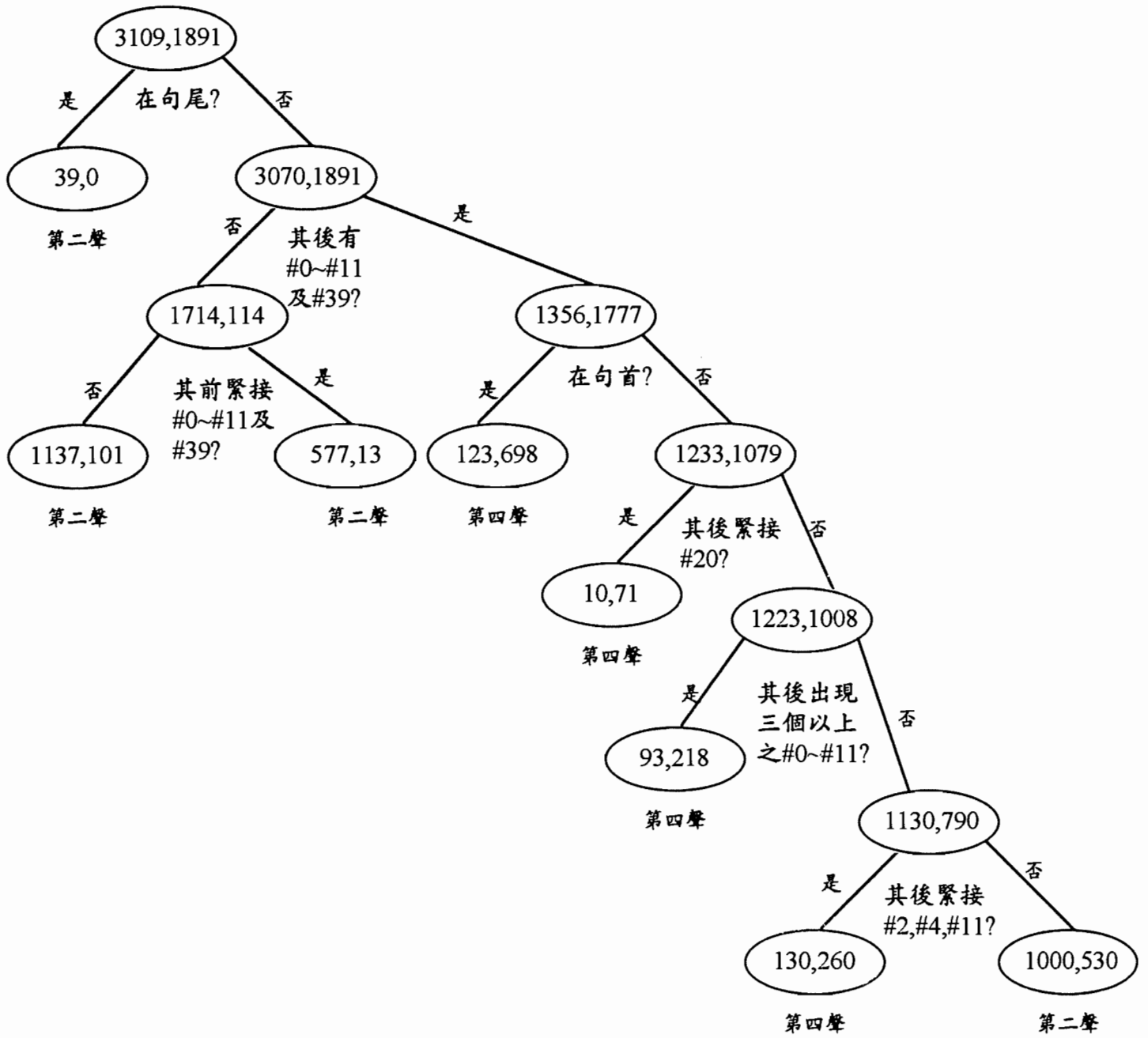
- [1] 中央研究院，資訊科學研究所，「中文書面語頻率辭典」。
- [2] K.J.Chen, S.H.Liu, L.P. Chang and Y.H. Chin, "A Practical Tagger for Chinese Corpora", ROCLING 1994, pp.111-126.
- [3] Michelle Wang and Julia Hirschberg, "Automatic Classification of Intonational Phrase Boundaries", Computer Speech and Language 1992 vol.6, pp.175-196.
- [4] W.J.wang, N.Campbell, N.Iwahashi and Y.Sagisaka, "Tree-based Unit Selection for English Speech Synthesis", ICASSP 1993, pp. 191-194.
- [5] 王文俊，李俊曉，「以詞類分析作破音字處理」，交通部電信研究所，內部報告。
- [6] L.Breiman, J.H.Friedman, R.A.Olsen and C.J.Stone, "Classification And Regression Trees", Monterey. CA: Wadsworth, 1984.
- [7] 李俊曉，李俊仁，黃英峰，「中文文句翻語音語言處理系統—工作文件 1」，交通部電信研究所，內部報告。
- [8] 蘇育新，「中文文句自動斷詞標詞類之研究與應用」，國立交通大學碩士論文。

表一、含「爲」字的句型整理

句 型	樣本總數	注音爲第二聲	注音爲第四聲
...爲了...	1133	0	1133
...爲之...	91	89	2
...爲由...	124	124	0
...以...爲...	769	756	13
...爲...去...	10	0	10
...爲...而	228	13	215
...爲的(卻,就)是.	7	0	7
...稱爲...	41	40	1
...爲..了...	97	11	86
...爲(此,免,使)...	316	6	310

表二、使用分類樹對於含「爲」字語料的影響

語料項目	資料分佈	預測正確率
訓練語料	62.18%	80%
測試語料	62.60%	79.175%



圖一、分類樹示意圖