# 國語廣播新聞語料轉述系統之效能評估

# Evaluation of Mandarin Broadcast News Transcription System

張隆勳、王逸如、陳信宏
國立交通大學電信工程系

摘要

在本論文中，使用國內自行錄製的國語廣播新聞語料庫，MATBN，製作一個基本的語音辨認系統以評估在國語廣播新聞環境下之國語語音辨認效能。在論文中所使用語音辨認器之聲學模型爲 100 韻母相關之聲母及 40 個韻母模型，另外也爲particles及超語言現象製作了聲學模型。在語言模式方面，論文中使用六萬詞之國語詞典及其雙連文模型；同時在論文中還加入了最簡單的韻律資訊－音節間靜音長度模型以期提升辨認器效能及詞、語句邊界的正確率。最後，對國語廣播新聞語料中的三種不同語者環境－主播、外場記者及受訪者，分別得到 86.9%、76.4%及 48.5%的詞辨認率。

## 一、 簡介

在 1995 年世界四個做語音辨認研究的著名單位(BBN, CMU, Dragon 及 IBM)開始參與一個在當年是一項創舉的語音辨認評比之語音資料庫建立工作，該語音資料庫稱做 Hub-4，在此項評比中希望能做到廣播新聞語料自動轉述(automatic broadcast news transcription)[1]。Hub-4 語料庫中也已陸續加入許多語料，事實上 Hub-4 語料庫中也已經有國語廣播新聞語料，其內容是由大陸中央台及洛杉磯中文台的廣播新聞節目錄製而成。由 1999 年 DARPA 所舉辦的語音辨認評比的結果可以看出世界各大語音辨認研究單位在廣播新聞語料自動轉述已獲得重大的進展；不只在語音辨認方面，在 segmentation、information extraction、topic detection 等技術都有許多成果。在英文廣播新聞語料語音辨認方面，在 DARPA Broadcast News (Hub-4) Evaluation [2]的 F0 評比項目 － 其訓練及測試環境是僅考慮無環境雜訊、背景音樂及無外國口音語者的廣播新聞語料，其語音辨識率已可達 7.8% 的詞錯誤率(word error rate, WER)；而在 F1 評比項目 － 其訓練及測試環境是 F0 再加上自發性廣播新聞語料(spontaneous speech)，也就是考慮了有不流利現象 (disfluencies) 的語料，其辨認結果也可達 14.4% 的詞錯誤率[2]。在國語廣播語料語音辨認部分，Dragon 公司在 1998 年發表的辨認結果可達 36%的詞錯誤率及 25%的字錯誤率(character error rate, CER)[3]。

在國內則從 2001 年起由台大、中研院、清大、成大及交大五個學術單位，在國科會的補助

感謝中研院王新民博士在MATBN語料庫標示內容上之協助及台師大陳柏琳教授所提供之詞典。

下開始了一項為期三年的國語廣播語料蒐集計畫。其中之一部分為蒐集國語新聞廣播語料庫(MATBN, Mandarin Across Taiwan – Broadcast News)[4,5]，三年計畫中共蒐集並轉述了 198 個小時的國語廣播新聞語料。這個國語新聞廣播語料，MATBN，現在正要由國科會技轉到語言學會中。

## 二、 國語新聞廣播語料庫(MATBN)

MATBN 計畫中所錄製的是「公視新聞深度報導」和「公視晚間新聞」兩個國語新聞廣播節目之內容，每次節目進行長度一個小時，錄製與處理標記共分三年進行，從 2001 年 11 月到 2004 年 7 月，錄製資料量如表一所示。

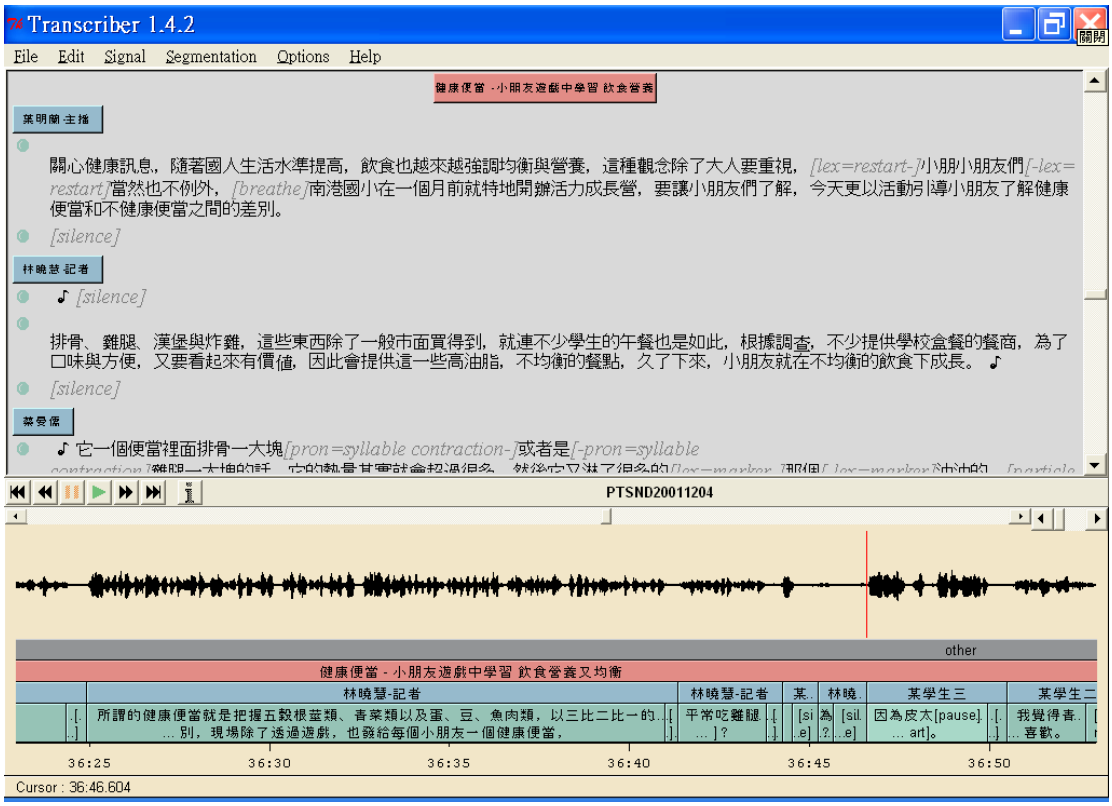
表一、MATBN 廣播語料庫之統計。

| 錄製時間 | 錄製資料量 |
|---|---|
| 第一年（2001 ~ 2002） | 40 小時 |
| 第二年（2002 ~ 2003） | 80 小時 |
| 第三年（2003 ~ 2004） | 78 小時 |
| 總計 | 198 小時 |

MATBN 語料的錄製過程是直接在電視台主控室利用 DAT（Digital Audio Tape）以 44.1 KHz 的取樣率和 16 bits 的精確度錄製，然後再做 down sampling，將取樣率降到 16 KHz。

MATBN 與 Hub-4 相同也是使用 LDC(Linguistic Data Consortium)所發展的廣播語料標示工具 Transcriber [6]來轉述的。廣播語料轉述軟體 Transcriber 是一套可以顯示聲音波形，並同時提供標記背景聲（Background Sound）、內容主題（Topic）、語者（Speaker）及語音內容的一套軟體，另外，還可以記錄除了一般文字之外的常見口語語音現象，例如：呼吸聲、particles 以及笑聲、嘆氣聲、砸嘴聲等超語言學現象（Paralinguistic Phenomena）。

這套廣播語料轉述軟體的編輯環境介面如下圖一所示，其中一段聲音的資訊標記均使用四層狀態來記錄，由上而下分別為背景聲、內容主題、語者以及語音內容，也正因為這四層標示資訊所以可以完整標記出廣播新聞的各所語音及其他聲音資訊。




圖一、標記軟體 Transcriber 之編輯介面。

在本論文中是使用與 Hub-4 新聞語料訓練及測試環境 F1 相同設定，首先將 MATBN 第一、二年語料依轉述資料依語者標示資料切割爲一個個語者項(speaker turn)，再將有環境雜訊或背景音樂之 speaker turn 去除。在論文中我們將 MATBN 廣播新聞語料，依據語者環境區分爲內場主播（Anchor）、外場記者（Reporter）和受訪者（Interviewee）三類，因爲不同的語者環境，其發音特性有一定程度的差異，例如：主播與記者大多因受過發音訓練而發音咬字比較正確、清晰，說話文字內容較符合文法規律性；然而受訪者則大多爲一般民眾，所以說話必較含糊、情緒化而且含有較多口語現象。且主播在攝影棚內其錄音環境也與外場記者及受訪者有異。在 MATBN 第一年與第二年的語料中，三種語者環境個別的語者個數統計大致如下：(1)內場主播：4 人，(2)外場記者：89 人及(3)受訪者：3429 人。而在 MATBN 第一、二年的語料庫中三種環境之含語音的語料所佔時間比例分別爲如下：(1)內場主播：18.23%，(2)外場記者：40.38%及(3)受訪者：41.39%。在論文中將可用語料中 9/10 將當作訓練語料，1/10 則作爲測試語料，依不同語者環境其統計資料如表二所示。

表二、各環境下的訓練語及測試料數量(表中數字分別爲爲 訓練/測試)。

| 語料環境 | Turn 數 | 中文字數 | 時間（小時） |
|---|---|---|---|
| 內場主播 | 2,071/190 | 175,194/14,906 | 10.1/0.84 |
| 外場記者 | 2,167/210 | 104,960/9,279 | 5.8/0.5 |
| 外場受訪者 | 1,666/18 | 99,039/10,377 | 6.4/0.63 |

由於廣播新聞語料的錄製不像於朗讀語料(read speech)有準備好的文字稿，因此其聲音特性比較類似口語語音(Spontaneous Speech)，所以語料中含有許多因爲說話口氣、思考及情緒等因素而產生的聲音。接下來，便列出幾個口語語音語料中比較常見的一些口語現象－

(1) Particles

口語語音中最常見的現象就是 particles，語言學上稱之爲「感嘆詞」，particle 又可分爲 discourse particle 與 grammatical particle 兩大類，但在 MATBN 中會將 discourse particles 及 grammatical particles 標示成一類，並無特別區分。Particle 常見的例子是：「爲什麼這樣 NEI？」，其中「NEI」便是一個 particle。

(2) Paralinguistic Phenomena

口語語音中另一個普遍的口語現象便是一些 paralinguistic phenomena，例如：笑聲、嘆氣聲、砸嘴聲等。

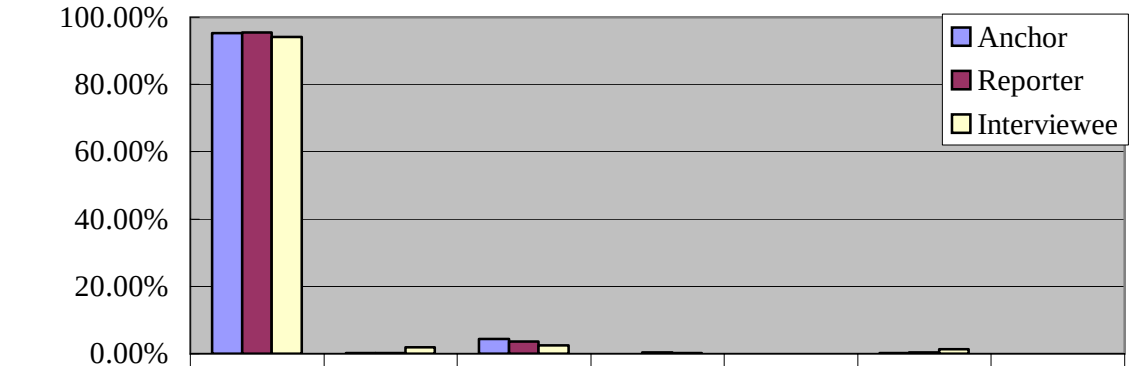
(3) Pronunciation Error

不同於一般 read speech，廣播新聞中語者的說話內容並沒有經過設計，因此發音不正確的情形便可能存在，例如發音偏差（Inappropriate Pronunciation）與音節合併（Syllable Contraction）等現象，各舉一個較常見的例子，如「發生」卻唸成「ㄏㄨㄚ生」與「這樣子」讀爲「ㄐㄧㄢˋ子」。

(4) Foreign Language

包含所有非國語的語言，因爲本土化及國際化的趨勢，即使是國語廣播新聞中也經常可聽到一些方言或外國語言穿插其中。

接著再由所選取的訓練語料中，分別統計國語 411 音、particles、呼吸聲、英語、閩南語這兩種較常見的外國語言跟方言、發音偏差以及其他現象（笑聲、砸嘴聲等 paralinguistic phenomena 及一些無法處理的聲音現象），並畫出其比例統計圖，從中觀察比較三種語者環境的特性差異。




| | 411 syllable | Particle | breath | English | MinNan | Inap_pron | Others |
|---|---|---|---|---|---|---|---|
| Anchor | 95.27% | 0.15% | 4.28% | 0.08% | 0.01% | 0.21% | 0.01% |
| Reporter | 95.36% | 0.24% | 3.65% | 0.36% | 0.02% | 0.36% | 0.01% |
| Interviewee | 94.20% | 1.81% | 2.48% | 0.13% | 0.07% | 1.31% | 0.09% |

圖二、各種語者環境之語料現象比例圖。

從圖二中可看出，內場主播的呼吸聲所佔的比例最高，這是因爲呼吸聲出現的次數通常隨著一句話的長度越長而增加，而三種環境的句子平均長度由長到短爲：內場主播、外場記者、受訪者，正好與此呼吸聲比例統計符合；外場記者與受訪者的語料中，非 411 音的比例均比內場主播高，而又以受訪者的 particle 和其他現象比例最高，也和預期相吻合。除此之外，我們再觀察發音偏差（Inappropriate Pronunciation）的現象，在各個語者環境的比例統計如下：(1)內場主播：0.21%，(2)外場記者：0.36%及(3)受訪者：1.31%。

因爲三種語者之語者環境無論在音質、語音內容上均有十分大的差異，所以接下均針對這三種語者環境的語料分開處理，對每種環境各別進行訓練其語音辨認模型。

## 三、 基本語音辨認系統之建立

在這節中將分別對MATBN中三種不同語者環境分別建立其基本國語 411 音節辨認器。在論文中，是使用Cambridge所發展的HTK(Hidden Markov Toolkit)及HLM(HTK language model tools)[7]語音辨認器發展工具來發展論文中的國語新聞廣播語料辨認系統。在系統中所使用的語音特徵向量爲 12 維MFCC、12 維ΔMFCC、12 維ΔΔMFCC、ΔE及ΔΔE，共 38 維；語音信號先經過 $1-0.97z^{-1}$的預強調後，取音框大小爲 30msec，每秒 100 個音框來求取參數；並使用了CMS(Cepstral mean subtraction)方法來去除部分語者及通道效應。

因爲 MATBN 廣播新聞語料庫之轉述資料是以 BIG5 碼形式標示，所以首先必須將 BIG5 轉成注音或拼音形式，在中文字轉音時會遇到破音字的問題，在此我們先對具有一字多音的破音字轉爲其最常使用的音；待建立語音辨認模式後，我們會利用語音辨認器去自動重新標示語料庫中各破音字之讀音。

在建立基本語音辨識系統時，由於輸入語音信號使用的是 speaker turn 爲單位，每段語料可能長達數百個音節，所以我們先使用由國語朗讀語料庫 － TCC-300 [8]所訓練的國語 411 音節 HMM 辨認模型做 force alignment 獲得 MATBN 訓練語料的音節切割位置，利用這些音節切割位置使用 isolated word 的訓練方法用音節爲單位來訓練廣播語料之初始 HMM 模型；對朗讀語料中不存在的辨認音節模型，如：particle，則使用相近音模型取代。Paralinguistic phenomena 資料方面，僅對出現較爲頻繁的呼吸聲，我們使用人工切割標示數百筆資料用以訓練起始 HMM 模型。

得到廣播新聞語料之初始 HMM 模型後，我們就可以去在訓練國語廣播新聞語料的 411 音節 HMM 辨認模型，並依資料多寡來調整 HMM 模型中各狀態下高斯分佈的個數(number of mixtures)；對訓練語料太少的模型，例如：particle，則先與相近音結合(tie)；最後無法找到相近音或聲音分佈差異極大者，如：閩南語、英語，我們建立了三個填充模型分別用來描述閩南語、英語及少見的 particle。在各環境下 HMM 參數設定如表三所示。

對三個語者環境下之 411 音節辨認率則如表四所示，其中統計辨認率時，是將非 411 音節之 particles、paralinguistic phenomena 及非國語語言由答案及辨認結果去除後，做結果與答案之比對已獲得國語廣播新聞語料之 411 音節辨認率。對於表四中之結果，我們可以發現：

(1) 內場主播因爲受過專業發音訓練，而且人數少、大多數語料均爲同一位語者(公視主播葉菊蘭小姐)的聲音，此外主播所在的環境安靜且錄音品質也較好，所以有較高的辨識率。

(2) 外場記者雖然咬字也很清晰，又因爲人數較多且處在較吵雜的環境、錄音器材也較差，所以辨識率不如內場主播。

(3) 受訪者因爲人數眾多，說話比較口語化、發音比較不正確；而且環境雜訊較多，所以辨識率與之前兩者相較之下有一段不小的差距，且插入及刪除型錯誤也大幅提昇。

由於內場主播與外場記者的語者數都不算多，因此建立的系統只能算是多語者(multi-speaker)辨識系統，且語音內容應該屬於 plain speech；但是受訪者的辨識器則是由上千名語者的聲音建立，真正是語者獨立(speaker independent)的辨識系統，而且是 spontaneous speech 且錄音品質較差；基於如此的差異，必定將造成受訪者的辨識系統會得到較低的辨識率。

表三、各環境下(anchor/reporter/interviewee)HMM 參數設定。

| HMM 模型種類 | 個數 | 狀態數 | Mixture/狀態 |
|---|---|---|---|
| 聲母 | 100(RCD) | 3 | 1 ~ 32 |
| 韻母 | 40 | 5 | 1 ~ 32 |
| Particle | 4/7/16 | 3 | 1 ~ 32 |
| Breath | 1 | 3 | 32 |
| Silence | 1 | 3 | 64 |
| SP (Tie to the middle state of Silence) | 1 | 1 | 64 |
| Garbage | 3 | 3 | 32 |

表四、各環境下之音節辨識率。

| 環境 | Sub | Del | Ins | Accuracy |
|---|---|---|---|---|
| 內場主播 | 18.51% | 2.88% | 0.85% | 77.76% |
| 外場記者 | 27.88% | 2.56% | 1.15% | 68.40% |
| 受訪者 | 45.73% | 6.87% | 5.08% | 42.32% |

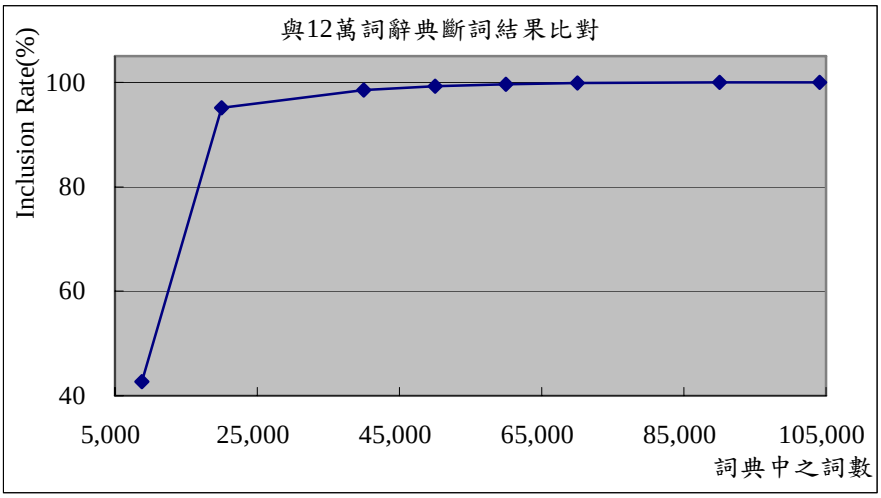
事實上，我們也做過未知環境之辨認實驗，對 anchor 及 reporter 之辨認率會下降 0.1%，而 interviewee 的之 411 音節辨認率還會提升 2%；而辨認爲正確環境的比例對 anchor/reporter/interviewee 環境分別爲 95%、93%及 83%。Interviewee 環境辨認率較低而音節辨認率會提高是因爲有些 interviewee 的錄音環境與 reporter 較爲相似，而使用 reporter 環境所建立之辨認器會得到較佳之辨認率。因爲未知環境之辨認所需之計算量非分之大，且辨認率相差不大，所以接著的實驗中均假設已知語者環境。

## 四、 加入語言模型之辨識系統

接著論文中將在國語廣播新聞語料語音辨認器中加入語言模型，做中文字與詞的辨認。

### 4.1、語言模式之建立

在論文中建立國語詞典時先建立一個由三個詞典－分別是中研院八萬詞詞庫[8]、交通大學語音實驗室自訂詞條與台師大陳柏琳教授所提供的詞典，聯集而成一個共計有十二萬四千多詞的原始詞典。但因爲記憶體容量及計算量大小等因素限制，在語音辨認器中無法使用此十二萬多詞的原始詞典。關於詞典大小的選擇，我們先利用交大語音實驗室的中文斷詞器[9]，使用上述十二萬多詞的原始詞典作爲斷詞器之詞典，將光華雜誌（Sinorama）、中文資訊檢索標竿測試集（CIRB030）以及中研院平衡語料庫（Sinica Corpus）[8] 之文字資料庫輸入做斷詞。再使用依詞頻高低來挑選後之較小的詞庫，去統計使用較小之詞典時之詞彙包含率(word inclusion rate)，結果如圖三所示。發現取六萬詞左右的詞典便能使詞彙包含率超過 99.6%，所以便將系統中之詞典大小設定爲六萬詞(59,787 詞)。上述三個中文文字資料庫除了在選擇詞典時使用外，也將作爲通用語言模型(general LM)建立的訓練資料，三個文字資料庫之統計資料則如表五所示。




圖三、選取詞典詞數與詞彙包含率之關係圖。

表五、通用語言模式(General LM)訓練語料統計。

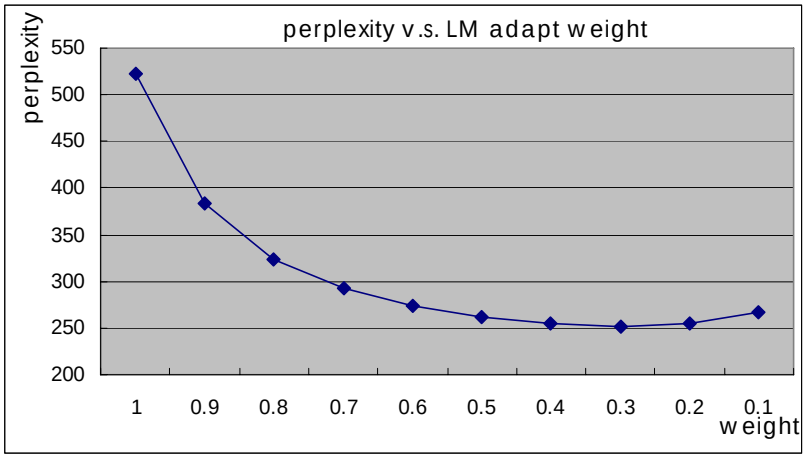
| 訓練語料 | 詞數（Word） | 字數（Character） |
|---|---|---|
| 光華雜誌 | 9,870,430 | 16,406,485 |
| 中文資訊檢索標竿測試集 | 124,442,861 | 206,847,107 |
| 平衡語料庫 | 4,796,163 | 7,972,113 |
| 合計 | 139,109,455 | 231,225,705 |

因爲前述三個中文文字資料庫並無法充分描述口語語音之語言模式，所以我們使用了 MATBN 中除被歸爲語音辨認語料部分其他所有語料之轉述文字(包含有環境雜訊及背景音樂部分語料)當調適語料來建立一個語言模型，並用來調適前述的通用語言模型，其中 particle、呼吸聲則被視爲兩個 classes；所使用的調適語料的資料統計則如表六所示。再加入語言模型到國語廣播新聞語音辨認器時，我們將調適語料之語言模型及通用語言模型加入權重值後相加，經加上不同權重值後再計算測試語料之 perplexity 後，發現爲調適語料語言模型之最佳權重值爲 0.3，此時

測試語料使用經調適後之語言模型可獲得最低之 perplexity，255.0。

表六、MATBN 調適資料之統計。

| MATBN 文字資料 | 中文詞數 | 中文字數 | Particle | 呼吸聲 |
|---|---|---|---|---|
| 數量 | 1,309,020 | 2,249,724 | 23,314 | 90,052 |




圖四、使用不同調適權重值之語言模型 perplexity。

### 4.2、MATBN 語料中破音字之重新標示

對訓練及辨認語料中所有破音字，我們使用了 force alignment 的技術去標示所有可能之讀音之分數，以獲的語料之正確讀音；對 anchor、reporter、interviewee 三種語者環境之語料分別更改了 0.8%、1.0%及 1.1%的次音節(sub-syllable)標示。

在辨認器中則依破音字出現頻率及其不同發音之 perplexity 高低，選擇加入 27 個破音字，因語言模式是由文字統計所以無法得知發音，於是對破音字我們加入了破音字讀音機率之參數，如下式所示：

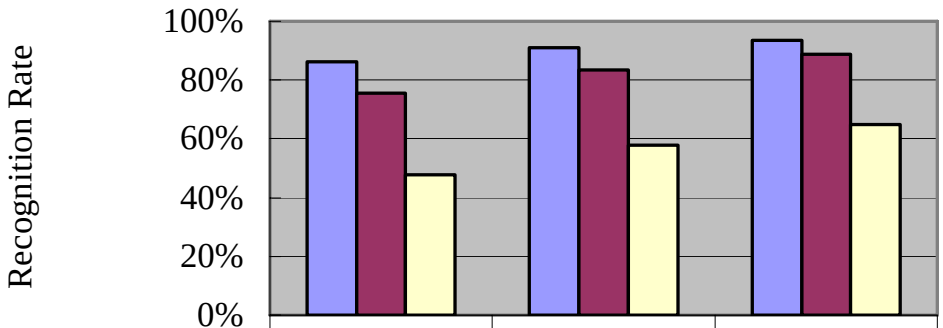
$$\log\left(P_{new}(w_i \mid w_{i-1} = Big5_j)\right)$$
$$= \log\left(P(w_i \mid w_{i-1} = Big5_j)\right) + \log\left(P(pinyin \mid big5_j)\right)$$

表七、辨認器加入的常用破音字。

| 系統中選取之一字詞破音字 |
|---|
| 了、地、行、佛、沒、那、和、的、長、重、哪、差、參、得、從、都、曾、朝、給、著、說、彈、樂、調、親、還、露 |

### 4.3、加入語言模型之辨識效能

更正破音字標音錯誤並加入語言模型後之語音辨認器效能如圖四所示，對 anchor/reporter/interviewee 的詞錯誤率約爲 14%、26%及 52%。前面提過，Anchor 及 reporter 的辨認結果可以說只是 multi-speaker 的辨認結果，尤其是 anchor 幾乎就是一個主播的語音，而且其內容應該是 plain speech 而非 spontaneous speech。對 Interviewee 的語料應當屬於 spontaneous speech，與 Hub-4 中的 Mandarin call home 語料庫之辨認結果比較[10,11]，詞錯誤率會略高但均在 50-60%左右；當然 call home 語料還到考慮到電話通道效應的影響。若以三種環境之辨認率平均，則平均詞與字錯誤率 30%、22%；與 Dragon 公司對 Hub-4 中之國語廣播語料之評估效能比較，詞與字錯誤率分別下降了 6%及 3%。




| | Word | Character | Syllable |
|---|---|---|---|
| Anchor | 86.29% | 91.04% | 93.62% |
| Reporter | 75.59% | 83.37% | 88.70% |
| Interviewee | 47.84% | 57.99% | 64.82% |

圖四、加入語言模型後之語音辨識率。

事實上在國語語音辨認的語言模式中還有許多的課題在此論文都還爲考慮，例如：『台』與『臺』、『的』與『地』、中文大小寫數字、…等同音同意義但在文字表示中可以使用不同的字的問題均未在此考慮。

## 五、 加入音節間靜音長度模型之辨認系統

音節中間靜音長度是韻律(prosody)參數中簡單但是又能改善語音辨認率的一項聲學參數。尤其在國語語音中它可以幫助辨認句子、片語及詞等單元的邊界，以提升國語語音辨認時詞及語句(utterance)邊界的正確率；這也就是論文中，語音辨認器之輸入音檔是以 speaker turn 爲單元，而不是以語句爲單元的原因，如此就可以統計語句邊界之辨認率。

在此先統計三種語者環境的平均說話速度（Speaking Rate），由快到慢依序是 anchor－5.55 音節/sec、reporter－5.27 音節/sec 與 interviewee－4.93 音節/sec，其中又以受訪者的語者說話速度差異較大而分布範圍最廣。因三種語者環境的說話速度有所差異，所以在此也將根據不同語者環境，分別建立符合其特性的音節間靜音長度模型。而音節間靜音長度則會因是否爲詞邊界或是否

存在標點符號(punctuation mark, PM)，或以語音信號觀點而言，是否爲一個韻律邊界而會有不同。所以我們在表八中先統計訓練語料中各標點符號出現之次數；平均每 12.6 個字會出現一個標點符號。

接著將 MATBN 訓練語料區分爲三種語者環境，統計詞內的音節間、詞間及標點符號處有靜音存在所佔有的比例，這裡我們將標點符號分爲三類，統計結果如表九所示。觀察表九中之數據發現，anchor 因說話速度較快，所以在標點符號處停頓的機率較小，而 interviewee 則有高達 60% 的機率在標點符號處會停頓；但三類語者環境在說話時都幾乎不會在詞間停頓。

表八、MATBN 語料標記之 PM 數量統計與分類

| MATBN | ， | 、 | 。 | ： | ； | ？ | ！ |
|---|---|---|---|---|---|---|---|
| 數量 | 124,520 | 4,318 | 46,320 | 56 | 79 | 2,950 | 24 |
| 分類 | COM | DOT | OTHERS | | | | |
| 數量 | 124,520 | 4,318 | 49,429 | | | | |

表九、MATBN 訓練語料詞內(Intra-word)與詞間(Inter-word)是否存在靜音之統計表。

| Environment | MATBN | Inter-word | | | | Intra-word |
|---|---|---|---|---|---|---|
| | | PM_COM | PM_OTHERS | PM_DOT | NON_PM | |
| Anchor | Total number | 8,676 | 1,763 | 239 | 94,956 | 77,706 |
| | With pause | 33.7% | 39.5% | 39.7% | 10.3% | 1.5% |
| | Without pause | 66.3% | 60.5% | 60.3% | 89.7% | 98.5% |
| Reporter | Total number | 5,309 | 629 | 373 | 59,011 | 44,414 |
| | With pause | 48.0% | 60.1% | 46.9% | 7.8% | 1.2% |
| | Without pause | 52.0% | 39.9% | 53.1% | 92.2% | 98.8% |
| Interviewee | Total number | 5,611 | 277 | 211 | 58,609 | 42,386 |
| | With pause | 60.2% | 66.4% | 49.3% | 13.2% | 2.0% |
| | Without pause | 39.8% | 33.6% | 50.7% | 86.8% | 98.0% |

在此將辨識系統之測試語料的文字內容留下分類後之標點符號標記，接著計算考慮標點符號之調適後語言模型之 perplexity，結果發現 perplexity 從原本不含標點符號時的 255.0 下降爲 249.2，由此可發現加入分類之標點符號後確實能夠令語言模型的效能有所提升。

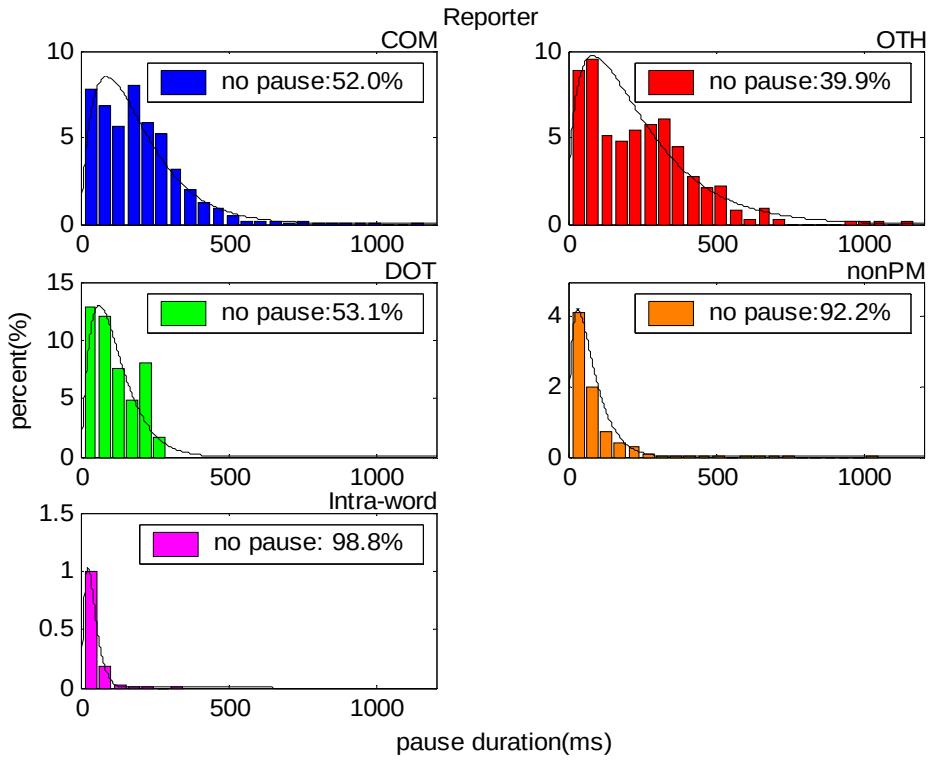
在論文中我們使用 Gamma distribution 來描述音節間靜音長度，但由表九結果得知，每種情形均有音節間靜音長度長度爲 0 情況，而且佔有相當程度的比重，所以最後使用下列機率來描述音節間靜音長度之分佈：

$$f_D(d)=\begin{cases} w & ,d=0 \\ (1-w)\,f(d) & ,d>0 \end{cases}$$

其中 $f(x)$ 是一個 Gamma distribution; $f(x)=\frac{\lambda(\lambda x)^{\alpha-1}e^{-\lambda x}}{\Gamma(\alpha)} \quad \forall x>0$ 。

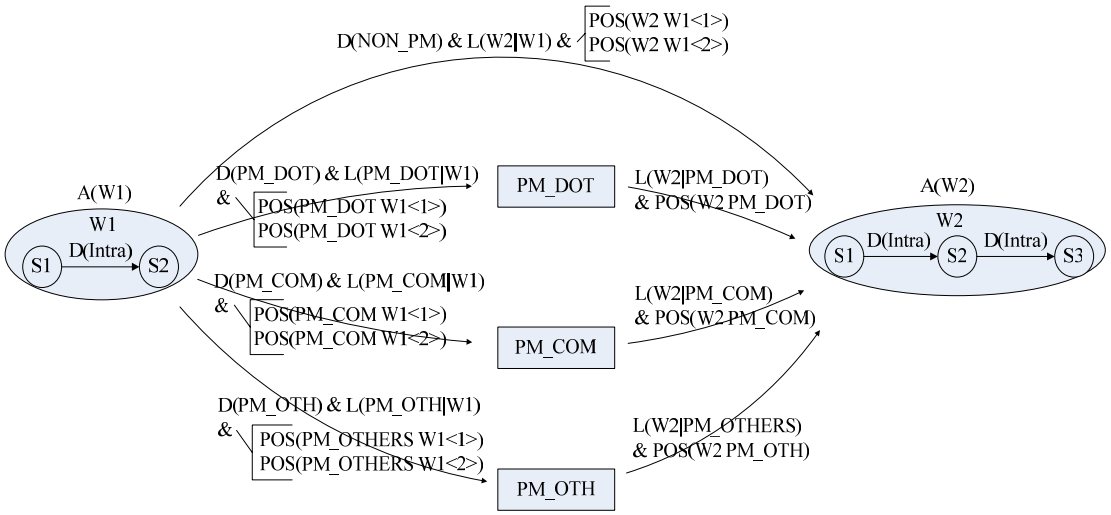
在此僅列出 reporter 環境下音節間靜音長度及其是否爲標點符號或詞邊界之關係，如圖五所

示。可以發現要單由音節間靜音長度去判斷是否有標點符號之存在或是否爲時邊界將是一件不可能的工作，但是與語言模式一起考慮後也許可以正確判斷出標點符號之存在，或者說是重要及次要的韻律或語句邊界(utterance boundaries)。




圖五、標點符號及音節間靜音長度之關係(reporter)。

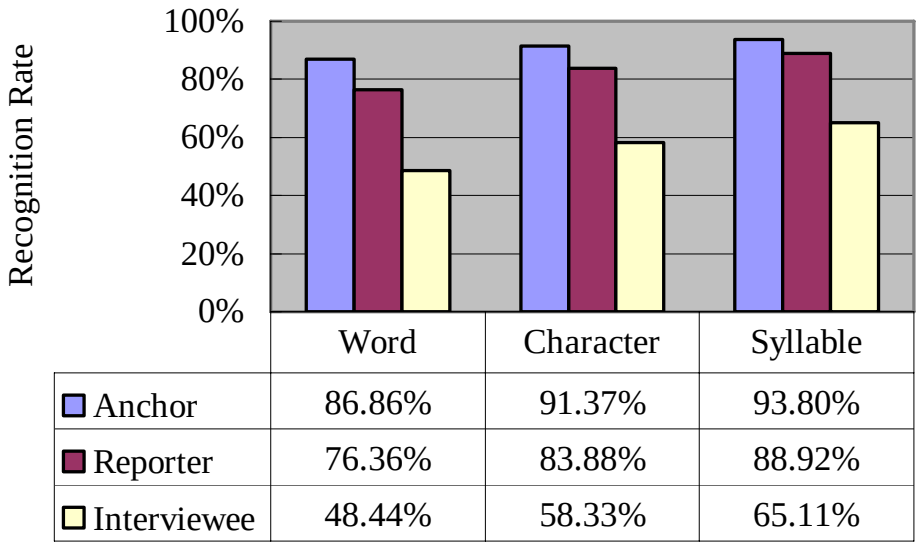
在加入音節間靜音長度之機率及標點符號之語言模式後，在辨認時 syllable 及 word 轉換時所要考慮的辨認分數可由圖五所示，在音節間轉移時須加上不同的音節間靜音長度機率分數，若是詞間轉移還需加上語言模型分數。




圖五、加入音節間靜音長度模型後之辨認分數之示意圖。

因爲 HTK 軟體的限制，在辨認過程我們是使用 two-pass 的方法來將音節間靜音長度之機率加入辨認結果。第一步先使用 HTK 找出待辨認輸入語料之 word lattice，其中設定同一時間允許的最大的 token 數爲 10；第二步再加上音節間靜音長度模型之分數；經重新計分(rescoring)後，找出最佳辨認結果。

在加入含標點符號之語言模式及音節間靜音長度模型後之語音辨認器之效能如圖五所示，對 anchor/reporter/ interviewee 的音節辨認率上昇 0.2-0.3%，字辨認率上昇 0.3-0.5%而詞辨認率亦可上昇 0.2-0.3%。




| | Word | Character | Syllable |
|---|---|---|---|
| Anchor | 86.86% | 91.37% | 93.80% |
| Reporter | 76.36% | 83.88% | 88.92% |
| Interviewee | 48.44% | 58.33% | 65.11% |

圖六、加入語言模型及音節間靜音長度後之語音辨識率。

但是我們想看一下加入標點符號之語言模式及音節間靜音長度資訊後，語音辨認器對廣播語料的切割(segmentation)是否有助益；若語音辨認器可以正確的標示標點符號的位置，將可把廣播

語料切割爲較 speaker turn 還小並具有語言意義的單位。由表十標點和號之辨認率中，高達 70-80% 的標點符號位置可以被辨認出來。由表十一，我們可以看出『、』號常被辨認成『，』號或者沒有辨認出來有標點符號的存在，尤其是說話速度較快的 anchor，所以語者說話時不一定會在『，』號停頓。對說話速度較慢的 interviewee，語句中次要韻律邊界『，』與其它表示主要韻律邊界或是說語句結束的標點符號(如『。』等)的辨識率可達 95%以上。

表十、標點符號辨識率。

| 環境 | Correct and substitution | correct | Substitution | Miss detection | False alarm |
|---|---|---|---|---|---|
| Anchor | 78.93% | 67.88% | 11.05% | 21.07% | 13.95% |
| Reporter | 83.99% | 77.11% | 6.88% | 16.01% | 19.94% |
| Interviewee | 67.91% | 66.18% | 1.73% | 32.09% | 24.37% |

表十一、標點符號標記辨認之 confusion table (anchor/reporter/interviewee)。

| 辨識結果 / 正確答案 | PM_COM | PM_OTHER | PM_DOT |
|---|---|---|---|
| PM_COM | 96.2/98.2/99.4% | 3.3/1.5/0.6% | 0.6/0.3/0.0% |
| PM_OTHER | 32.0/11.3/4.7% | 68.0/88.7/95.3% | 0.0/0.0/0.0% |
| PM_DOT | 100.0/93.8/50.0% | 0.0/0.0/50.0% | 0.0/6.3/0.0% |

## 六、 結論及未來展望

在本論文中，對國內自行錄製的國語新聞廣播語料庫，MATBN，做了基本的語音辨認之效能評估對國語廣播新聞中的三種不同語者環境－主播、外場記者及受訪者，分別得到 86.9%、76.4% 及 48.5%的詞辨認率。就語音辨認器的觀點，本論文中的辨認系統就還有許多課題值得探討，例如：加入音節長度模型、加入基頻資訊就是可以立即提高語音辨認器效能的方法。國內有一個 MATBN 這樣大型的國語新聞廣播語料庫相信在 segmentation、information extraction、topic detection 或語言學方面也可以進行許多有趣的研究課題。

## 七、 參考文獻


1. Richard Stern, “ Specification of the 1996 Hub 4 Broadcast News Evaluation,”, 1997 DARPA Broadcast News Workshop.
2. Fiscus, John S. Garofolo, Alvin Martin, Mark A. Przybocki, David S. Pallett, Jonathan G., “ 1998 Broadcast News Benchmark Test Results,”, 1999 DARPA Broadcast News Workshop.
3. Puming Zhan, Steven Wegmann, Steve Lowe, “ Dragon Systems' 1997 Mandarin Broadcast News

System,”, 1999 DARPA Broadcast News Workshop.

4. Hsin-Min Wang, Berlin Chen, Jen-Wei Kuo and Shih-Sian Cheng, „ MATBN: A Mandarin Chinese Broadcast News Corpus, “, Computational Linguistics and Chinese Language Processing, Vol. 10, No. 2, June 2005, pp. 219-236.
5. http://sovideo.iis.sinica.edu.tw/SLG/corpus/MATBN-corpus.htm.
6. C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, “Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech, “, *First International Conference on Language Resources and Evaluation (LREC)*, pp. 1373-1376, May 1998. also http://www.ldc.upenn.edu/mirror/Transcriber/.
7. S. Young, G.. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, *The HTK Book*（for HTK Version 3.2.1）.
8. Speech Database in *The Association for Computational Linguistics and Chinese Language Processing*, http://www.aclclp.org.tw/corp_c.php.
9. 江振宇，『中文斷詞器之改進』，交大碩士論文，2004。
10. Fu-Hua Liu, Michael Picheny, etc. “Speech recognition on Mandarin Call Home: a large-vocabulary, conversational, and telephone speech corpus”, pp. 157-160, ICASSP 96.
11. Ming-yi Tsai, Lin-shan Lee, "Pronunciation modeling for spontaneous speech by maximizing word correct rate in a production-recognition model", in *SSPR-2003*, MAP6.